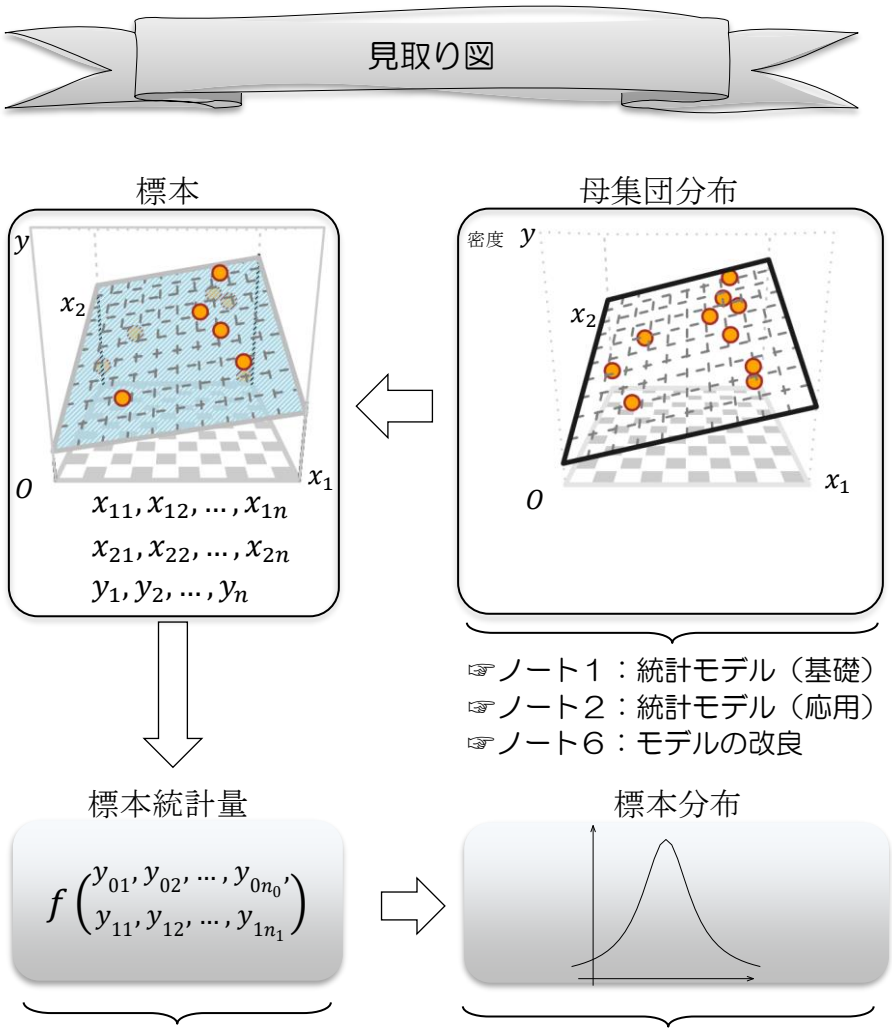
第5講　重回帰分析

学習の目標

- □ 回帰モデルにおいて独立変数の数が複数含まれるものを**重回帰モデル**ということが分かる。
- □ 回帰モデルを**数式や図によって表現**することができる。
- □ 研究者の興味や仮説に応じて母集団において柔軟なモデルを立てる意義ややり方についての基礎が理解できる。
- □ ある独立変数が別の変数を経由して従属変数に**間接効果**を持つとき仲介をなす変数を**媒介変数**ということが分かる。
- □ 複数の要因が連動して変化するためどちらに影響したか判断できない状況を**交絡**と呼ぶことが分かる。
- □ **調整変数**の存在によって独立変数単体の**主効果**では説明できない**交互作用効果**が生じることがあることが分かる。
- □ 独立性の仮定を乱すクラスターを持つデータ構造に対して**階層モデル**を想定する必要性が理解できる。
- □ 重回帰モデルの各独立変数にかかる係数を**偏回帰係数**と呼び、その解釈の仕方を理解することができる。
- □ 独立変数にかかる係数の大きさを比較する手段として**標準偏回帰係数**を用いることの意義が分かる。
- □ 推定した回帰モデルのデータへの適合度を測る基準として**重相関係数**と**決定係数**を用いることができる。
- □ **重回帰式の検定**および各**偏回帰係数の検定**、そして、**信頼区間**や**予測区間**についても理解をし、使用できる。
- □ **過学習**、**解釈可能性**、**多重共線性**などの問題が懸念される際に、独立変数を選択しモデルの改良を行うことができる。
- □ **罰則を設けた最小二乗推定**について理解することができ、**ラッソ回帰**を用いた**モデル縮約**を行うことができる。
- □ **モデル比較**の必要性を理解し、**情報量基準**や**クロスバリデーション**を用いてモデルを選択することができる。
- □ 多重共線性に解決の一つの対策として、**主成分分析**を援用し独立変数間の相関をゼロにしたモデルを提案できる。



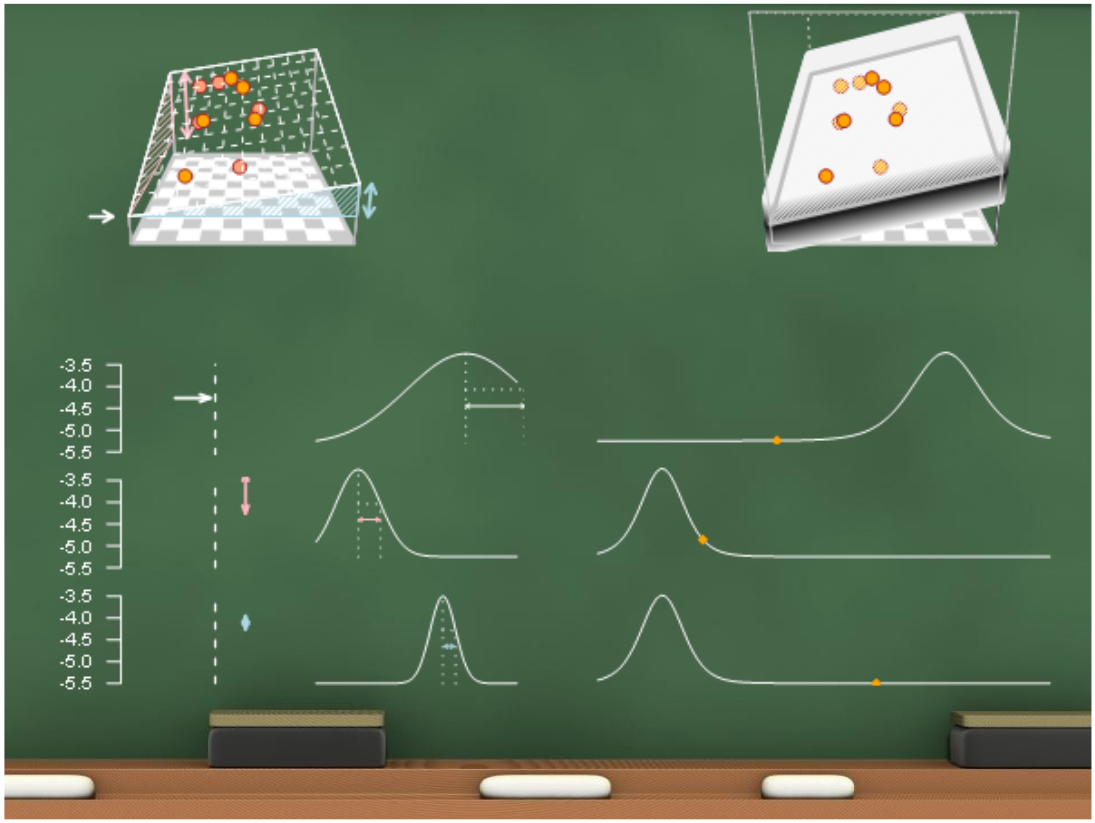
データの形式

ID	予測変数１	予測変数２	…	予測変数ｐ	応答変数
1	0.3	0	…	-4	2.1
2	0.1	1	…	-10	3.2
⋮	⋮	⋮	⋮	⋮	⋮
$\boldsymbol{n}$	1.2	0	…	29	1.5

