2016/11/8 (Tue) 社会調査実習授業

第2弾

社会学のための統計分析基礎

——統計ソフトSPSSを用いて

麦山 亮太 (mugiyama@l.u-tokyo.ac.jp)

東京大学大学院人文社会系研究科
社会学専門分野博士課程

［付記］ 授業準備にあたり、東京大学大学院人文社会系研究科・文学部社会学研究室が2015年度社会調査実習にて実施した調査（「A団地のくらしと地域づくりに関するアンケート」）の個票データの使用許可を得ました。記して感謝いたします。

SPSSの使用の手順

1. savファイル（またはcsvファイル）を準備する
2. データを開く
3. シンタックスファイルを開く
4. シンタックスを書き、実行
5. 結果を確認する
6. 4.と5.を繰り返す
7. 分析が終わったらシンタックスファイル**だけ**を保存
8. SPSSを閉じる

SPSSの見方｜データビュー

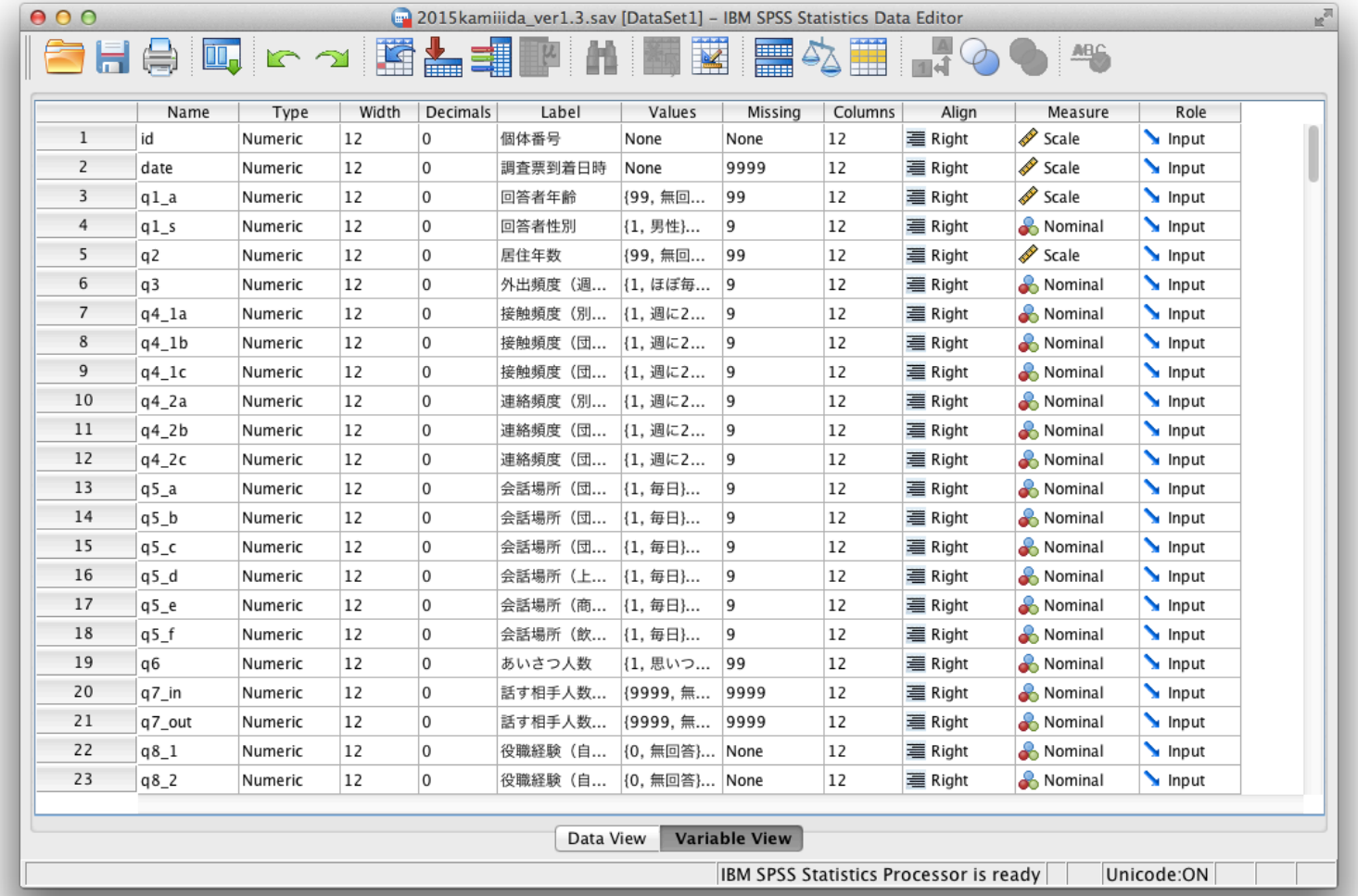
列＝変数

行＝
個体

SPSSの見方｜変数ビュー

列＝変数の情報（変数ラベル、値ラベル、変数型など）

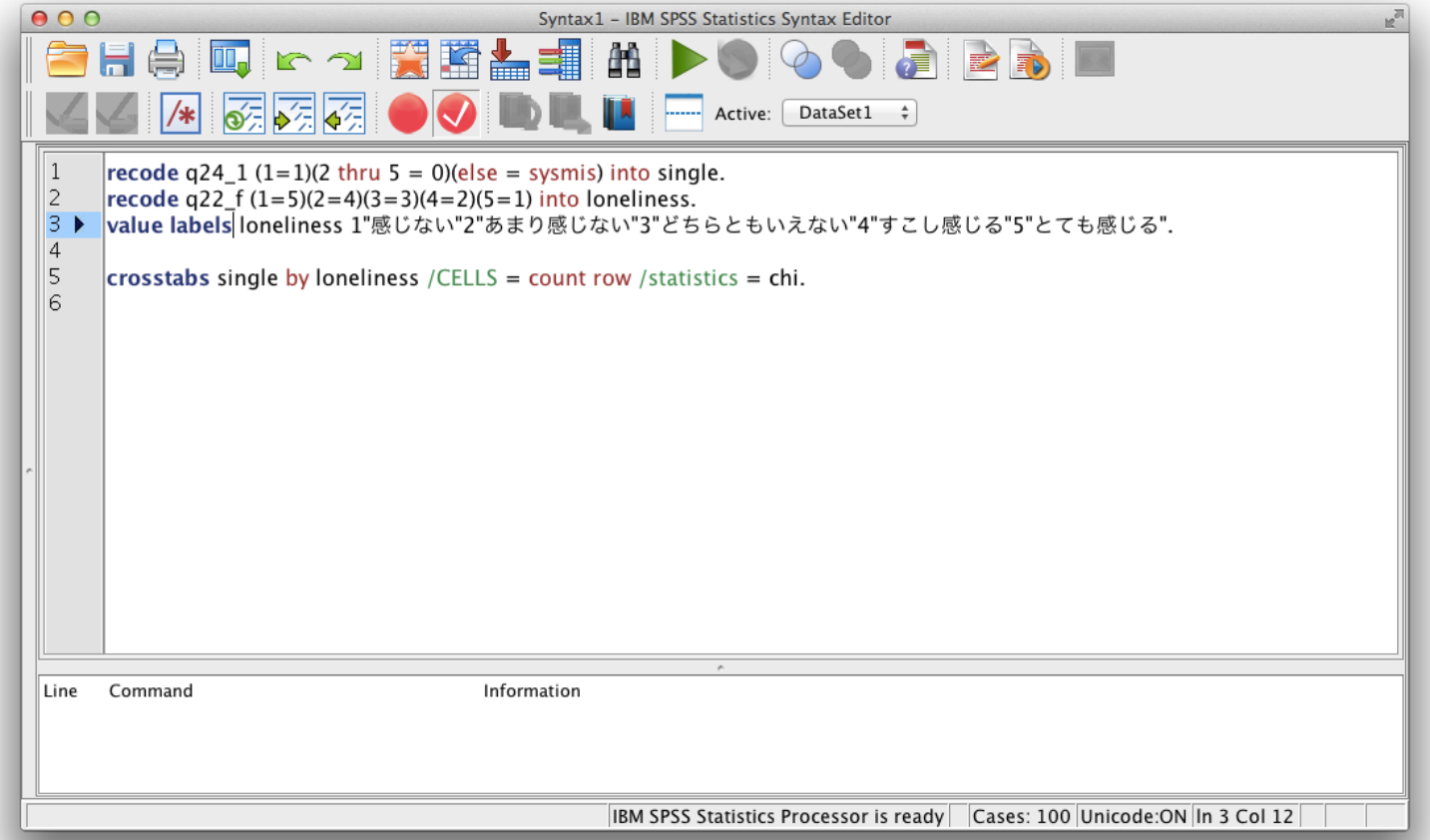
行＝
変数



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	id	Numeric	12	0	個体番号	None	None	12	Right	Scale	Input
2	date	Numeric	12	0	調査票到着日時	None	9999	12	Right	Scale	Input
3	q1_a	Numeric	12	0	回答者年齢	{99, 無回...	99	12	Right	Scale	Input
4	q1_s	Numeric	12	0	回答者性別	{1, 男性}...	9	12	Right	Nominal	Input
5	q2	Numeric	12	0	居住年数	{99, 無回...	99	12	Right	Scale	Input
6	q3	Numeric	12	0	外出頻度（週...	{1, ほぼ毎...	9	12	Right	Nominal	Input
7	q4_1a	Numeric	12	0	接触頻度（別...	{1, 週に2...	9	12	Right	Nominal	Input
8	q4_1b	Numeric	12	0	接触頻度（団...	{1, 週に2...	9	12	Right	Nominal	Input
9	q4_1c	Numeric	12	0	接触頻度（団...	{1, 週に2...	9	12	Right	Nominal	Input
10	q4_2a	Numeric	12	0	連絡頻度（別...	{1, 週に2...	9	12	Right	Nominal	Input
11	q4_2b	Numeric	12	0	連絡頻度（団...	{1, 週に2...	9	12	Right	Nominal	Input
12	q4_2c	Numeric	12	0	連絡頻度（団...	{1, 週に2...	9	12	Right	Nominal	Input
13	q5_a	Numeric	12	0	会話場所（団...	{1, 毎日}...	9	12	Right	Nominal	Input
14	q5_b	Numeric	12	0	会話場所（団...	{1, 毎日}...	9	12	Right	Nominal	Input
15	q5_c	Numeric	12	0	会話場所（団...	{1, 毎日}...	9	12	Right	Nominal	Input
16	q5_d	Numeric	12	0	会話場所（上...	{1, 毎日}...	9	12	Right	Nominal	Input
17	q5_e	Numeric	12	0	会話場所（商...	{1, 毎日}...	9	12	Right	Nominal	Input
