「回帰分析（1）：とにかくやってみる」

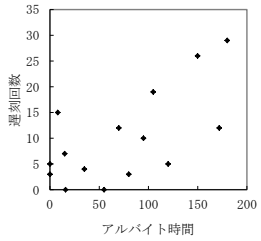
■回帰分析の目的と魅力〔テキスト pp.87-89〕

・応用的な分析技法で圧倒的によく用いるのは**回帰分析**（regression analysis）。
・ある 1 つの変数の値を、他の変数の値で説明したい⇒回帰分析を使う。

独立変数 ——説明——→ **従属変数**
（independent variable）　　（dependent variable）

例）授業の遅刻回数の多さをアルバイト時間の長さで説明したい。

・発想は、散布図の上に直線を引いてみることと同じ。



・直線を引く発想をしつこく考えると……
⇒「X と Y の関係は、本来この直線のような関数で表せるのではないだろうか。実際のデータがこの直線からいくらかずれているのは、何らかの誤差によるものだろう」
⇒だいたい Y=a+bX という関数が成り立つが、実際は **Y=a+bX+e** と誤差がある。
⇒最適な a や b を特定すれば、最適な直線が引けるはず。
例）Y=4.5+0.1X+e

・**回帰線**（regression line）……予測される直線のこと
・**回帰式**（regression equation）……回帰線を表す式のこと
切片である a を回帰式の定数項、
傾きである b を回帰式の**回帰係数**（regression coefficient）と呼ぶ

・独立変数が複数の場合（重回帰分析とも呼ぶ）でも、考え方はまったく変わらない。
$Y=a+b_1X_1+b_2X_2+b_3X_3+\cdots\cdots+e$

・回帰分析の目的……「回帰線を最適に調整することを通して、ある変数の値が、その原因と考えられる変数によってどのように説明できるのかを統計的に明らかにする」
・回帰分析の魅力……X と Y の間に何らかの因果関係を想定したときに、実際に X が Y にどれだけの影響を与えるのか「具体的な」関係を知ることができる。例）アルバイト時間が 1 時間増えるごとに遅刻回数は 0.1 回増える

作業課題①

(1) テキスト p.259 の Web ページから「重回帰分析のデータ」をダウンロードする

(2) 以下の設定で回帰分析を実行してみる

従属変数 Y＝老後幸福感
独立変数 X_1＝年齢
X_2＝世帯人数
X_3＝年間世帯所得

(3) 結果を回帰式で表して、意味を読み取る。

■回帰分析の結果の要点〔テキスト pp.89-90〕

(1) もっともよい線を引く。

⇒最適な定数項 a と回帰係数 b_1, b_2,……を読み取る。※最重要！

(2) その線は全体としてどのくらいよい線であるかを評価する。

⇒決定係数 R^2 で説明力を％で表す。

(3) 母集団についても同様の線を引く価値があるかどうかを判断する。

⇒F 値を用いた検定の有意確率を確かめる。

(4)［重回帰分析の場合］各独立変数の効果が母集団でも有効といえるか判断する。

⇒t 値を用いた検定の有意確率を確かめる。

作業課題②

(1) 作業課題①の出力をもう一度見直して、下の表を完成させよう。

	B	t	p
（定数）			
X_1　年齢			
X_2　世帯人数			
X_3　年間世帯所得			

n=＿＿＿＿＿、調整済み R^2=＿＿＿＿＿、F=＿＿＿＿＿、p＝＿＿＿＿＿

(2) 適切な数値を読み取って、以下の文章を穴埋めしよう。

老後幸福感の得点を 3 つの独立変数で予測する回帰分析を行った。その結果、老後幸福感は年齢が 1 歳上がるごとに約＿＿＿＿点｛上がり・下がり｝、世帯人数が 1 人多いごとに約＿＿＿＿点｛上がり・下がり｝、年間世帯所得が 1 万円多いごとに約＿＿＿＿点｛上がる・下がる｝ことがわかった。たとえば、86 歳、3 人家族、所得 700 万円の人は、老後幸福感が＿＿＿＿点と予想される。ただし、このうち有意水準 5%で有意な効果が認められたのは、＿＿＿＿＿＿＿＿だけであった。全体として、この回帰式で老後幸福感の個人差が＿＿＿＿％説明できる。この説明力は有意水準 5%で統計的に有意｛である・ではない｝。