（1）目的（リサーチクエスチョン）

複数の独立変数に基づいて従属変数の値を予測するモデルを作り、他の変数の影響が統制された下で各変数の効果量がどのくらいかを議論する。

（2）考え方



独立変数が一つしかモデルに登場していない単回帰分析を発展させて、独立変数の数を複数に増やした回帰分析を重回帰分析と呼ぶ。

このため、独立変数を表す軸が増加する。黒板では独立変数が二つの場合にどうなるのかを示している。すると、“東西方向”と“南北方向”のように二つの方向でそれぞれどのくらい傾いているのかを考えることになり、単回帰のような直線ではなく、平面を取り扱うことになるのが分かるだろう。この平面は黒板の白い矢印（切片）と二つの傾き（ピンクと水色の矢印）で特定出来るので、私たちはこれらの効果量の大きさを検討していくことになる。だが、検討の仕方に新しい方法があるわけではなく、それぞれの回帰係数に t 検定を行ったり、信頼区間や予測区間を検討する。

だが、単回帰と違い、独立変数を複数取り入れることで、かえって注意を向けないといけない問題も存在する。回帰係数の解釈やモデルの比較などの実践的な注意事項をぜひしっかり学んでおこう！

📖 ノート１ 母集団に対する仮定：基本的な統計モデル

（１）統計モデル

母集団に対する統計モデルを少しずつ複雑なものにしていくことで、細やかなモデルを提案することができる。

例１：二群の差の検定（対応なしの t 検定）

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

例２：単回帰分析

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

例３：重回帰分析

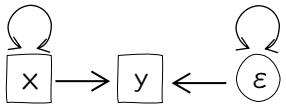
$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{3i} + \beta_2 x_{3i} + \cdots + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

（2）図による表現１：パスモデルとパス解析
これは、変数の関係を中心に発想する視覚的表現。

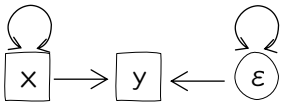
例１：二群の差の検定（対応なしの t 検定）

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$



例２：単回帰分析

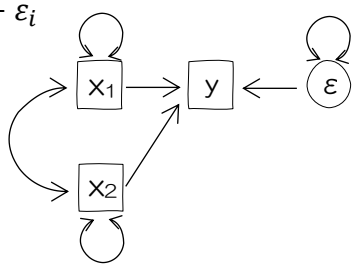
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$



例３：重回帰分析

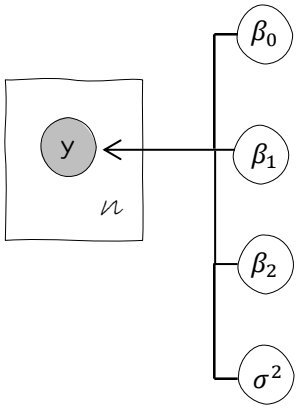
$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

	x_1	x_2
x_1	r_{11}	r_{12}
x_2	r_{21}	r_{22}



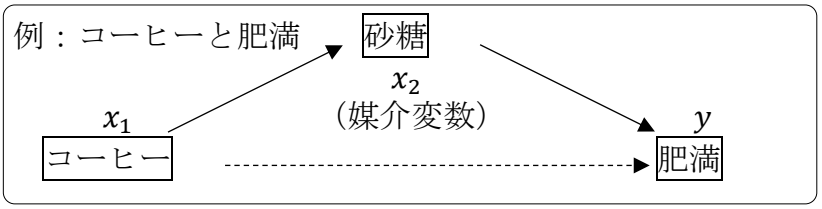
（3）図による表現２：グラフィカルモデリング
これは、パラメータの関係を中心に発想する視覚的表現。

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$



🕮 ノート２ †母集団に対する仮定：発展的なモデル

（１）媒介モデル Mediation Model



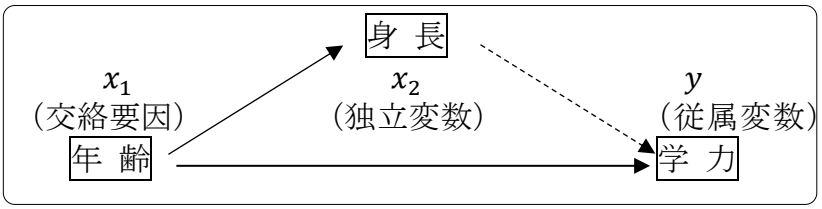
① 直接効果 Direct effects
独立変数から従属変数への直接的な影響の指標。

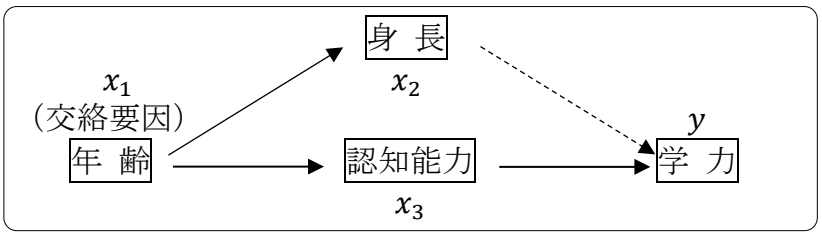
② 間接効果 Indirect effects
ある独立変数から別の変数（**媒介変数／中間変数**）を経由して従属変数に伝わる影響のこと。

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

$$x_{2i} = \gamma_1 x_{1i} + u_i$$

（２）交絡要因／共通因子を持つモデル　　（☞ 多重共線性）



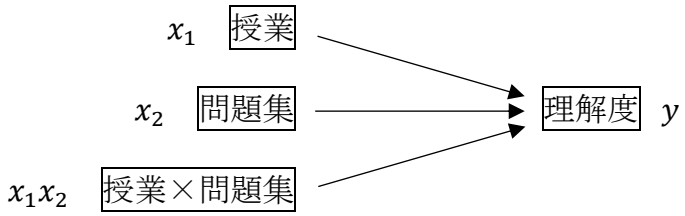


① 交絡 Confounding
二つ以上の要因が連動して変化するため、そのうちどれが結果に影響したかが判断できない状態になること。

② 交絡変数 Confounding variables
独立変数x_2の上流側にある、独立変数x_2と目的変数yの両者に影響をもたらす要因のこと。

（3）交互作用モデル　Interaction Model

例：授業と理解度（問題集が調整変数）



① 調整（抑制）変数 moderator (suppressor) variable
x_1とyの関係が第三の変数x_2の値によって変化するとき、このx_1と交互作用を持つx_2を調整変数という。

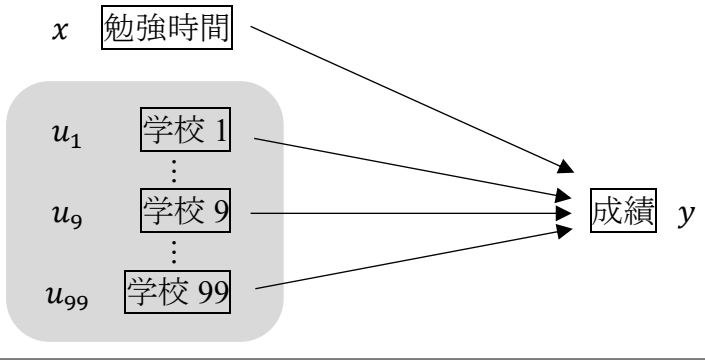
② 交互作用 Interaction
二つ以上の変数が組み合わさって生まれるもともとの変数単体の効果（主効果）では説明できない効果のこと。