18	q5_f	Numeric	12	0	会話場所（飲...	{1, 毎日}...	9	12	Right	Nominal	Input
19	q6	Numeric	12	0	あいさつ人数	{1, 思いつ...	99	12	Right	Nominal	Input
20	q7_in	Numeric	12	0	話す相手人数...	{9999, 無...	9999	12	Right	Nominal	Input
21	q7_out	Numeric	12	0	話す相手人数...	{9999, 無...	9999	12	Right	Nominal	Input
22	q8_1	Numeric	12	0	役職経験（自...	{0, 無回答}...	None	12	Right	Nominal	Input
23	q8_2	Numeric	12	0	役職経験（自...	{0, 無回答}...	None	12	Right	Nominal	Input

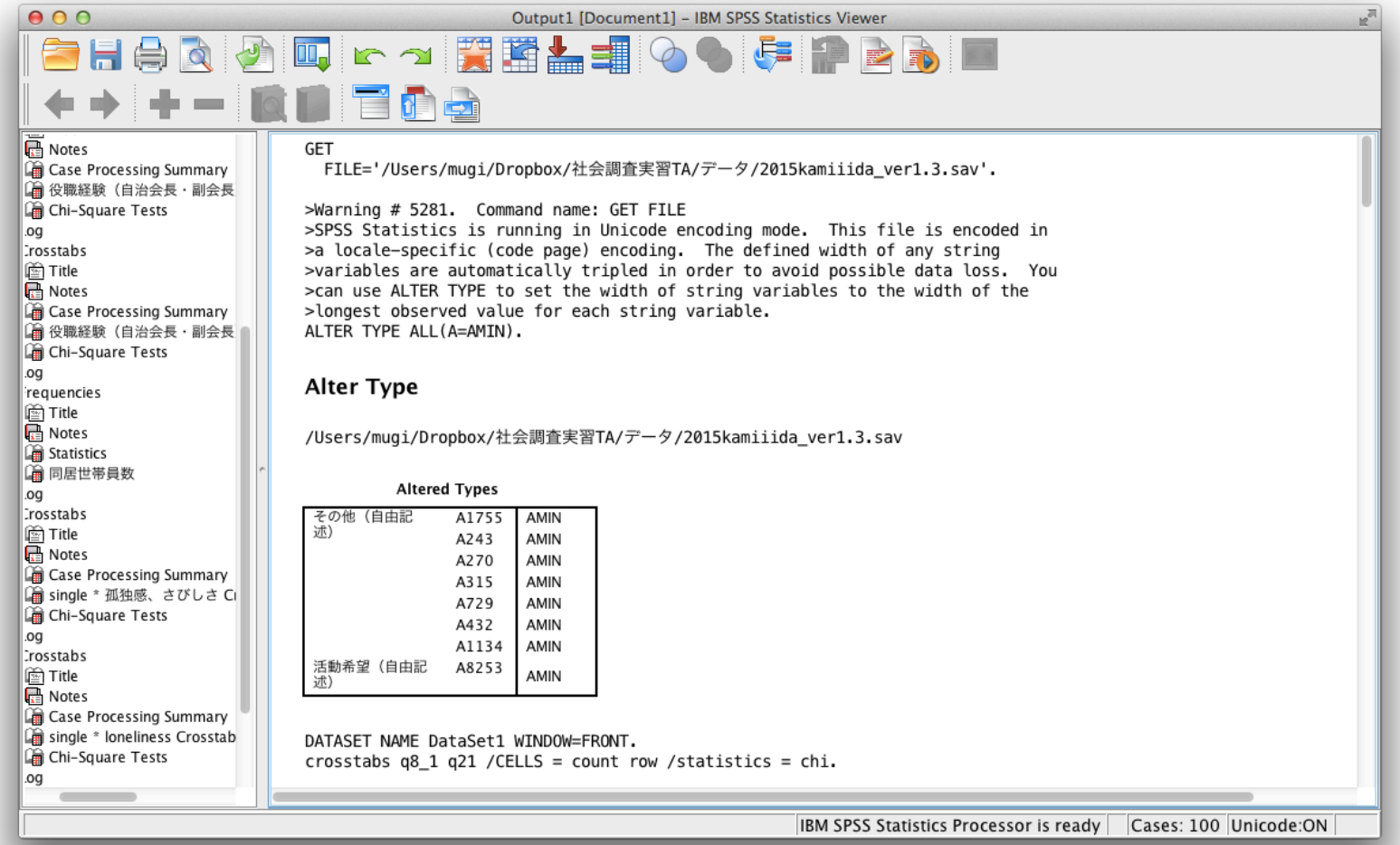
SPSSの見方｜シンタックスエディタ

上部メニュー「ファイル」→「新規作成」→「シンタックス」で、シンタックスファイルを開く。



SPSSの見方｜出力ビュー

走らせたシンタックスと、実行結果が表示される。



シンタックスの基本ルール

- シンタックスウインドウにコマンドを記述し、走らせたいコマンドをドラッグして選択し、右上の▷をクリック、またはctrl + Rでコマンドを実行。
- それぞれのコマンドははは基本的に「COMMAND /SUBCOMMAND … .」というような形をとる。 **サブコマンドの“=”はあってもなくてもよい。**
- コマンドはすべて半角英数字。ただし、変数や値に名前をつけるときは全角文字も使用できる。
- コマンドのなかに全角スペースはあってはいけない。
- 大文字と小文字を区別せず、どちらも同じ文字と判断する。
- コマンドの最後には必ず「.」が必要。
- 「.」までを1つのコマンドとみなすので、長いコマンドの途中で改行しても構わない。ただし1 行空いた場合はそこでコマンドが終了するとみなす。
- 何も起こらないときのおまじない、EXECUTE.

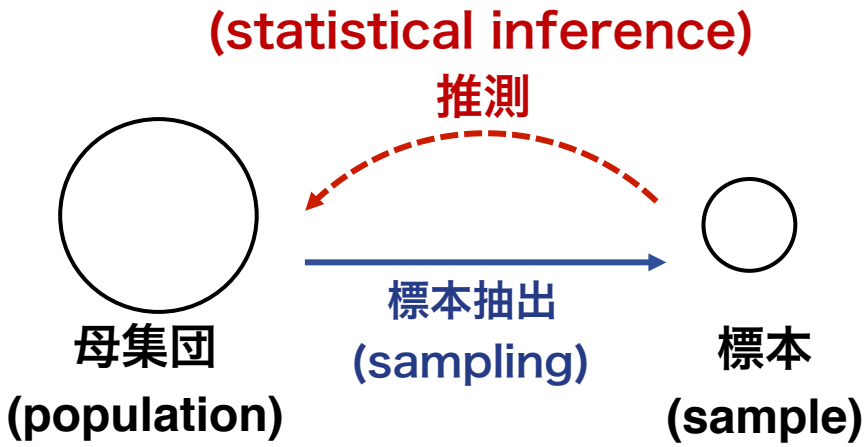
目次

目次

“量的”データと統計的分析の利点

“質的”データと比較したときの“量的”データ、およびその統計的な分析の特長

1. 変数による比較が容易である
2. 数値化に適している
3. (いくつかの仮定を満たせば) 母集団の特徴を確率的に推論できる
   →**(統計的)推測**



統計的検定の論理

頻度論的統計学における検定の論理は**背理法**による

帰無仮説（H_0）｜母集団において変数間に関連がない

対立仮説（H_1）｜母集団において変数間に関連がある

H_0が成り立つと仮定すると、サンプルから計算された統計量が得られる確率は極めて低い（慣習的には5%未満）

→ゆえに、H_0を棄却し、H_1を採択する

サンプルから計算された統計量についてH_1が採択 (H_0が棄却) されることを「**統計的に有意**である」という

関連の強さと統計的有意性

比喩的に、統計的有意性は以下のように表せる

統計的有意性＝関連の強さ×サンプルサイズの大きさ

表1　性別と投票の関連(1)

	前回の選挙で	
性別	投票した	投票せず
男性	50	50
女性	51	49

Pearson's $\chi^2 = 0.02$ ($p = 0.888$)

表2　性別と投票の関連(2)

	前回の選挙で	
性別	投票した	投票せず
男性	50000	50000
女性	51000	49000

Pearson's $\chi^2 = 22.0$ ($p < 0.001$)