■SPSS でやってみよう〔テキスト pp.90-108〕

・SPSS での回帰分析の操作

①メニューから、分析➜回帰➜線型

②説明したい変数（Y）を［従属変数］、説明に使う変数（X）を［独立変数］欄へ移動

（②’　質的変数を独立変数にする場合は、あらかじめダミー変数に変換すること）

③ＯＫボタン

モデル要約 ❷

モデル	R	R2 乗	調整済み R2 乗	推定値の標準誤差
1	.662[a]	.438	.394	6.97320

a. 予測値: (定数)、work アルバイト時間。

分散分析[a] ❸

モデル		平方和	自由度	平均平方	F 値	有意確率
1	回帰	491.868	1	491.868	10.115	.007[b]
	残差	632.132	13	48.626		
	合計	1124.000	14			

a. 従属変数 late 遅刻回数

b. 予測値: (定数)、work アルバイト時間。

読み取るポイント

❶最適な回帰式の a、b

❷調整済み決定係数

❸全体的な検定結果

（重回帰分析の場合）

❹各独立変数の
影響力の検定結果

係数[a]

❶ モデル		標準化されていない係数 B	標準誤差	標準化係数 ベータ	t 値	有意確率 ❹
1	(定数)	3.009	2.841		1.059	.309
	work アルバイト時間	.095	.030	.662	3.180	.007

a. 従属変数 late 遅刻回数

・SPSS の結果を 1 つの表にまとめる

モデル要約

モデル	R	R2 乗	調整済み R2 乗	推定値の標準誤差
1	.568[a]	.322	.283	81807.816

a. 予測値: (定数)、op5schpf 中学3年生の頃の成績, ageb 年齢, xjobyr 勤続年数。

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	1.656E+11	3	55195002905	8.247	.000[b]
	残差	3.480E+11	52	6692518720		
	合計	5.136E+11	55			

a. 従属変数 szpaymox 月給

b. 予測値: (定数)、op5schpf 中学3年生の頃の成績, ageb 年齢, xjobyr 勤続年数。

係数[a]

モデル		標準化されていない係数		標準化係数	t 値	有意確率
		B	標準誤差	ベータ		
1	(定数)	-18553.808	151096.766		-.123	.903
	ageb 年齢	1620.406	3997.425	.046	.405	.687
	xjobyr 勤続年数	6772.435	1925.550	.409	3.517	.001
	op5schpf 中学3年生の頃の成績	33703.636	12112.178	.325	2.783	.007

a. 従属変数 szpaymox 月給

↓

表 1　回帰分析の結果［詳細な表記の例］

	B	t	p
（定数）	-18553.81	-.123	.903
X_1　年齢	1620.41	.405	.687
X_2　勤続年数	6772.44	3.517	.001
X_3　中 3 時の成績	33703.64	2.783	.007

n=56、調整済み R^2=.283、F=8.247、p=.000

表 2　回帰分析の結果［簡潔な表記の例］

	B
（定数）	-18553.81
X_1　年齢	1620.41
X_2　勤続年数	6772.44 ***
X_3　中 3 時の成績	33703.64 **

n=56、調整済み R^2=.283 ***

+ p<.10,　* p<.05,　** p<.01,　*** p<.001

有意確率（p 値）が 5%、1%、0.1%未満のそれぞれの水準をクリアしていたら、*印の数を増やす（星が多いほど統計的な有意性が保証される）。

10%水準をクリアしていれば有意な「傾向」があるものとして、「+」または「†」（ダガー）の記号を使うこともある。

文章課題

老化意識を、年齢・年間世帯所得・目標設定・向上心で説明する回帰分析を行い、結果を表にまとめた上で、口頭で説明しなさい。

「回帰分析（2）：比較の視点」

■回帰分析の比較の視点

・回帰分析の魅力は、各独立変数の影響力が「数量で具体化」されていること（回帰係数）
　⇒数量は「比較できる」ことに意味がある

・代表的な比較の視点
　①独立変数間の比較（標準化係数）
　②モデル間の比較（独立変数の増減）
　③対象者グループ間での比較

■比較① 独立変数間の比較（標準化回帰係数）〔テキスト p.104〕

・複数の独立変数の間でどれが一番重要な影響力を持つのか？
　⇒単純に、回帰係数を比べてはならない
　　扱っている事柄の規模が違うから
・変数の規模をそろえて（平均 0、標準偏差 1 の標準化して）から回帰分析
　⇒標準化回帰係数 β
・標準化回帰係数は、単純に数値の大小を比べて「何倍影響力が強い」と読める
・ただし、具体性は読み取りにくくなる

作業課題①

JGSS-2000 の 30 代女性データを用いて、「月給」を「年齢」「勤続年数」「中 3 時の成績（5 段階評価）」の 3 変数で説明する回帰分析を行いなさい。そして、どの独立変数が月給をもっとも強く規定するのか説明しなさい。