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

$$\beta_2 = \gamma_0 + \gamma_1 x_{1i}$$

（4）階層モデル Hierarchical Model

例：学校というグループ要因を考察する事例



$$y_{ij} = \beta_{0j} + \beta_1 x_i + \varepsilon_i$$

$$\beta_{0j} = \gamma_0 + u_j$$

📖 ノート３ 統計量：偏回帰係数の解釈

（解釈のコツ１）係数の比較（標準偏回帰係数の利用）
（解釈のコツ２）点推定された偏回帰係数の意味

（1）解釈のコツ１：単位の影響を除く！

データ１

X1	X2	X3	Y
18.77	37	244	1.62
17.13	38	257	1.68
⋮	⋮	⋮	⋮
23.38	41	304	1.46

データ２

X1	X2	X3	Y
-1.21	-1.18	-1.67	1.62
-1.49	-1.12	-1.59	1.68
⋮	⋮	⋮	⋮
-0.42	-0.94	-1.29	1.46

① 非標準偏回帰係数（データ１を分析）
生データの単位をそのまま利用した回帰分析の結果。

	推定値	標準誤差	t 値	p 値	
$\hat{\beta}_0$	0.565	0.233	2.428	0.017	*
$\hat{\beta}_1$	0.044	0.012	3.781	0.000	***
$\hat{\beta}_2$	-0.004	0.004	-0.894	0.374	
$\hat{\beta}_3$	0.002	0.000	5.899	0.000	***

② 標準偏回帰係数（データ２を分析）
データを標準化したうえで行った回帰分析の結果。
※ 標準偏差の比が 1 となり、相関係数のみに基づく。

	推定値	標準誤差	t 値	p 値	
$\hat{\beta}_0$	2.709	0.05	54.15	2e-16	***
$\hat{\beta}_1$	0.259	0.07	3.78	0.000	***
$\hat{\beta}_2$	-0.064	0.07	-0.89	0.374	
$\hat{\beta}_3$	0.385	0.07	5.90	5.23e-08	***

（2）解釈のコツ２：偏回帰係数はモデル相対的！

① 解釈上の注意点

同じデータを用いても、ある条件を満たさない限り、重回帰の回帰係数は、単回帰の回帰係数とは一致しない

単回帰分析　$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + e_i$

重回帰分析　$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + e_i$

② 理由

同じモデルに含まれているそのほかの要因（変数）に相対的に値が決まるため。

③ 解釈

他の独立変数の影響が取り除かれた時、その独立変数が一単位増加すると従属変数がどの程度変化するのか。

④ 例外

独立変数同士に、関連性がない（無相関である）場合は、単回帰分析の結果と一致する。

※ただし、検定や信頼区間の結果は異なる。

例１：現地調査

ID	背丈 y	気温 x1	湿度 x2
1	10	20	40
2	11	18	50
3	11	19	50
4	12	22	65
5	13	26	60
6	15	28	70
7	14	27	80
8	20	35	85

① 単回帰分析　$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + e_i$

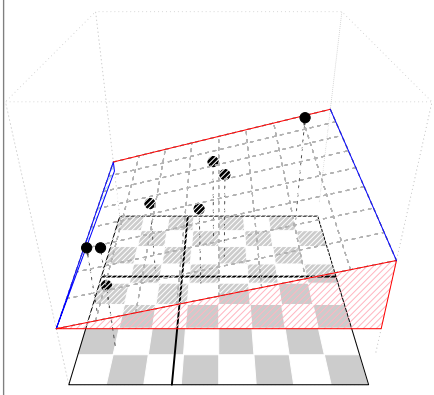
	推定値	標準誤差	t値	p値	
beta0	0.181	1.564	0.116	0.912	
beta1	0.536	0.062	8.554	0.000	***

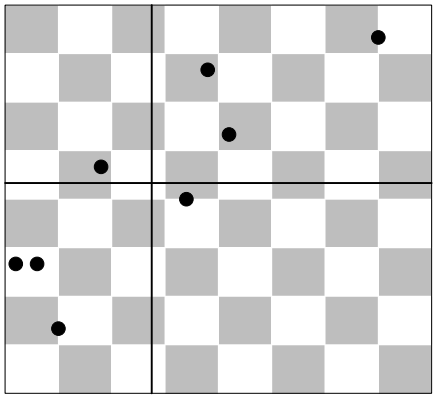
② 単回帰分析　$y_i = \hat{\beta}_0 + \hat{\beta}_2 x_{2i} + e_i$

	推定値	標準誤差	t値	p値	
beta0	2.037	2.603	0.782	0.464	
beta2	0.179	0.041	4.423	0.005	**

③ 重回帰分析　$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + e_i$

	推定値	標準誤差	t値	p値	
beta0	0.053	1.676	0.031	0.976	
beta1	0.467	0.138	3.390	0.020	*
beta2	0.029	0.050	0.571	0.593	





例２：温室での調査

ID	背丈 y	気温 x1	湿度 x2
1	10	20	40
2	11	20	80
3	11	25	40
4	12	25	80
5	13	30	40
6	15	30	80
7	14	35	40
8	20	35	80

① 単回帰分析　$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + e_i$

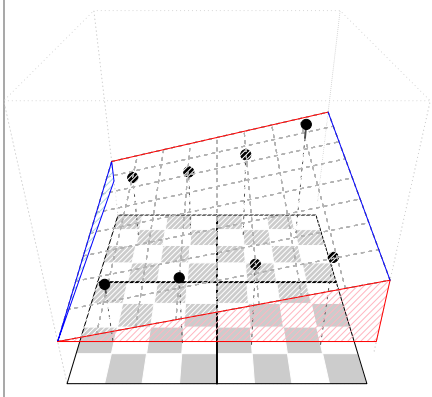
	推定値	標準誤差	t値	p値
beta0	1.150	3.482	0.330	0.753
beta1	0.440	0.124	3.546	0.012 *

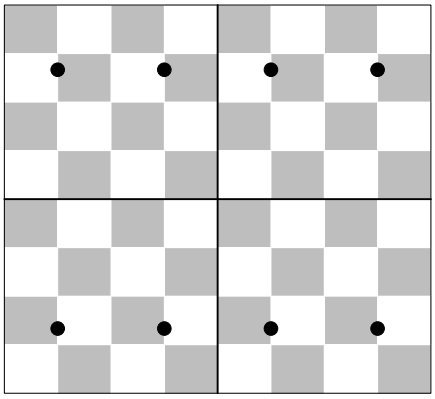
② 単回帰分析　$y_i = \hat{\beta}_0 + \hat{\beta}_2 x_{2i} + e_i$

	推定値	標準誤差	t値	p値
beta0	9.500	3.506	2.710	0.035 *
beta2	0.063	0.055	1.127	0.303

③ 重回帰分析　$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + e_i$

	推定値	標準誤差	t値	p値
beta0	-2.600	3.010	-0.864	0.427
beta1	0.440	0.092	4.778	0.005 **
beta2	0.063	0.026	2.428	0.060 .