表2は統計的に有意（性別と投票との間に関連がないという帰無仮説が棄却）→ 関連は強いと言えるかは**自分たちの基準に照らして判断**

有意性をどこで見ればよいか

- χ^2値とかF値とかt値とかいろいろあるけど、とりあえず（これらを確率で表現したところの） p値を見とけばOK

p値 (p-value)｜母集団における統計量が0であるという帰無仮説のもとで、サンプルから推定された統計量よりも大きい値が現れる確率。
p値が十分に小さい→帰無仮説を棄却

- 回帰分析などでは、p値が小さい係数の右肩に印をつけたりする。
  † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$ もしくは
  * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ あたりが標準的
- **P値が大きい ≠ 効果が大きい・関連が強い**
  ということにくれぐれも注意！

*すごい雑に説明してるので家に帰ったら基礎統計を復習しておいてください。

目次

復習｜変数の尺度水準

表　4つの尺度水準

	種類	順序	加減	乗除
(1) 名義尺度	○	×	×	×
(2) 順序尺度	○	○	×	×
(3) 間隔尺度	○	○	○	×
(4) 比率尺度	○	○	○	○

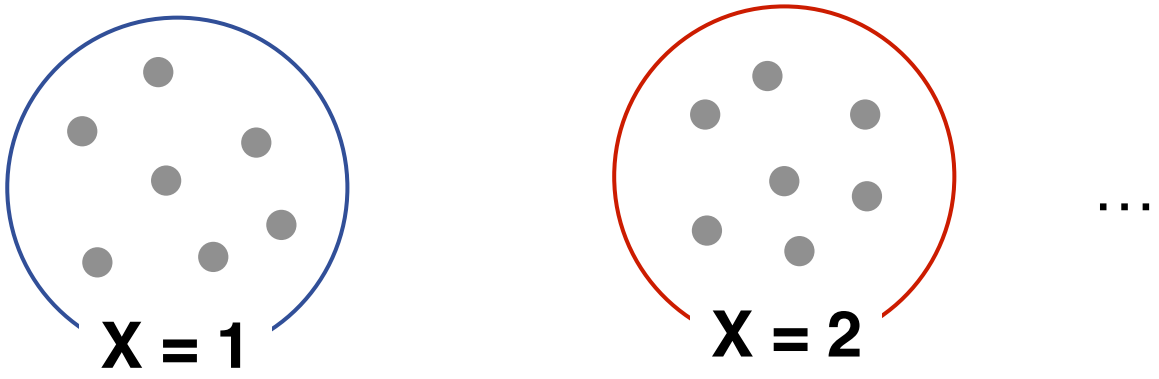
カテゴリカル変数
categorical variable

連続変数
continuous variable

- 社会学の場合、変数の多くはカテゴリカル変数。
- ただし、一定の仮定のうえで順序尺度を間隔尺度とみなす場合はしばしばある。

復習｜集団を比較する

2変量関連の分析は、集団間の比較をイメージして考えるとよい。



→それぞれについてYの分布（要約統計量）を比較

例) 男性と女性で所得はどの程度異なるのか？

カテゴリカル変数の分布を比較する＝**クロス表**

連続変数の統計量を比較する＝**平均値の比較**

復習｜度数分布表

度数分布表 (frequency table)｜変数の分布をそのまま見る方法。

→2変数の関連を分析する前に、まずは必ず1つの変数の分布を確認！

→cf. FREQUENCIES

表　孤独感の度数分布表

	度数	%	累積%
1. 感じない	134	33.5	33.5
2. あまり感じない	109	27.3	60.8
3. どちらともいえない	53	13.3	74.1
4. すこし感じる	85	21.3	95.4
5. とても感じる	19	4.8	100.0
合計	400	100.0	

出所）「A団地のくらしと地域づくりに関するアンケート」

クロス表

表　居住形態と孤独感のクロス表

居住形態	孤独感					
	感じない	あまり感じない	どちらともいえない	すこし感じる	とても感じる	Total
独居	71	78	28	61	16	254
	(28.0)	(30.7)	(11.0)	(24.0)	(6.3)	(100.0)
非独居	63	31	25	24	3	146
	(43.2)	(21.2)	(17.1)	(16.4)	(2.1)	(100.0)
Total	134	109	53	85	19	400
	(33.5)	(27.3)	(13.3)	(21.3)	(4.8)	(100.0)

Pearson's $\chi^2 = 18.07$ $(p < 0.01)$ →**推測統計のために必要な情報**

注）括弧内は行%を示す。

出所）「A団地のくらしと地域づくりに関するアンケート」

クロス表における独立性の検定

ピアソンのカイ2乗値（Pearson' χ^2）

$$\chi^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(n_{ij}-E_{ij})^2}{n_{ij}} \qquad \left(E_{ij} = \frac{n_{i.}n_{.j}}{n_{..}}\right)$$

χ^2が（自由度を考慮しても）十分に大きい → H_0（2変数が独立）を棄却

図　2変数が独立である場合に得られるクロス表（期待度数）

	孤独感					
居住形態	感じない	あまり感じない	どちらともいえない	すこし感じる	とても感じる	Total
独居	85.1	69.2	33.7	54.0	12.1	254
	(33.5)	(27.3)	(13.3)	(21.3)	(4.8)	(100.0)
非独居	48.9	39.8	19.3	31.0	6.9	146
	(33.5)	(27.3)	(13.3)	(21.3)	(4.8)	(100.0)
Total	134.0	109.0	53.0	85.0	19.0	400
	(33.5)	(27.3)	(13.3)	(21.3)	(4.8)	(100.0)

クロス表を使う際のポイント

- **行%または列%が、知りたい情報と一致しているか？**

どの変数の分布を、どのようなグループに分けて比較したいのかを考えることが大事
例) 居住形態別に孤独感を比較→孤独感の分布が分かるように

- **カテゴリが多すぎないか？**

たとえば、5×5のクロス表を見て解釈するのは難しい。できるだけカテゴリの数が少なくなるように統合するのがよい →cf. RECODE

- **2変数間に有意な関連があり、かつそれは実質的に意味のある違いといえるか？**

微小な差異を強調しすぎない、どのように違いがあるのかを確認

CROSSTABS｜クロス表(1)

```
CROSSTABS x BY y
/CELLS = COUNT ROW
/STATISTICS = CHISQ
.
```

/*補足*/
*xは行の変数、yは列の変数を表す.
*上はもっともポピュラーな場合の書き方.

CROSSTABS｜クロス表(2)

CELLSサブコマンドで使用

SYNTAX	意味	SYNTAX	意味	SYNTAX	意味
COUNT	度数	TOTAL	総%	ARESID	調整済み
ROW	行%	EXPECTED	期待度数		標準化残差
COLUMN	列%	RESID	標準化残差	ALL	すべて

STATISTICSサブコマンドで使用

SYNTAX	意味
CHISQ	カイ2乗値
PHI	CramerのV
RISK	相対リスク（オッズ比。2×2表に限る）
CORR	Pearsonの相関係数、Spearmanの順序相関係数
GAMMA	Goodman=Kruskalのγ
ALL	すべて

平均値の比較

表　居住形態別・孤独感の統計量

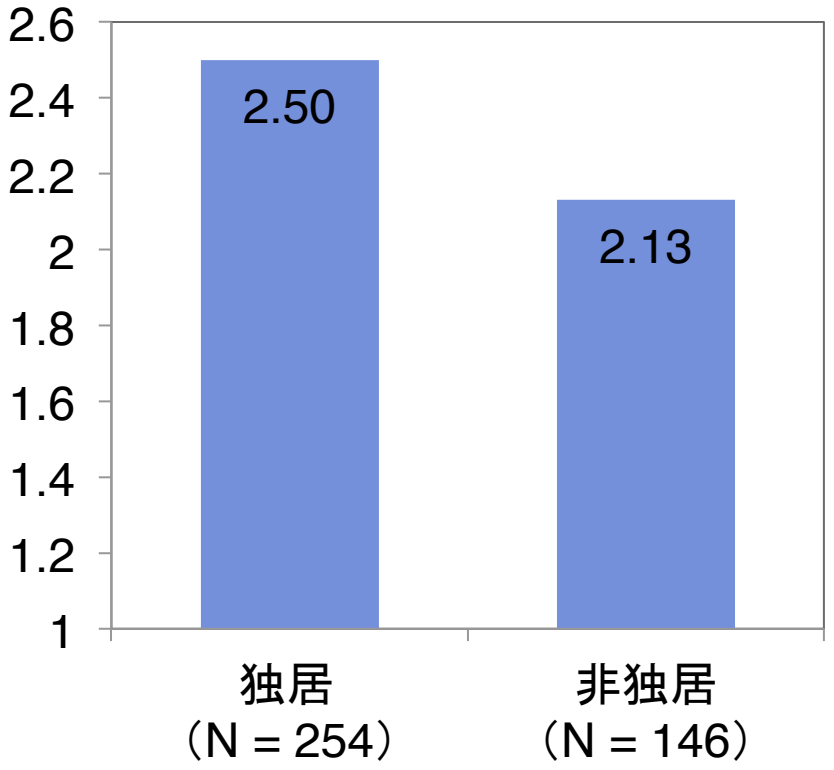
居住形態	平均	標準偏差	N
独居	2.50	1.29	(254)
非独居	2.13	1.20	(146)
Total	2.37	1.27	(400)

$F = 7.99$ ($df = 1$, $p < 0.01$)