	B	標準化係数 β
（定数）		
X_1　年齢		
X_2　勤続年数		
X_3　中 3 時の成績		

n=＿＿＿＿、調整済み R2=＿＿＿＿＿＿

* p<.05　** p<.01　*** p<.001

■比較② モデル間の比較（独立変数の増減）〔テキスト pp.104-108〕

・「モデル」とは？…… (a)変数間の関係性の形と、(b)扱う変数をどう仮定しているか

例）幸福感が収入と余暇時間に比例して上昇すると仮定するモデル

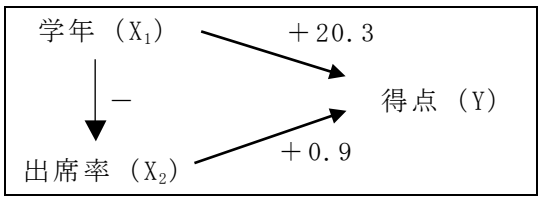
＝幸福感が従属変数（Y）で、収入と余暇時間が独立変数（X）の線形回帰モデル

・通常は、独立変数を変更した場合の比較が、回帰分析でのモデル比較

（(a)は同じで、(b)を変える）

・モデルが変わると、同じ独立変数でも影響力（回帰係数）が変わることがある

【重回帰分析の図式】



$$Y = -55.6 + 20.3X_1 + 0.9X_2 + e$$

$$Y = 22.5 + 11.5X_1 + e$$

・共線性の問題

あまりにも似たよう事項を同時に独立変数にすると、おかしな分析結果を示す

形式的には、共線性の指標 VIF が 2.0 以上だと要注意

作業課題②

（1）JGSS-2000 の 30 代女性データを用いて、次の 2 つのモデルで結果を比較しなさい。

モデル 1：「月給」を「年齢」で説明する

モデル 2：「月給」を「年齢」と「勤続年数」で説明する

	モデル 1	モデル 2
（定数）		
X_1　年齢		
X_2　勤続年数		
n		
調整済み R^2		

＊ p<.05　　＊＊ p<.01　　＊＊＊ p<.001

（2）なぜ、2 つのモデルで年齢の影響力（回帰係数）が異なるのか考察しなさい。

■比較③ 対象者グループ間での比較

・同じ回帰モデルを異なる対象者グループに適用して結果を比較する。

例）出席率が成績に影響する程度は、男子学生と女子学生でどう異なるのだろう？

・変数間の関係が「数量」で具体化されることの面白さがよくわかる。

文章課題

JGSS-2000 のデータを用いて、「月給」を「年齢」「勤続年数」「中 3 時の成績」の 3 変数で説明する回帰分析を行う。ただし、以下のように男女別に 20～50 代にグループ分けして、8 個のグループで結果を比較しなさい。

わかることを、次の点に気を付けて文章化すること。

・結果と考察を区別する

・ややこしい結果の表現は GEE アプローチに留意する

	男性 20 代	30 代	40 代	50 代
（定数）				
X_1　年齢				
X_2　勤続年数				
X_3　中 3 時の成績				
n				
調整済み R^2				

＊ p<.05　＊＊ p<.01　＊＊＊ p<.001

	女性 20 代	30 代	40 代	50 代
（定数）				
X_1　年齢				
X_2　勤続年数				
X_3　中 3 時の成績				
n				
調整済み R^2				

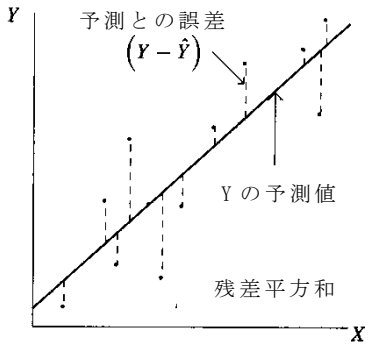
＊ p<.05　＊＊ p<.01　＊＊＊ p<.001

「回帰分析（3）：最小二乗法の理解＋ダミー変数」

■最適な回帰線はどうやって導かれているのか〔テキスト p.89〕

・回帰分析はデータから「最適な回帰線」を算出している。
・簡単にその仕組みは理解しておいた方がよい。
・もっともよい回帰線
　実際のデータと予測値のずれ（残差）$Y-\hat{Y}$ が最小になる線
⇒全体での残差の量は「残差の二乗の合計（残差平方和）」にまとめられる
⇒このやり方を**最小二乗法**（method of ordinary least squares; OLS）と呼ぶ
⇒数学的には微分方程式だが、目分量でも同じような作業はできる