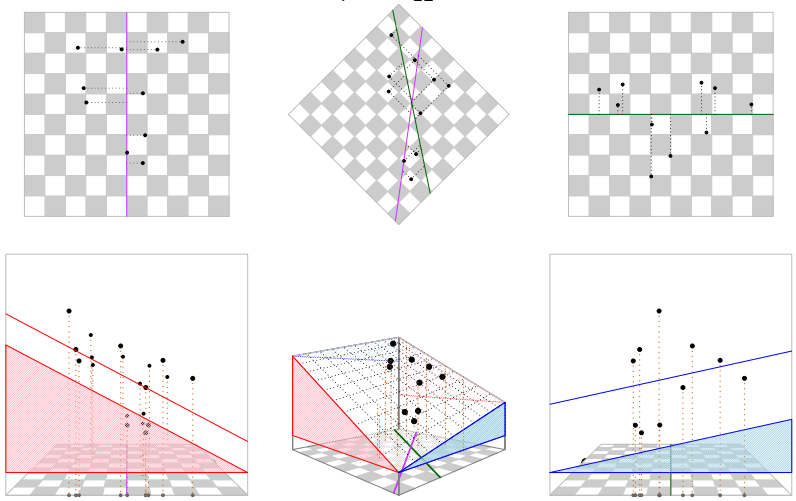


（３）†偏回帰係数と部分相関係数/偏相関係数の関係

① 部分相関係数 Part correlation coefficient

一つの変数から第三の変数の影響を除いた後で、二つの変数の相関を求めたもの。

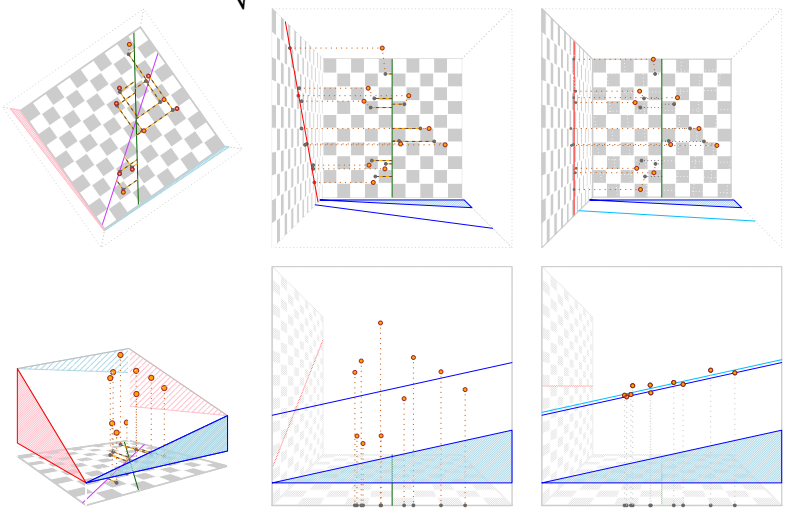
$$r_{y(2|1)} = \frac{r_{y2} - r_{y1}r_{12}}{\sqrt{1 - r_{12}^2}}$$



② 偏相関係数 Partial correlation coefficient

二つの変数それぞれから第三の変数の影響を除いた後で、それら二つの変数の相関を求めたもの。

$$r_{y2|1} = \frac{1}{\sqrt{1 - r_{y1}^2}} \times \frac{r_{y2} - r_{y1}r_{12}}{\sqrt{1 - r_{12}^2}}$$



③ 偏回帰係数 Partial regression coefficient

ある独立変数からそれ以外の独立変数の影響を除いた残差変数によって従属変数を予測するときの回帰係数。

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

点推定値

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2$$

$$\hat{\beta}_1 = r_{y(1|2)} \times \frac{s_y}{s_{1|2}} \qquad s_{1|2} = s_1 \sqrt{1 - r_{21}^2}$$

$$\hat{\beta}_2 = r_{y(2|1)} \times \frac{s_y}{s_{2|1}} \qquad s_{2|1} = s_2 \sqrt{1 - r_{12}^2}$$

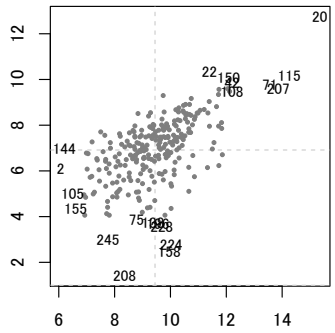
※ 偏回帰係数と部分／偏相関係数の関係

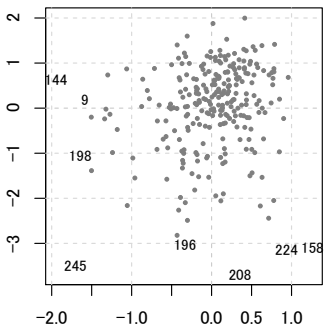
縦軸と横軸の(i) 相関係数$r_{y(1|2)}$と(ii)標準偏差の比$s_y/s_{1|2}$から構成されている（参照：第 4 講）

偏相関プロット

研究の動機：過去形と現在完了形の違いを調べたい。

	VERB	PRESENT	FUT (WILL)	**PAST**	**PRS PRF**	PST PRF	FUT PRF	PRS PROG	PST PROG	FUT PROG	TOTAL
A	BECOME	32,211	6,043	**89,105**	**31,547**	10,340	55	6,203	2,208	2	291,488
	EVOLVE	942	209	**1,813**	**2,222**	336	7	374	84	3	12,421
	CHANGE	14,003	2,531	**16,976**	**10,903**	3,303	38	3,090	768	22	139,217
	INCREASE	6,993	2,050	**6,883**	**4,962**	676	31	1,245	248	13	69,476
B	SCREAM	1,202	42	**4,212**	**4**	100	0	367	862	4	18,699
	NOD	1,229	21	**20,951**	**12**	79	0	75	219	1	30,482
	SMILE	2,498	60	**22,302**	**16**	165	0	421	1,212	11	45,951
	STARE	3,088	40	**15,996**	**35**	140	1	644	1,612	2	43,096
C	REMEMBER	48,731	970	**14,108**	**40**	242	2	141	224	5	121,157
	LIKE	71,705	385	**22,536**	**78**	259	0	28	18	0	207,626
	NEED	135,622	5,366	**42,551**	**165**	229	0	26	19	15	323,740
	Total	19,915,160	571,588	**14,262,925**	**966,351**	531,979	2,922	508,233	406,085	10,578	60,877,297





2: accumulate
9: affect
20: be
22: become
42: come
75: drink
105: fulfil
108: get
115: have
138: lean
144: limit
150: make
155: monitor
158: nod
196: remember
198: require
207: say
208: scream
224: smile
228: stare
245: thank

📖 ノート４　分析の評価１：作った統計モデルの正確さを評価

（参照：第４講ノート５）

データへの適合度

独立変数を複数取り入れたことで、従属変数の予測はどれくらい正確になったの？

（１）平方和／分散の直交分解

予測値$\hat{y}$と残差eは互いに無相関であるので、その和である$y = \hat{y} + e$の分散を次のように分解することができる。

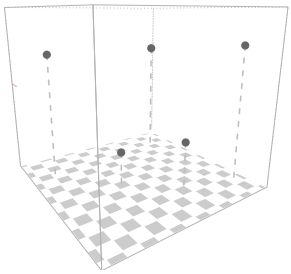
① 分散の直交分解

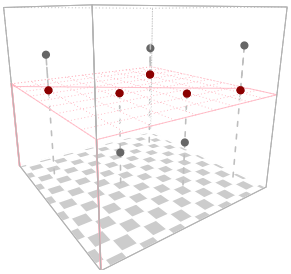
$$s_y^2 = s_{\hat{y}}^2 + s_e^2$$

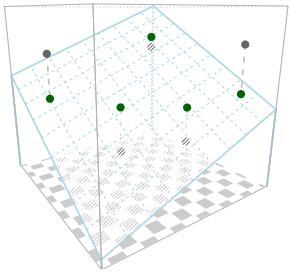
② 平方和の直交分解

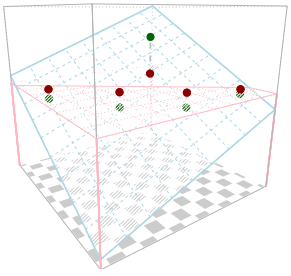
$$\frac{1}{n}ss_y = \frac{1}{n}ss_{\hat{y}} + \frac{1}{n}ss_e$$