→推測統計のために必要な情報

出所）「A団地のくらしと地域づくりに関するアンケート」

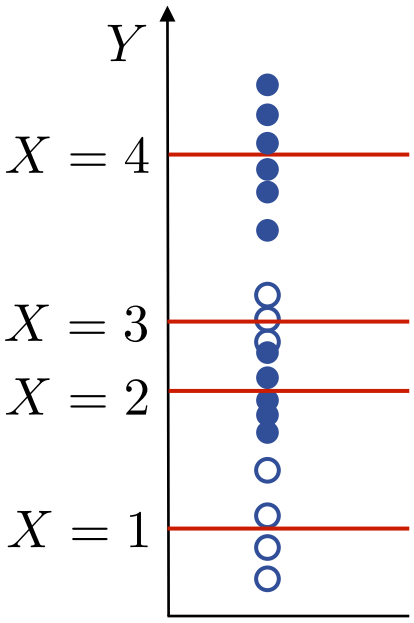
図　居住形態別・孤独感の平均値



条件つき期待値と総分散の定理

条件つき期待値（conditional expectation）

ある条件のもとで求めた平均値を意味し、$E(Y|X=1)$ などと表記する。



総分散の定理（law of total variance）

変数Yの分散は、Xを条件づけることで以下の2つの要素の和に分解できる。

$$\mathrm{Var}(Y) = \underbrace{\mathrm{Var}(\mathrm{E}[Y|X])}_{(1)} + \underbrace{\mathrm{E}[\mathrm{Var}(Y|X)]}_{(2)}$$

(1) **集団間分散** (between variance)
Xで条件づけたYの平均値の分散

(2) **集団内分散** (within variance)
Xで条件づけたYの分散の平均値

F値を用いた平均値の差の検定

F値 (分散比)

$$F = \frac{\text{Between variance}}{\text{Within variance}} = \frac{\text{Var}(\text{E}[Y|X])}{\text{E}[\text{Var}(Y|X)]}$$

Fが自由度を考慮しても十分に大きい→H_0 (母集団において集団間で平均値に差がない)を棄却

表　平均値に差がないときの統計量

居住形態	平均	標準偏差	N
独居	2.37	1.27	(254)
非独居	2.37	1.27	(146)
Total	2.37	1.27	(400)

以上の検定は (一元配置) 分散分析などと言われたりする

平均値の比較を使う際のポイント

- **知りたいことは平均値で表すのが適切か？**

順序尺度のなかでもいくつかのカテゴリを統合して、比率で表すほうが適切な場合も多い
例）孤独感の平均値→孤独感を感じるかどうかの2カテゴリに分割し、孤独感を感じる人の割合を調べる

- **カテゴリが多すぎないか？**

たくさんの集団を比較するほど、解釈が難しくなる。できるだけカテゴリの数が少なくなるように統合するのがよい →cf. RECODE

- **集団間の分散（標準偏差）はほぼ同じくらいか？**

等分散性の仮定を満たさない場合は適切な検定ができない

ONEWAY｜平均値の比較

```
ONEWAY y BY x
/STATISTICS = DESCRIPTIVES HOMOGENEITY
/PLOT = MEANS
.
```

/*補足*/
*前回は平均値の比較を行うシンタックスとしてMEANSを紹介したが、ONEWAY（分散分析のコマンド）でも同様にできるのでこちらも紹介.
*YをXで分ける。DESCRIPTIVESを入れないと平均値や標準偏差が出ないことに注意。HOMOGENEITYは等分散性が成り立つという帰無仮説の検定結果を出す。PLOTは視覚的に見たいときに使う。

連続変数どうしの関連

散布図 (Scatter plot)
2つの連続変数の値を2次元座標上にプロットした図

相関係数 (Correlation coefficient)
(主に) 連続変数同士の一致度合いを測定する指標

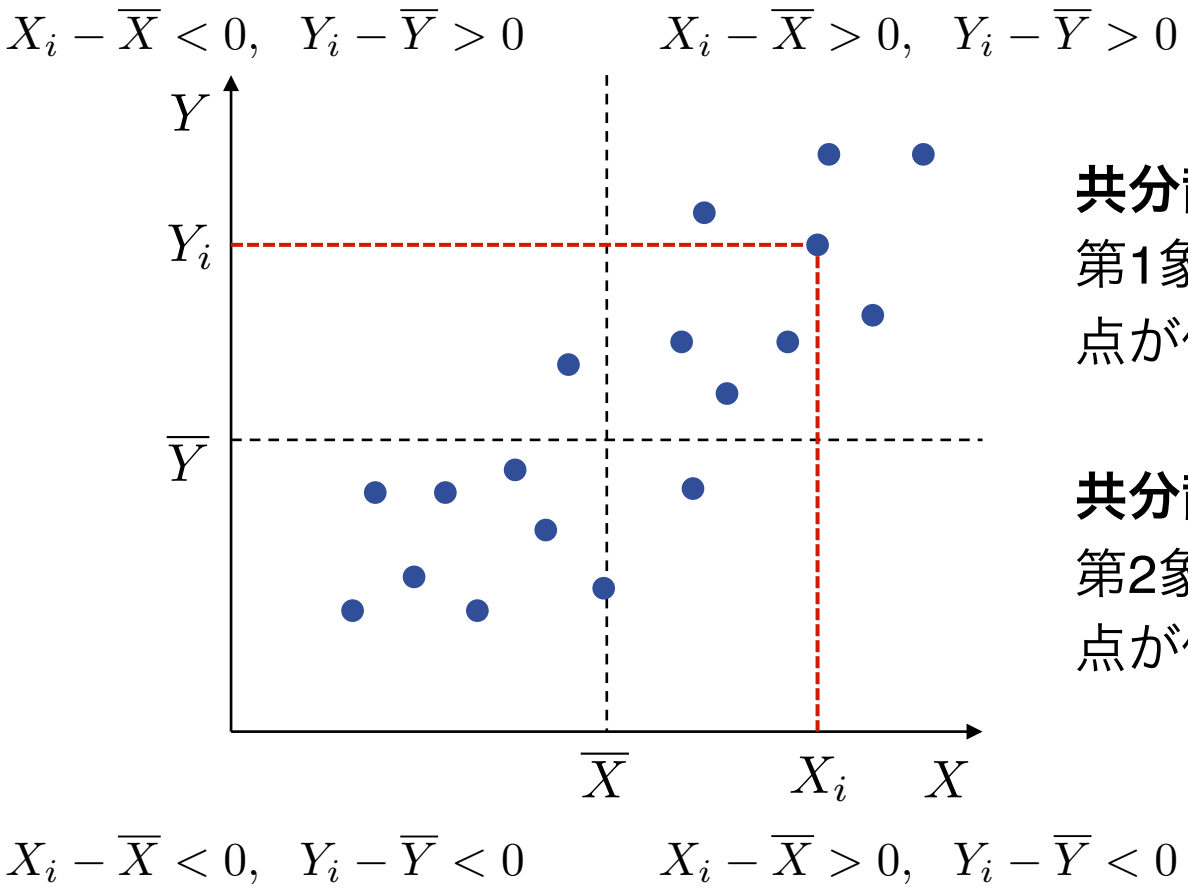
Pearsonの積率相関係数は以下のように定義される

$$\mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Sd}(X)\mathrm{Sd}(Y)} = \frac{\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i-\bar{Y})^2}}$$

-1（完全な負の相関）←0（相関なし）→+1（完全な正の相関）

散布図と相関係数の関係(1)

図　正の相関を示す散布図の例



共分散が正の場合

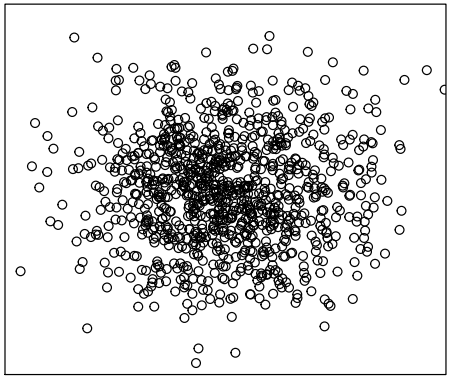
第1象限または第3象限に点が位置する

共分散が負の場合

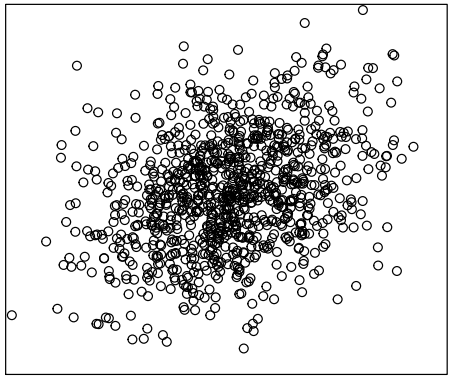
第2象限または第4象限に点が位置する

散布図と相関係数の関係(2)