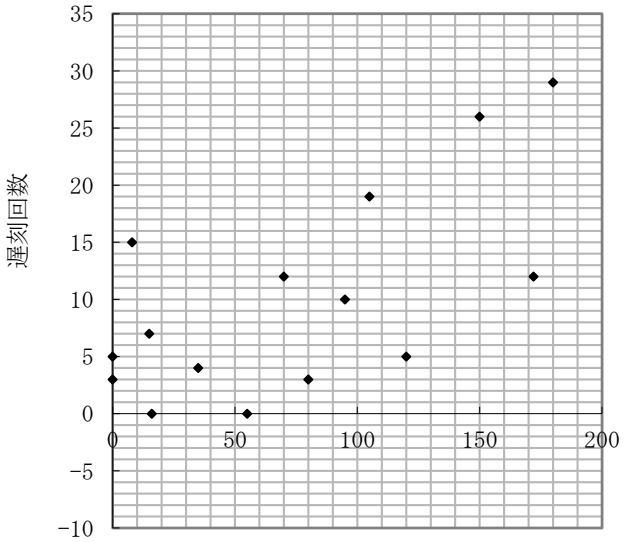
回帰式を用いた Y の予測



作業課題①

(1) 自分が最適だと思う直線を、散布図の上に定規で引いてみよう。

(2) その直線の切片と傾きを読み取って、式に表わしてみよう。



【自分が引いた回帰線】

$\hat{Y} = a + bX$

↓

$\hat{Y} =$ _____ + _____ X

(3) 自分が引いた直線について、残差平方和を求め、周りの人と比較してみよう。

	アルバイト時間 X	遅刻回数（観測値） Y	自分が引いた直線		
			予測値 $\hat{Y}$	残差 $Y-\hat{Y}$	残差平方 $(Y-\hat{Y})^2$
1 人目	55	0			
2 人目	35	4			
3 人目	180	29			
4 人目	172	12			
5 人目	150	26			
6 人目	8	15			
7 人目	80	3			
8 人目	95	10			
9 人目	0	3			
10 人目	15	7			
11 人目	16	0			
12 人目	120	5			
13 人目	105	19			
14 人目	70	12			
15 人目	0	5			
				（合計） ↓	

残差の二乗の合計 ＝ _____

(4) SPSS による分析結果と比べてみよう（分散分析表の「残差」「平方和」の欄）。

■**ダミー変数の活用〔テキスト pp.111-115〕**

・質的変数を独立変数にしたいことがある。

・質的変数は**<u>ダミー変数</u>**（dummy variable）に変換する。

└──→0 と 1 のどちらかしか取らず、量的変数としても扱える変数

・もともと 2 値の場合

	元の変数		男性ダミー		女性ダミー
男性	1	→	1	または	0
女性	2	→	0		1

例）Y が遅刻回数、X_1が学年、X_2が男性ダミーの回帰分析

$$\hat{Y} = 2.0 + 3.9X_1 + 2.2X_2$$

・3 値以上の質的変数の場合

	元の変数		文学部ダミー	法学部ダミー	工学部ダミー
文学部	1	→	1	0	0
法学部	2	→	0	1	0
工学部	3	→	0	0	1
医学部	4	→	0	0	0

例）Y が遅刻回数、X_1が学年、X_2が男性ダミーとして、
さらに X_3、X_4、X_5が文学部ダミー、法学部ダミー、工学部ダミーの回帰分析

$$\hat{Y} = 1.2 + 4.0X_1 + 0.2X_2 + (1.2X_3 - 3.2X_4 + 5.2X_5)$$

文学部　1.2＋4.0×2＋0.2×0＋（1.2×1－3.2×0＋5.2×0）＝10.4
法学部　1.2＋4.0×2＋0.2×0＋（1.2×0－3.2×1＋5.2×0）＝ 6.0
工学部　1.2＋4.0×2＋0.2×0＋（1.2×0－3.2×0＋5.2×1）＝14.4
医学部　1.2＋4.0×2＋0.2×0＋（1.2×0－3.2×0＋5.2×0）＝ 9.2

・選択肢（カテゴリー）よりも 1 つ少ない個数のダミー変数しか要らないことに注意。

・省略したカテゴリーは、比較の基準になるので重要。

参照カテゴリー［基準カテゴリー］（reference category）と呼ぶ。

・参照カテゴリーには、ある程度人数が多く明確な内容のものをあてる。（×「その他」）

作業課題②

(1) JGSS-2000 のデータを用いて、「月給」について回帰分析をする。「結婚状況」を独立変数に加えたい。適切なダミー変数に変換しなさい。

(2) 30 代男性の「月給」について、「年齢」「勤続年数」「中 3 時の成績」「結婚状況（2つのダミー変数)」を独立変数にして回帰分析をしなさい。

(3) 同じ分析を 30 代女性についておこないなさい。

	30 代男性	30 代女性
（定数）		
X_1　年齢		
X_2　勤続年数		
X_3　中 3 時の成績		
X_4		
X_5		
N		
調整済み R^2		

＊ p<.05　＊＊ p<.01　＊＊＊ p<.001

文章課題

作業課題②で得られた表を見て、男性と女性の結果の違いを記述しなさい。そのうえで、この結果について考察しなさい（なぜこのような違いが出たと思うか、15 年前と比べて現在はどうなっていると思うか、など)。