$$ss_y = ss_{\hat{y}} + ss_e$$



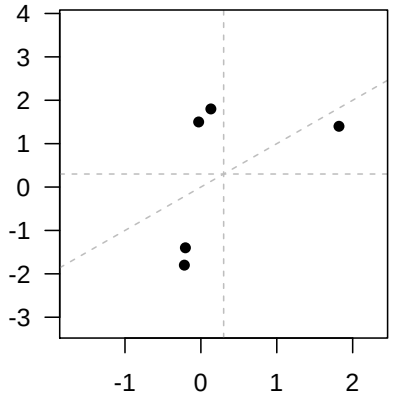






（２）重相関係数R Multiple correlation coefficient
これは、実測値yと 予測値$\hat{y}$の相関係数であり、Rと表す。

$$R = r_{y\hat{y}}$$



（３）決定係数 Coefficient of determination
これは独立変数がどれだけ従属変数の値を決定するかを示す指標。別名：分散説明率 Proportion of variance accounted for

$$R^2 = 1 - \frac{s_e^2}{s_y^2}$$

（４）決定係数の性質
独立変数を増やすと追加した独立変数が従属変数の予測に寄与しているかに関係なく決定係数は単調に増加する。

Model 1　$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$

Model 2　$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$

Model 3　$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$

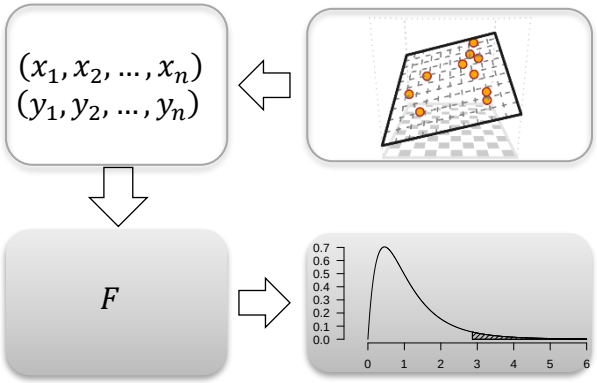
📖 ノート５ 分析の評価２：係数の推定値の正確さの評価

ステップ１　　重回帰式「全体」の検定
データから何らかの統計量を計算し、その標本分布を用いて「$\beta_1 = \beta_2 = \cdots = \beta_p = 0$（係数が全部 0）」という帰無仮説を検証したい。

ステップ２　　「個別」の回帰係数の検定
データから何らかの統計量を計算し、その標本分布を用いて「i 番目の偏回帰係数について$\beta_i = 0$だ」という帰無仮説を検証したい。

ステップ３　　重回帰式・偏回帰係数の区間推定
重回帰式やそれぞれの偏回帰係数について、信頼区間や予測区間を計算し、幅を持った推定を行う。

（１）検定１：重回帰式「全体」の検定　　ステップ1



① 帰無仮説と対立仮説

H0：「$\beta_1 = \beta_2 = \cdots = \beta_p = 0$（すべての係数が０）」
H1：「$\beta_1, \beta_2, \ldots, \beta_p$の少なくとも一つはゼロではない」

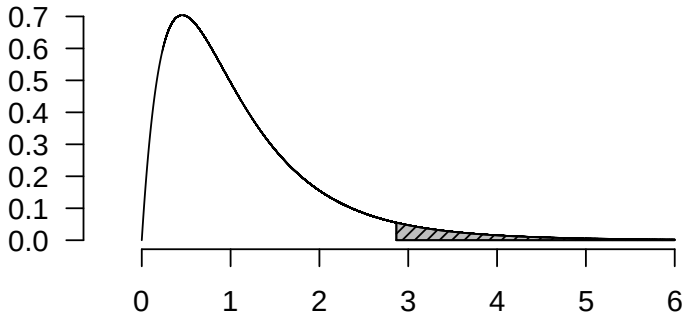
② 検定統計量Ｆ（重回帰式の検定における検定統計量）

$$F = \frac{ss_{\hat{y}}/p \quad \text{「}\hat{y}_i\text{の全体平均からのばらつき」}}{ss_e/(n-p-1) \quad \text{「}y_i\text{の}\hat{y}_i\text{からのばらつき」}}$$

（※pは独立変数の数）

③ Ｆ分布

H0 が真なら F 値は自由度$p, n-p-1$の F 分布に従う。



当然出てくるであろう疑問

疑問１：Ｆ分布ってどこから登場したの？

⇒ $N(0,1)$からカイ二乗分布を経由して登場

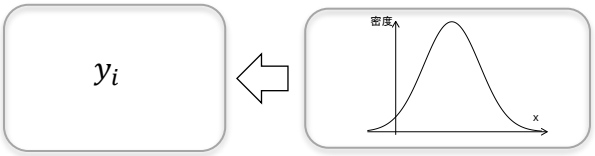
疑問２：自由度ってなにものなの？

⇒ カイ二乗分布の再生性に起因して登場

※ カイ二乗分布とＦ分布

① 正規分布

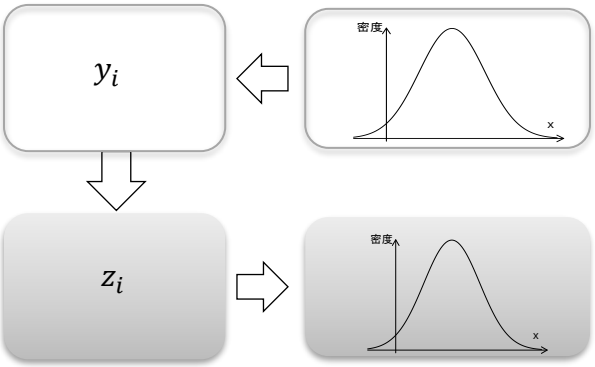
これは、ランダムな誤差が積み重なって登場する分布。



$$y_i \sim N(\mu, \sigma^2)$$

② 標準正規分布

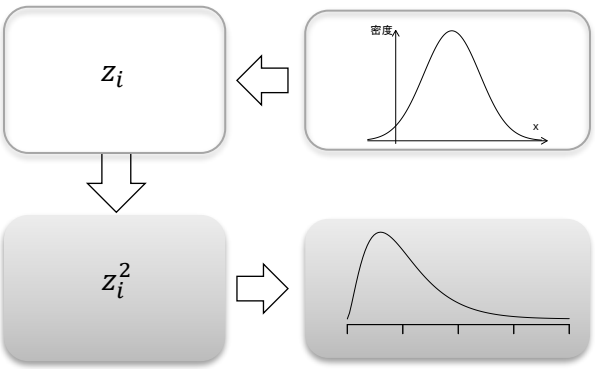
これは、平均が 0、分散が 1 の正規分布$N(0,1)$。正規分布$N(\mu, \sigma^2)$に従う変数y_iを標準化したz_iが従う分布。



$$z_i \sim N(0,1)$$

③ カイ二乗分布

これは$N(0,1)$に従う変数を二乗した統計量が従う分布。



$$z_i^2 \sim \chi^2(1)$$

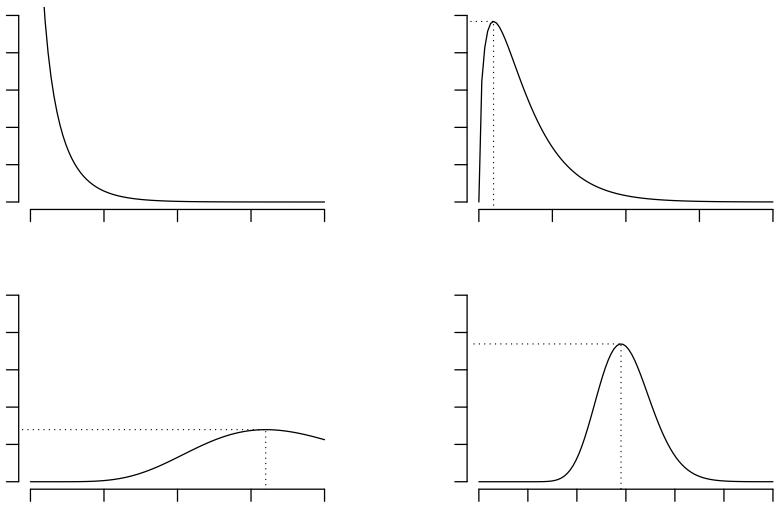
再生性（正規分布）

$$
\begin{aligned}
&y_1 \sim N(\mu_1, \sigma_1^2)\\
&y_2 \sim N(\mu_2, \sigma_2^2)\\
&\quad\vdots\\
+)\ &y_n \sim N(\mu_n, \sigma_n^2)\\
\hline
&y_1 + y_2 + \cdots + y_n\\
&\quad \sim N(\mu_1 + \mu_2 + \cdots + \mu_n, \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2)
\end{aligned}
$$