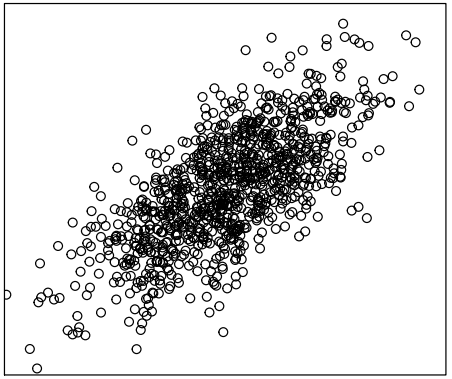
Cor ≒ 0



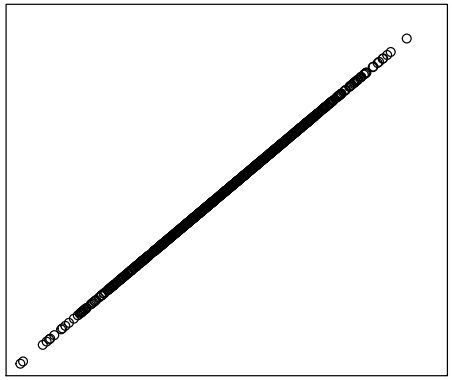
Cor ≒ + 0.3



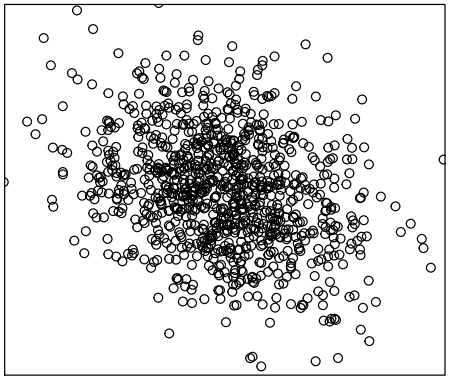
Cor ≒ + 0.7



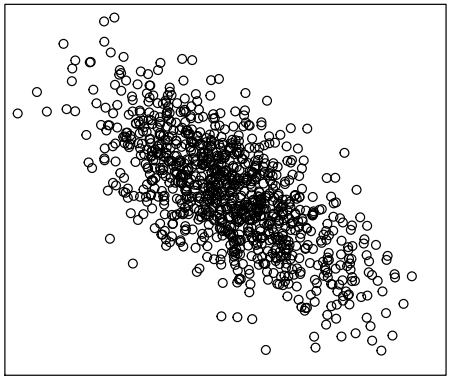
Cor ≒ + 1



Cor ≒ - 0.3



Cor ≒ - 0.7



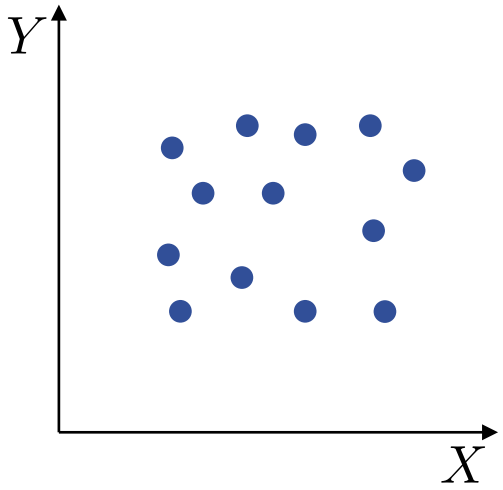
相関係数の独立性の検定

以下のt値を用いて検定。

$$t = \frac{\mathrm{Cor}(X,Y)\sqrt{n-2}}{\sqrt{1-\mathrm{Cor}(X,Y)^2}}$$

Tが十分に大きい→H_0 (母集団において2変数が無相関) を棄却

図　2変数が無相関の場合

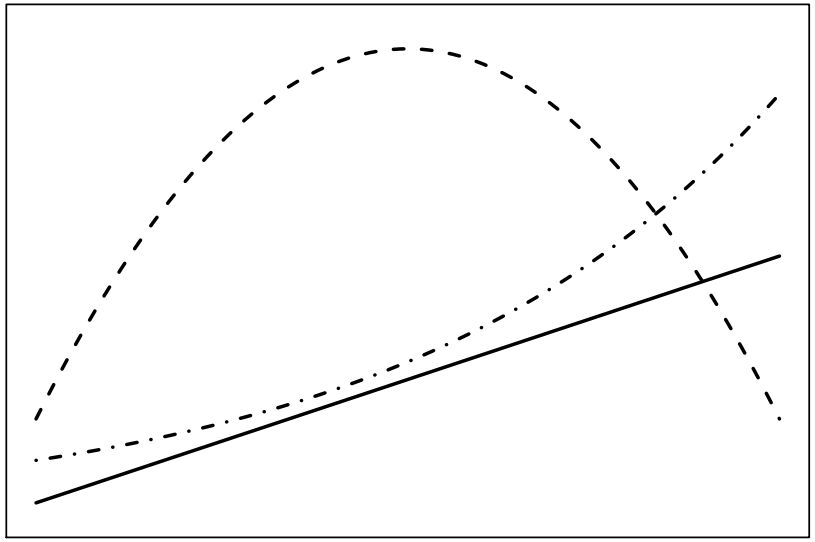


相関係数を使うときのポイント

- **それぞれの変数が適度にばらつき、かつ不自然な値がないか？**

変数が極端な値に偏っていると、そのケースに値が大きく左右され、
実際の関連を誤って判断してしまう可能性→散布図を描いて確認

- **2変数の関連が直線的であるか？**



$$\log Y = \beta_0 + \beta_1 X$$

$$Y = \beta_0 + \beta_1 X$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2$$

相関係数は、上図の実線のような関連が想定される場合に有効

SCATTERPLOT｜散布図を描く

```
GRAPH
/SCATTERPLOT = x WITH y
.
```

/*補足*/

GRAPHというさまざまなグラフを描くコマンドのオプションとして散布図を指定する形になっている.

座標上で同じ位置にある点は被ってしまい、度数がわからなくなる点に注意.このようなときは、クロス表で度数を見てみるとよい.

yのあとに“BY z”と入れると、第3変数で層化した散布図を描くことができる.

CORRELATIONS｜相関係数の計算

```
CORRELATIONS x1 x2 …
/STATISTICS = DESCRIPTIVES
/MISSING = PAIRWISE
.
```

/*補足*/

2行目をオプションとして入れると各変数の平均・標準偏差・度数を出してくれる.

MISSINGは欠損値の処理方法を指定するオプション。PAIRWISEがデフォルトで、LISTWISEは指定した変数群のうち1つでも欠損値がある場合は除いて相関係数を計算する.

目次

多変量解析とは

社会科学においては実験によって条件の統制ができないことが多く、
変数間関係は通常2変数間の関連だけでは捉えきれないほどに複雑

↓

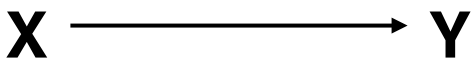
多変量解析

3つ以上 (場合によっては2つ以上) の変数を組み合わせる分析の総称

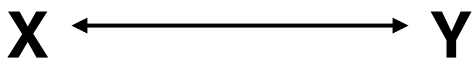
今回はとくに、2変量関連を拡張して、多変量解析で最もポピュラーな方法である回帰分析への導入を行う。

因果関係と共変関係

因果関係



共変関係



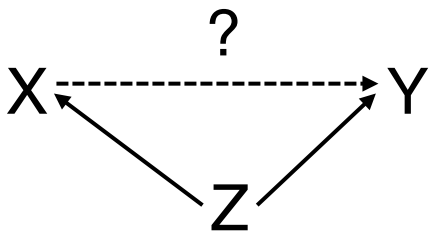
統計分析はそれ自体では因果関係と共変関係を区別することはできず、何らかの理論を準備する必要がある

Lazarsfeldの因果関係に関する3つの基準

1. XとYの間に時間的な先後関係がある（Yよりも先にXが起こる）
2. 2つの変数の関連が経験的に確認できる
3. 観察された2変数の関連が、XとYの両方をもたらす第3変数Zによっては説明されない

第3変数を統制する

第3変数を入れることで、2変数間の因果関係をより厳密に検証できる



第3変数Zを一定としてもなお、XとYの間に関連はあるか？

例）

X = 婚姻状態（既婚または未婚）、Y = 生活満足度、Z = 世帯年収

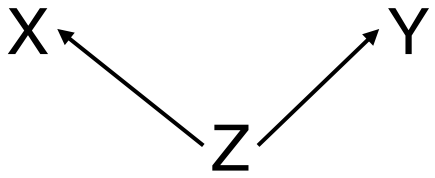
- 既婚者は未婚者よりも生活満足度が高い。
- 年収が高い者は結婚しやすく、かつ生活満足度が高い傾向がある。

→年収を一定としても、既婚者は生活満足度が高いのだろうか？

第3変数とのさまざまな関連

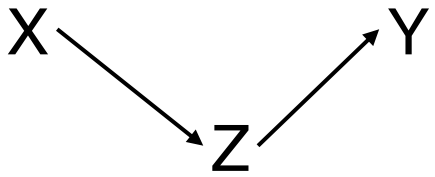
第3変数を入れることで、より洗練された問い・答えを得られる

擬似関係



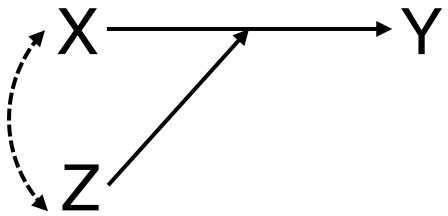
XとYの関連はZによって説明できる
例）いつも朝食を食べる小学生は成績が良いが、これは朝食を食べさせる親が教育熱心なためである。

媒介関係



XはZを通してZと関連している
例）裕福な家庭に生まれた子どもは高い学歴を得やすく、それゆえに高い収入を得られる。

交互作用関係



XとYの関連はZの有無により異なる
例）多くの友人をもつことは孤独感を弱めるが、その効果は同居家族がいない場合により強くなる。

*例はいずれも別の関係の可能性があり得るので注意。

線形モデルの考え方

変数間の関係を$Y = f(X)$という関数型で表現する方法を総称して、
一般化線形モデル (generalized linear model) という。

YとXの呼び方

Y	X
従属変数 dependent variable	独立変数 independent variable
被説明変数 explained variable	説明変数 explanatory variable

通常、Yが1つ、Xが複数あるものと想定される。

線形回帰分析(1)