再生性（カイ二乗分布）

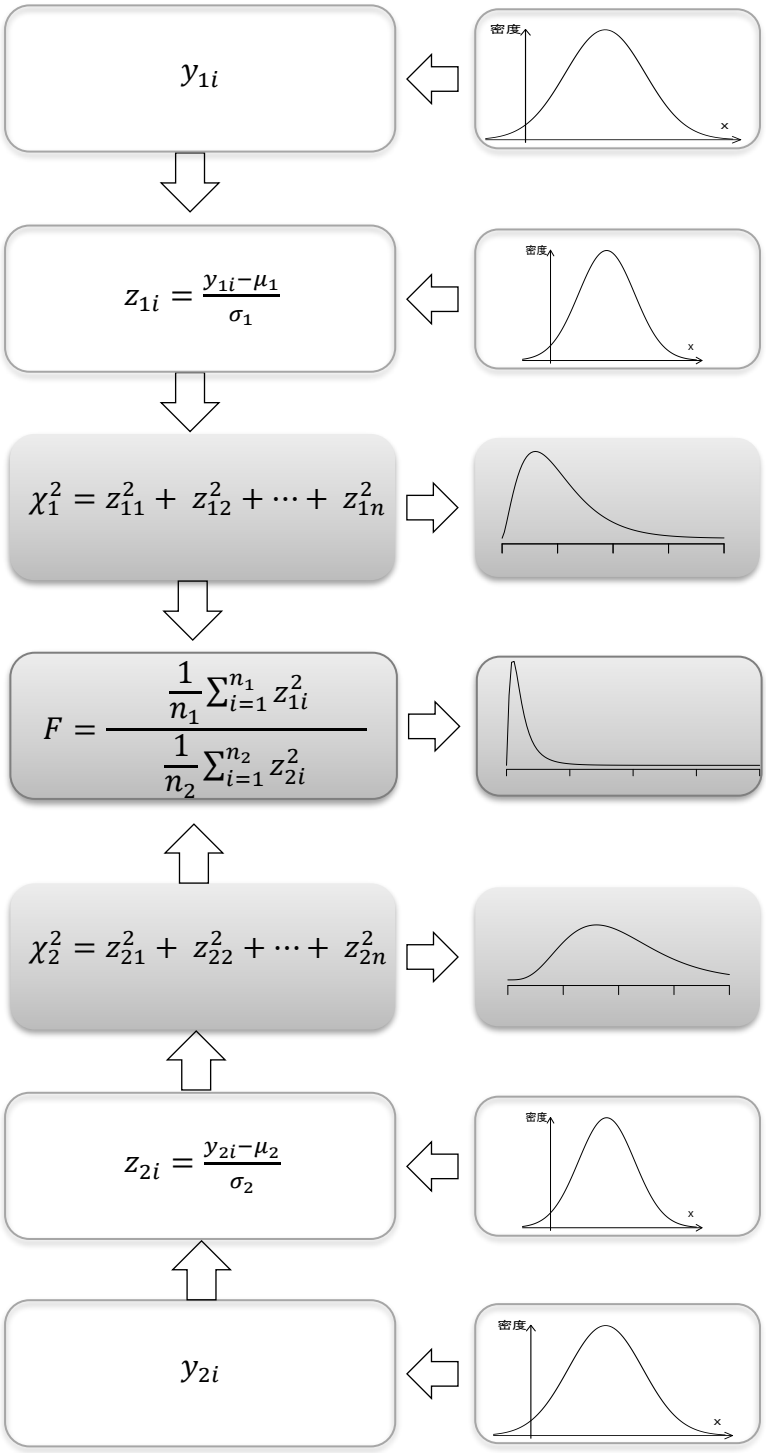
$$
\begin{aligned}
&z_1^2 \sim \chi^2(1)\\
&z_2^2 \sim \chi^2(1)\\
&\quad\vdots\\
+)\ &z_n^2 \sim \chi^2(1)\\
\hline
&z_1^2 + z_2^2 + \cdots + z_n^2 \sim \chi^2(n)
\end{aligned}
$$

カイ二乗分布



④ F 分布

これは、互いに独立な、二つのカイ二乗分布に従う変数 $\chi_1^2 \sim \chi^2(n_1), \chi_2^2 \sim \chi^2(n_2)$の比が従う分布。



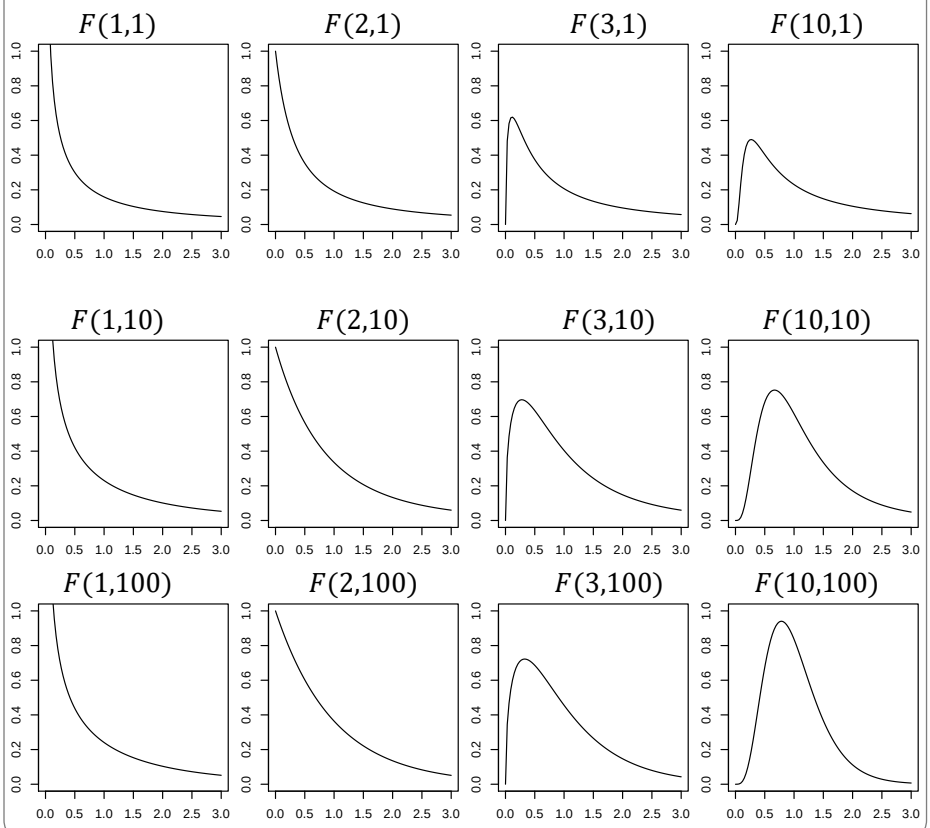
定義：F 値

$$F = \frac{\frac{1}{n_1}\chi_1^2}{\frac{1}{n_2}\chi_2^2} = \frac{\frac{1}{n_1}\sum_{i=1}^{n_1} z_{1i}^2}{\frac{1}{n_2}\sum_{j=1}^{n_2} z_{2j}^2}$$

$$= \frac{\frac{1}{n_1}\left(z_{1i}\text{の平方和}\right)}{\frac{1}{n_2}\left(z_{2j}\text{の平方和}\right)}$$

$$= \frac{z_{1i}\text{の分散}}{z_{2j}\text{の分散}}$$

F 分布



A 標準化する際には中心を確認！（統計量y_iを標準化）

$$y_i \sim N(\mu, \sigma^2)$$

人間の視点

$$\frac{y_i - \bar{y}}{\sigma} \not\sim N(0,1)$$

$$\sum_{i=1}^{n}\left(\frac{y_i - \bar{y}}{\sigma}\right)^2 \not\sim \chi^2(n)$$

全知全能の視点

$$\frac{y_i - \mu}{\sigma} \sim N(0,1)$$

$$\sum_{i=1}^{n}\left(\frac{y_i - \mu}{\sigma}\right)^2 \sim \chi^2(n)$$

B 標準化する際には中心を確認！　（統計量y_iを変換）

$$y_i \sim N(\mu, \sigma^2)$$

$$\sum_{i=1}^{n}\left(\frac{y_i - \bar{y}}{\sigma}\right)^2 \sim \chi^2(n-1)$$

C 標準化する際には中心を確認！（統計量 $\bar{y}$を標準化）

$$y_i \sim N(\mu, \sigma^2)$$

$$\bar{y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{\bar{y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$\left(\frac{\bar{y} - \mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi^2(1)$$

B において自由度が n-1 になる理由

$$\underbrace{\sum_{i=1}^{n}\left(\frac{y_i-\mu}{\sigma}\right)^2}_{\text{A}} = \underbrace{\sum_{i=1}^{n}\left(\frac{y_i-\bar{y}}{\sigma}\right)^2}_{\text{B}} + \underbrace{\left(\frac{\bar{y}-\mu}{\sigma/\sqrt{n}}\right)^2}_{\text{C}}$$

$\chi^2(n)$

ということは
再生性から

$\chi^2(n-1)$

$\chi^2(1)$

証明

$$\sum_{i=1}^{n}\left(\frac{y_i-\mu}{\sigma}\right)^2$$

$$=\sum_{i=1}^{n}\left\{\frac{(y_i-\bar{y})+(\bar{y}-\mu)}{\sigma}\right\}^2$$

$$=\sum_{i=1}^{n}\left\{\left(\frac{y_i-\bar{y}}{\sigma}\right)^2+\frac{2(y_i-\bar{y})(\bar{y}-\mu)}{\sigma^2}+\left(\frac{\bar{y}-\mu}{\sigma}\right)^2\right\}$$

$$=\sum_{i=1}^{n}\left(\frac{y_i-\bar{y}}{\sigma}\right)^2+\frac{2(\bar{y}-\mu)}{\sigma^2}\sum_{i=1}^{n}(y_i-\bar{y})+\sum_{i=1}^{n}\left(\frac{\bar{y}-\mu}{\sigma}\right)^2$$

$$=\sum_{i=1}^{n}\left(\frac{y_i-\bar{y}}{\sigma}\right)^2+\frac{2(\bar{y}-\mu)}{\sigma^2}\sum_{i=1}^{n}(y_i-\bar{y})+n\left(\frac{\bar{y}-\mu}{\sigma}\right)^2$$

$$=\sum_{i=1}^{n}\left(\frac{y_i-\bar{y}}{\sigma}\right)^2+\frac{2(\bar{y}-\mu)}{\sigma^2}\sum_{i=1}^{n}(y_i-\bar{y})+\left(\frac{\bar{y}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2$$

$$=\sum_{i=1}^{n}\left(\frac{y_i-\bar{y}}{\sigma}\right)^2+\left(\frac{\bar{y}-\mu}{\sigma/\sqrt{n}}\right)^2$$

（２）検定２：「個別」の回帰係数の検定　　ステップ２

① 帰無仮説と対立仮説

H0：$\beta_1 = 0$

H1：$\beta_1 \neq 0$

② 検定統計量 t

☞ t 値（回帰係数の検定における検定統計量）

$$t = \frac{\hat{\beta}_i \text{「}\beta_i\text{の推定値」}}{\hat{\sigma}_{\hat{\beta}_i} \text{「}\beta_i\text{の推定値」の標準誤差の推定値}}$$

	点推定値	標準誤差	t value	Pr(>|t|)	
$\hat{\beta}_0$	-0.49	0.08	-6.30	1.13e-06	***
$\hat{\beta}_1$	1.02	0.15	6.66	4.63e-07	***

F-statistic: 36.57 on 3 and 98 DF, p-value: 6.032e-16

（３）区間推定　　ステップ３

① 偏回帰係数の信頼区間

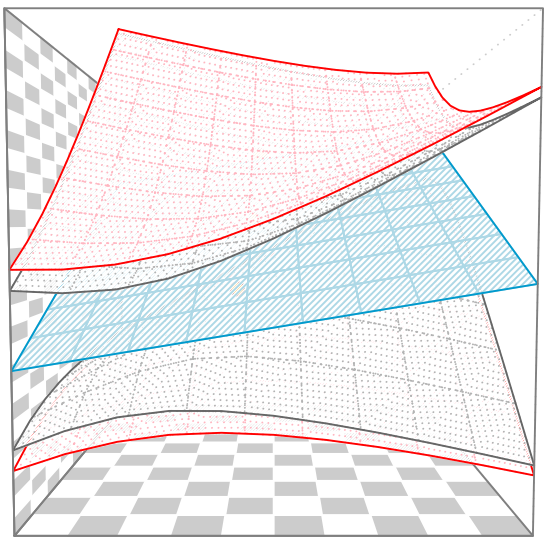
標本を繰り返し抽出した時、この範囲を設けておけば100回に95回、真のβ_0, β_1を含むだろう、という区間。