線形モデルのなかでも最も基本的なモデルが**線形回帰分析** (linear regression model) であり、以下のように定義される。

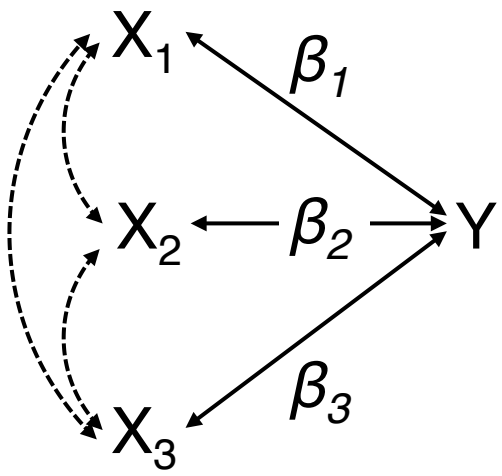
$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + e_i$$

切片 (intercept)　傾き (slope)　残差 (residual)
誤差 (error)
係数 (coefficient)

独立変数が1つだけの場合を単回帰分析、2つ以上の場合を重回帰分析と言ったりする。

線形回帰分析(2)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$



限界効果 (partial effect)

β_1は、他の変数X_2, X_3を一定としたうえで、X_1が1単位増加したときのYの増加量（X_1がYに与える直接効果）を表す。

- 擬似関係をもたらすような変数X_2, X_3を統制する
- 独立変数間の効果の大きさを比較する

といった使い方ができるが、それ自体が直接に変数間の因果関係を検証する方法ではないことに注意

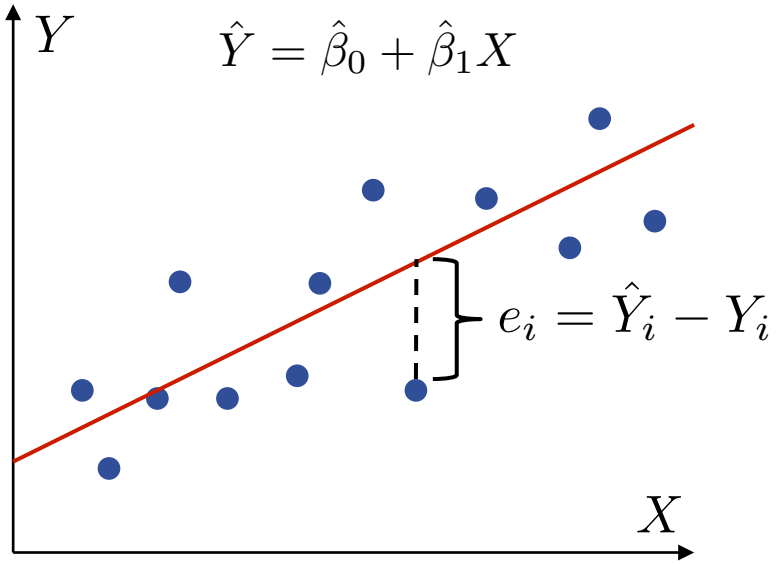
最小二乗法｜係数の推定のしかた

最小二乗法 (ordinary least square, OLS)

残差の2乗和を最小とするように係数 $\beta_0, \beta_1, \cdots, \beta_k$ を決める

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \cdots - \hat{\beta}_k X_{ik})^2$$

$\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$について上式を微分し、それぞれ = 0 となるようにk+1個の連立方程式を作り (OLSの一階の条件)、これを解く

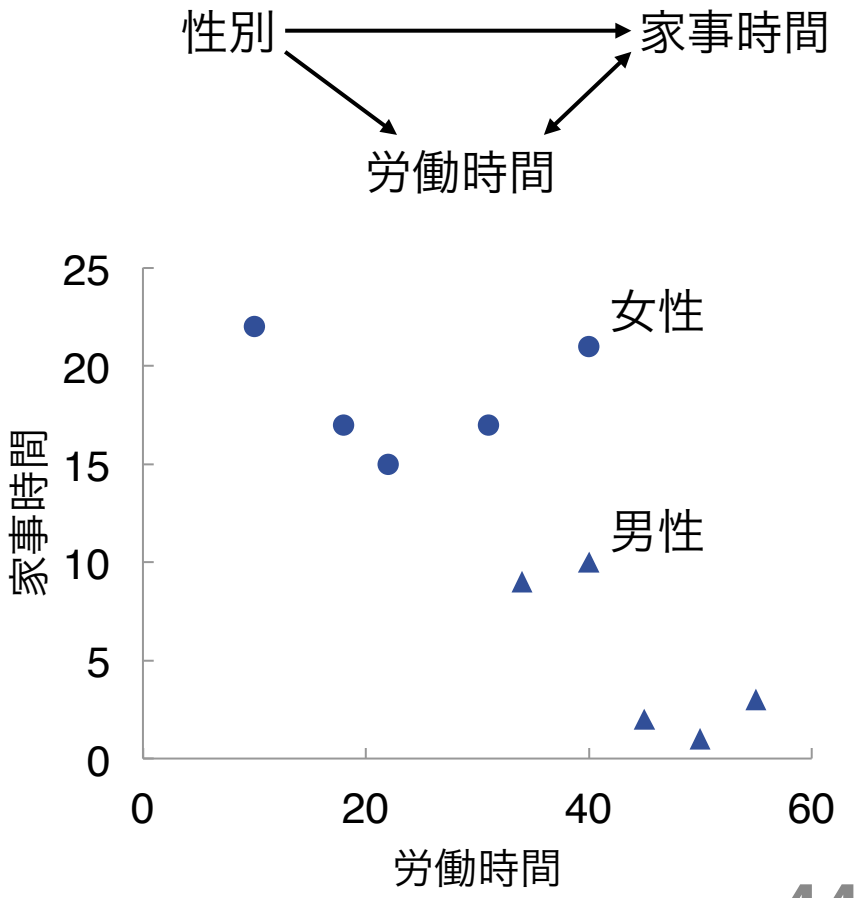


他の変数を一定とすることの意味

問い｜労働時間を一定としても女性の家事時間は長いのだろうか？

架空データ：性別と週あたり時間

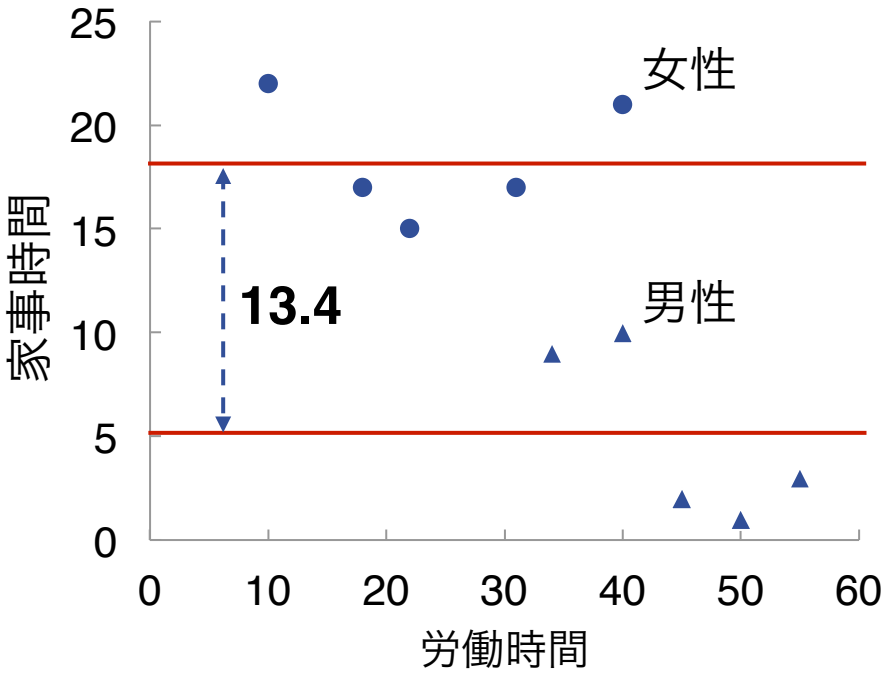
ID	性別	労働時間	家事時間
1	男性	50	1
2	女性	40	21
3	女性	22	15
4	男性	45	2
5	女性	31	17
6	男性	40	10
7	男性	55	3
8	女性	10	22
9	男性	34	9
10	女性	18	17



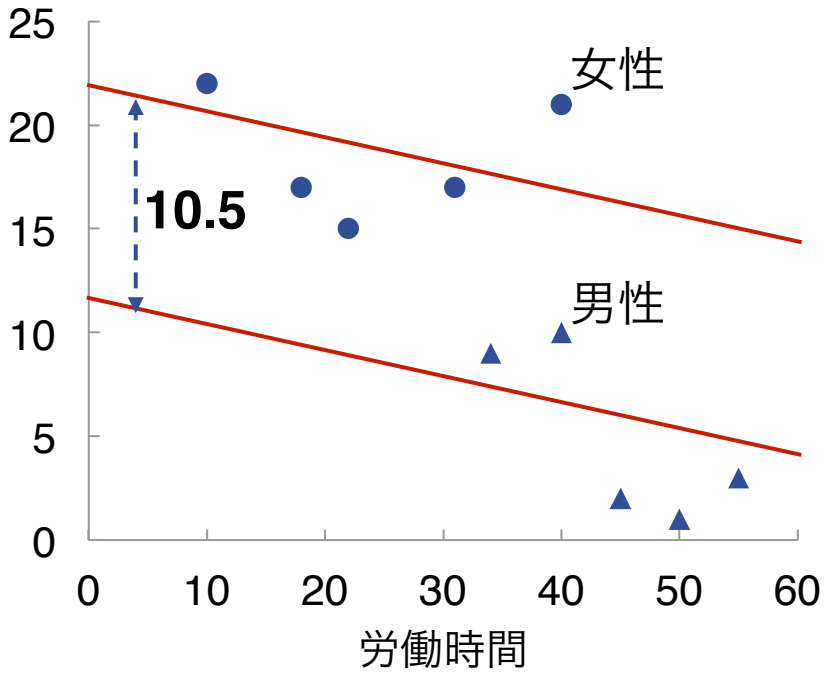
回帰式の当てはめ

重回帰分析のイメージ ＝「労働時間が増えるほど家事時間は減る」という傾向を考慮したうえで男女の平均値を比較する

$$\mathrm{HW} = 5 + 13.4 \times \mathrm{Female}$$



$$\mathrm{HW} = 11.2 + 10.5 \times \mathrm{Female} - 0.138 \times \mathrm{Work}$$



例からわかること

1つめの式から得られる結論

女性は男性と比較して家事時間が長い。

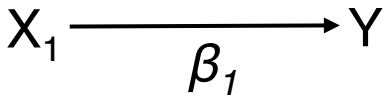
2つめの式から得られる結論

労働時間が長いほど家事時間が減るという傾向を考慮しても、女性は男性と比較して家事時間が長い。

- 回帰分析は、平均値の比較をより一般的な形に拡張したものと考えることができる。
- 回帰分析を適切に使うことで、議論の説得力や自分の問いをより洗練させることにつながる（統制変数の選択も重要）。
- 一方で、適当にいろいろ変数を入れて「Aが有意だった、Bは...」というような「とりあえず回帰分析」から得られるものは乏しい。

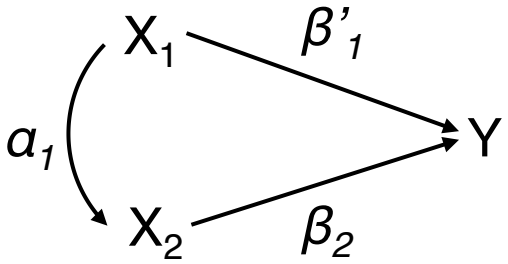
重回帰分析における係数の意味

単回帰分析 $Y = \beta_0 + \beta_1 X_1$ の場合



$$\hat{\beta}_1 = \sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}) / \sum_{i=1}^{N}(x_i - \overline{x})^2$$

重回帰分析 $Y = \beta_0' + \beta_1' X_1 + \beta_2 X_2$ の場合



1. $X_1 = \alpha_0 + \alpha_1 X_2$ を推定して残差の予測値 $\hat{r}_i$ を得る。
2. Y を $\hat{r}_i$ で回帰すると β_1' が得られる。

$$\hat{\beta}_1' = \sum_{i=1}^{N}\hat{r}_i(y_i - \overline{y}) / \sum_{i=1}^{N}\hat{r}_i^2$$

以上2つの回帰分析において、以下の関係が成り立つ。

$$\beta_1 = \beta_1' + \alpha_1 \times \beta_2$$

決定係数｜モデルの説明力を評価する

Yの総変動 (total sum of squares) は、回帰式により予測されるYの変動 (explained sum of squares) と、予測されないYの変動 (residual sum of squares) に分解される。→cf. 総分散の定理

$$\sum_{i=1}^{N}(Y_i - \overline{Y})^2 = \sum_{i=1}^{N}(\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{N}\hat{e}_i^2$$

$$\mathrm{SST} = \mathrm{SSE} + \mathrm{SSR}$$

R2乗値、決定係数 (R-squared, coefficient of determination)

Yの総変動のうち、回帰式により予測されるYの変動の占める割合。

$$R^2 \equiv \frac{\mathrm{SSE}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSR}}{\mathrm{SST}}$$

なお単回帰分析におけるR2乗値は相関係数の2乗に一致する。

Gauss-Markovの仮定

以下の5つの仮定が満たされるとき、$\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$ は $\beta_0, \beta_1, \cdots, \beta_k$ の最良線形不偏推定量 (best linear unbiased estimators, BLUEs)となる。

1. サンプルだけでなく母集団においても従属変数と独立変数は線形の関係にある。
2. サンプルが母集団から無作為に抽出されている。
3. 説明変数の分散が0でない。
4. 説明変数を条件づけたときの残差の期待値が0である。
5. 説明変数の値によらず分散の大きさは一定 (等分散性)。

4と5を合わせると、**残差が正規分布にしたがう**ことを意味する。

$$e_i | X_1, \cdots, X_K \sim N(0, \sigma^2)$$

標準誤差と検定

標準誤差（standard error）

サンプルから推定された係数がばらつく程度（係数の標準偏差）

$$\mathrm{Se}(\hat{\beta}_j) = \sqrt{\frac{1}{N-k-1}\sum_{i=1}^{N}\hat{e}_i^2 \Big/ \sum_{i=1}^{N}(X_{ij}-\overline{X}_j)^2}$$

t値 $t = \hat{\beta}_j / \mathrm{Se}(\hat{\beta}_j)$ が十分に大きい→H_0（母集団において係数β_jの値が0でない）を棄却

標準誤差は、サンプルサイズ（N）が大きく、残差が小さく、独立変数X_jの分散が大きいほど、小さくなる

REGRESSION｜回帰分析

```
REGRESSION
/DEPENDENT = y /*従属変数を指定*/
/METHOD = ENTER x1 x2 /*独立変数を指定*/
/DESCRIPTIVES /*記述統計量を出力*/
.
```

/*補足*/

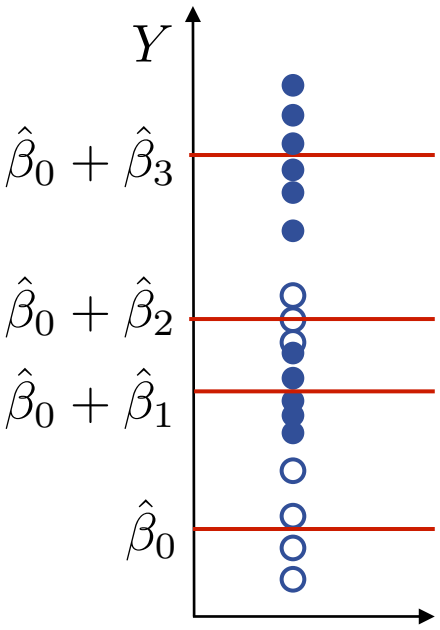
サブコマンドのMETHOD = ENTERの部分は複数指定することができ，それ以前の独立変数にさらに変数を追加するモデルを一度に推定できる．

どれか1つでも変数が欠損しているケースは推定に使用されない(このことをリストワイズlist-wise削除という)．

カテゴリカル変数を独立変数にする

カテゴリカル変数を独立変数とする場合は、ある値を取る場合に1、
そうでない場合に0となるような変数 (ダミー変数) を作成する。
ダミー変数を入れる際には、どこか1つのカテゴリを基準とする。

例）学歴4分類を独立変数とする回帰分析



$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{High} + \hat{\beta}_2 \mathrm{JuniCol} + \hat{\beta}_3 \mathrm{University}$$

学歴	中学	高校	短大高専	大学
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1

ダミー変数と参照カテゴリ

参照カテゴリ (reference category)

ダミー変数を用いる際にモデルに投入しないカテゴリのこと。投入しないカテゴリが、カテゴリ間の比較の参照点となる。

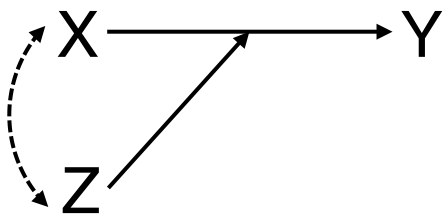
ダミー変数と参照カテゴリの選び方のガイドライン

- 主張したいことに対応させる　　例）「女性は家事時間が長い」→女性ダミー。「男性は家事時間が短い」→男性ダミー。
- 参照カテゴリは明確なほうがよい
  「その他」や「分からない」との比較は難しい
- 参照カテゴリは特殊すぎず、ある程度該当ケースも多いほうがよい
- カテゴリ間に順序が想定される場合には最も端orまんなかを参照カテゴリとする

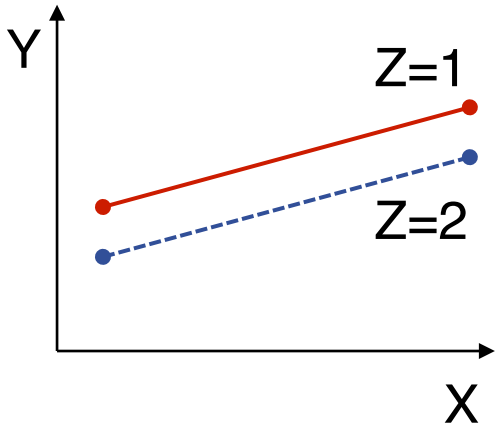
交互作用とは

交互作用効果（interaction）

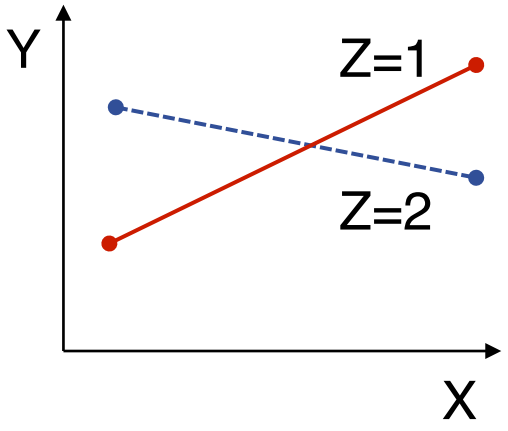
ある変数の効果が、別の変数の有無や程度によって変化すること



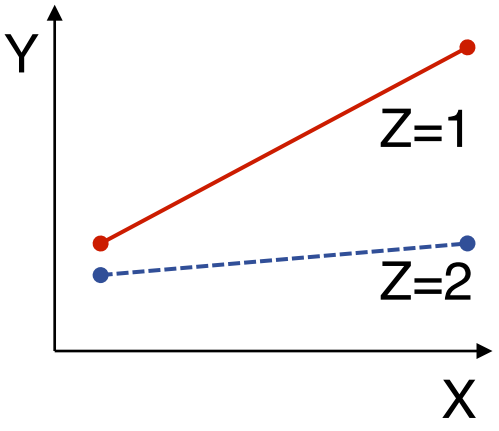
交互作用がない



交互作用がある(1)



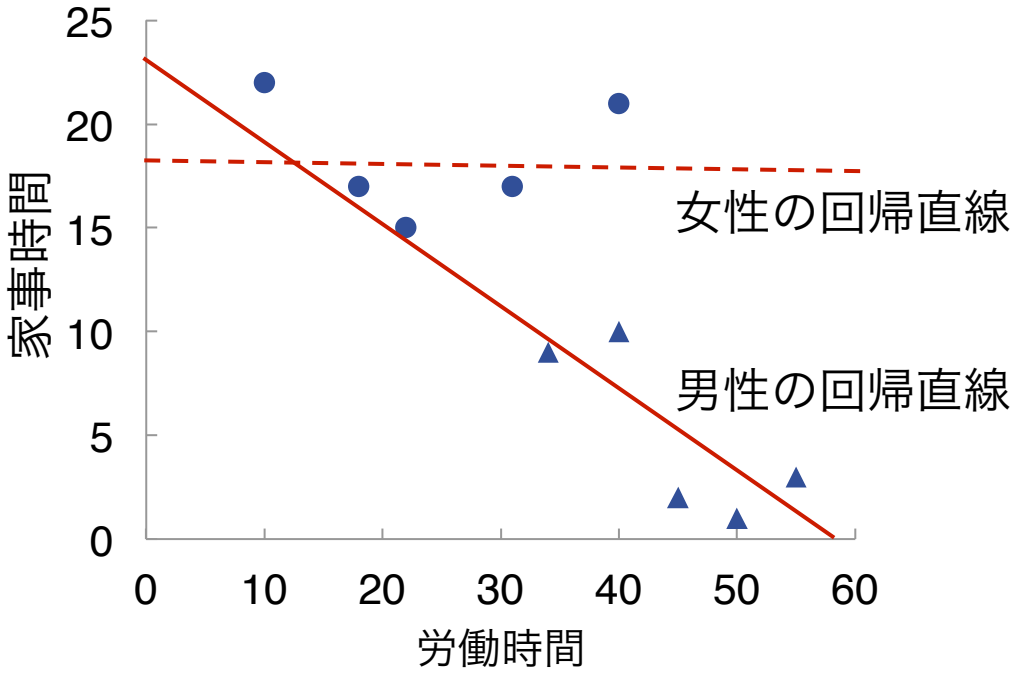
交互作用がある(2)



交互作用関係を検討する

問い｜労働時間が家事時間に与える影響は男女によって異なるか？

$$\mathrm{HW} = 23.0 - 4.5\mathrm{Female} - 0.403\mathrm{Work} + 0.396\mathrm{Female} \times \mathrm{Work}$$



男性は労働時間が多いほど家事時間が少なくなる傾向があるが、女性は労働時間と家事時間の間にほとんど関連がみられない。

交互作用項の作り方と解釈

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

主効果 (main effect)　　交互作用効果 (interaction)

交互作用を表す変数は2つの変数をかけ合わせて作る →cf. compute

β_1は「**$X_2 = 0$のもとで**、X1が1単位増加したときのYの増加量」を表す。

- 交互作用項を入れると主効果の係数がしばしば大きく変わったりするのはそのため。
- 交互作用項を入れる前のモデルと、入れた後のモデルを比較しながら係数の意味を考えるのがよい。

カテゴリカル変数を従属変数にする

カテゴリカル変数への効果を見たいときはどうするか？

例）夫の年収が高いほど妻は就業しにくいのだろうか？

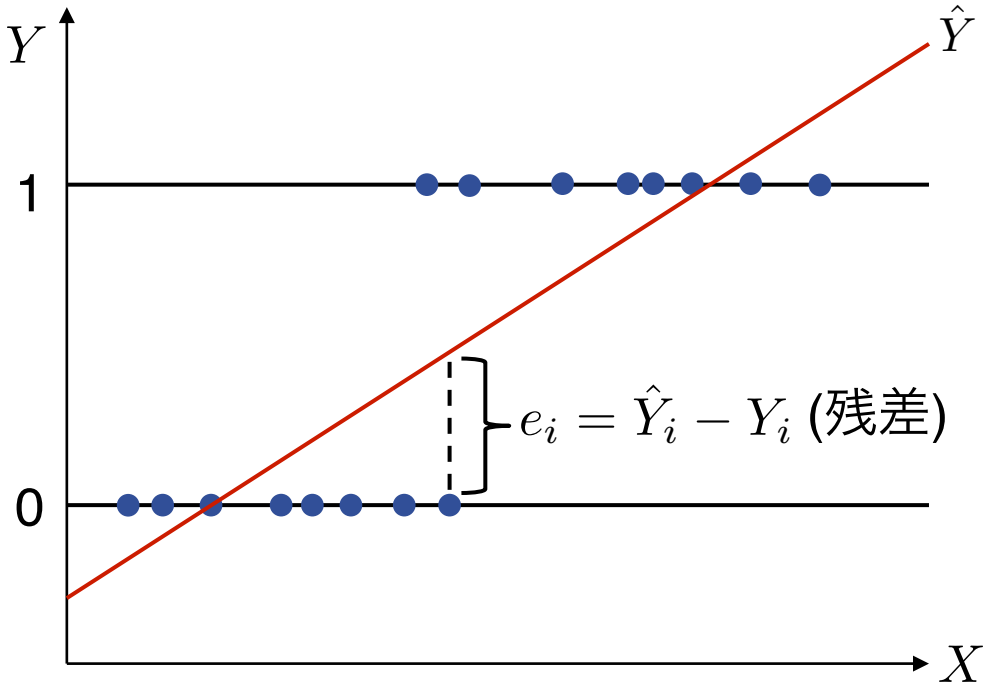
↓

一番素直な拡張は、これまでの回帰分析の従属変数を2値変数に置き換えることで対応する

2値変数 (0 or 1)を従属変数とする線形回帰分析をとくに**線形確率モデル** (linear probability model) という。

線形確率モデルの解釈

係数は独立変数が1単位増加したときのYの確率(比率)の増加量を表す



$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

Xが1単位増加→
Yが1をとる確率が
β_1だけ増加

残差の等分散性の仮定への違反や予測値の一部が0または1を超える値となりうるといった問題があるが、解釈はしやすい

回帰分析の結果のまとめかた(一例)

表　孤独感の規定要因に関する回帰分析

	model 1		model 2	
	B	(SE)	B	(SE)
独居(ref: 非独居)	.549**	(.150)	-.025	(.417)
挨拶人数	-.147**	(.040)	-.225**	(.066)
独居×挨拶人数			.120	(.081)
男性(ref: 女性)	.129	(.130)	.128	(.129)
居住年数	.009*	(.005)	.009†	(.005)
年齢	.002	(.011)	.000	(.011)
切片	2.314**	(.806)	2.808**	(.872)
N	385		385	
R^2	.064		.069	

注）** $p < .01$, * $p < .05$, † $p < .1$

出所）「A団地のくらしと地域づくりに関するアンケート」

結果をまとめるときのポイント

- ダミー変数については参照カテゴリを示しておくと結果が読みやすい
- 絶対値1未満の値が多く出てくる場合は0を省略してもよい
- 複数のモデルを立てる場合はモデル間でサンプルサイズ（N）を揃える
  →cf. FILTER または SELECT IF
- 有意水準については自分で定めて、有意確率をもとに印をつけるのが一般的（有意確率そのものは載せない）
- 係数と標準誤差については必ず欲しい
- モデルの統計量についてはNとR2乗値は最低限欲しい。そのほかは関心に応じて入れる
- 回帰分析を行う場合は、用いる変数の記述統計量（平均および標準偏差）を変数の説明などの節に載せておく

回帰分析etcをする際にありがちなミス

- **ダミー変数のコーディングを間違える**
  例）男性 = 0、女性 = 1とするべきところを男性 = 1、女性 = 2のまま性別変数として投入してしまう
- **外れ値を含んだ連続変数を入れてしまう**
  例）年齢を示す変数の無回答（999）を欠損値指定しないで投入してしまう
- **変数の数値が意図した並びとは逆になっている**
  例）孤独感を感じるほど高い値、感じないほど低い値にするはずが逆になっており、まったく逆の解釈をしてしまう

おまけ｜カテゴリカル変数・制限従属変数を従属変数とする回帰分析

カテゴリカル変数 (正規分布に従わない連続変数を含む) の分析

- ロジットモデル・プロビットモデル
- 順序ロジット、多項ロジットモデル
- 対数線形モデル (ログリニアモデル)
- ポワソン回帰・負の二項回帰

制限従属変数 (ある値を閾値としてデータが欠損している場合)の分析

- トービット・モデル
- Heckmanのサンプル・セレクションモデル