② 予測値の95%信頼区間　　（ 図の灰色の区間 ）

標本を繰り返し抽出した時、点x_jに対してこの範囲を設けておけば100回に95回μ_jを含むだろう、という区間。

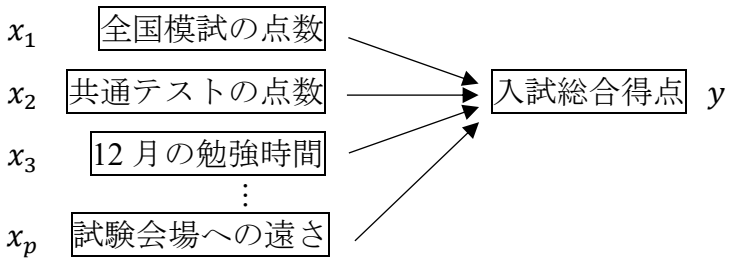
③ データの95%予測区間　　（ 図の赤色の区間 ）

標本を繰り返し抽出した時、点x_jに対してこの範囲を設ければ100回に95回データを含むだろうという区間。



🕮 ノート６ 分析の評価３：想定した統計モデルの適切さの評価
（モデルの改良）

例：入試当日の得点の予想



（１）目的：素性選択 Feature selection
重要な独立変数（素性）のみに絞りたい！

（動機１） 過学習 overfitting
独立変数が多いと訓練データへの依存度が強まり新標本への対応力が落ちてしまう。

（動機２） 解釈可能性 interpretability
独立変数が多いと解釈が難しくなる。

工学的興味：予測精度の向上 ⇕ 理論的興味：有意味な解釈の追及

（動機３） 多重共線性 Multicollinearity
独立変数が多いと独立変数が互いに強い関連性を持つ非理想的な状況が生まれやすい。

⇒ 偏回帰係数の推定量が不安定になる。

多数存在する独立変数を適切に操作しより良いモデルを模索するための方法として次の三つの方略がよく紹介される。

案１（モデル縮約）： 重要度の低い変数を削除する
案２（モデル比較）： よりスマートなモデルを探す
案３（合成指標） ：変数同士を合成し新指標を作る

（２）多重共線性　Multicollinearity

独立変数が互いに強い関連性を持ち、偏回帰係数の推定量が不安定になってしまう状態のこと。

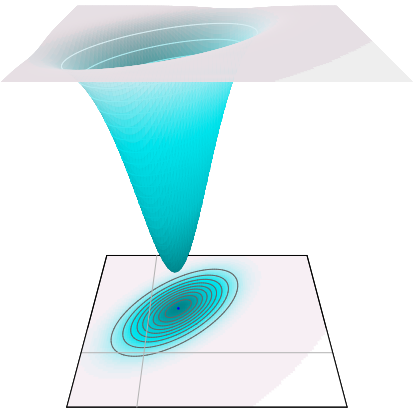
① 独立変数同士の相関が高い場合

例：店舗の売り上げの予測

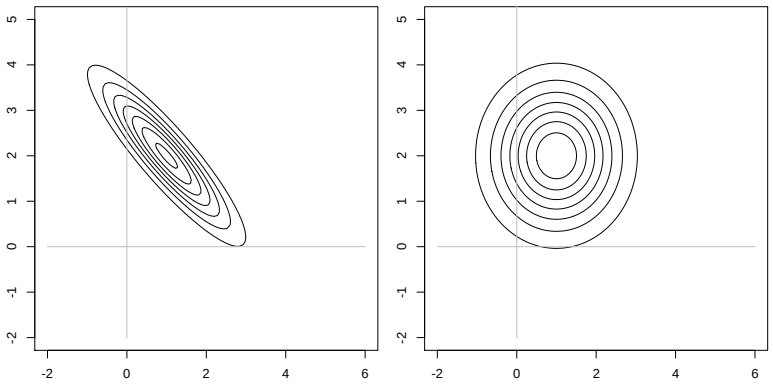
x_1 月間の雨天の日数 → y

x_2 月間の降雨量 → 店舗の売り上げ

② 点推定の幾何的な理解



（ケース１）　（ケース２）



③ 分散拡大要因 VIF（Variation Inflation Factor）

多重共線性の診断に使われる指標。10 以上のものは危険だが、4 程度の数値でも多重共線性に注意するといい。

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{x_1|x_2}}$$

（２）モデル縮約 Model Shrinkage（正則化 Regularization）

① 基本的なアイデア

モデルの複雑さを反映する罰則（正則化項）を設けた上で推定を行い、偏回帰係数を最小化・削除を行う。

通常の最小二乗推定

$$L(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} e_i^2$$ を最小化

正則化項付きの最小二乗推定

$R(\beta_0, \beta_1, \ldots, \beta_p) \leq s$という制約を満たす範囲で

$$L(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} e_i^2$$ を最小化

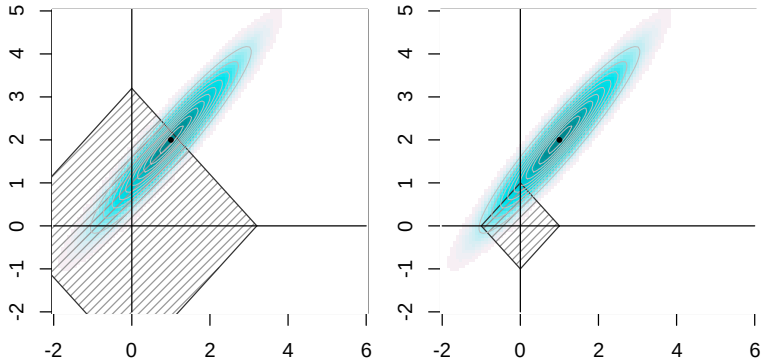
② ラッソ回帰 Lasso Regression

$R(\beta_0, \beta_1, \ldots, \beta_p)$の部分にL1 ノルムを用いたもの。

$\sum_{i=1}^{n} \beta_i \leq s$という制約を満たす範囲で

$$L(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} e_i^2$$ を最小化

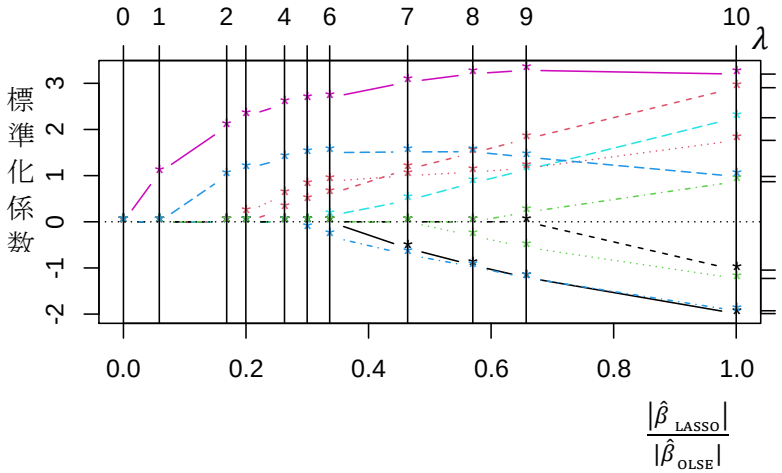
λを大きくすると正則化項の罰則が増大。β_jたちは大きい値を取りづらくなり効果の小さいものから0になる。



③ 実践上の手続き

A. チューニング

λの値によって、通常の最小二乗推定量（OLSE）と比べてどのくらい係数の値が小さくなるのかを吟味する。



※係数の比較：標準化回帰係数にしておくのがコツ。

B. 変数の数の選択　　（☞ クロスバリデーション）

クロスバリデーションを行い、最もパフォーマンスが高いモデルを採用することが多い。

```
> library(glmnet)
> data(QuickStartExample)

> fit <- glmnet(x, y)
> plot(fit)
> coef(fit, s = 0.1)

> cvfit <- cv.glmnet(x, y)
> plot(cvfit)
> coef(cvfit, s = "lambda.min")
> predict(cvfit, newx = x[1:5,], s = "lambda.min")
```

https://glmnet.stanford.edu/articles/glmnet.html

（3）モデル選択 Model selection

これは、考えられるモデルたちの中で、最もよいパフォーマンスを持つモデルを探す試みのこと。

モデル選択のステップ

手順１：モデルを比較するための基準を決める
手順２：考えられるモデルたちのパフォーマンスを測る
手順３：最良のパフォーマンスを持つモデルを解釈する

① モデルのパフォーマンスを測る指標

（系統１）訓練データの誤差を修正した指標

[1] 赤池情報量基準（AIC） 小

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$

[2] ベイズ情報量基準（BIC） 小

$$BIC = \frac{1}{n}(RSS + \log(n)\, d\hat{\sigma}^2)$$

[3] Mallow’ｓ C_p 小

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

[4] 自由度調整済み決定係数 大

$$R^2_{adjusted} = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

（系統２）テストデータの誤差を近似した指標

[1] Leave-One-Out Cross-Validation 小

$$CV_{(n)} = \frac{1}{n}\sum_{j=1}^{n} MSE_j^2$$

[2] K-fold CV 小

$$CV_{(k)} = \frac{1}{k}\sum_{j=1}^{k} MSE_j^2$$

② ベストなモデルを見つけるためのアルゴリズム

（アルゴリズム１）Best Subset Selection
2^p個のモデルすべてを比較する

（アルゴリズム２）Stepwise Selection
モデルの部分集合を作り勝ち抜き
試合式で$1 + p(p+1)/2$個を比較

総当たり法

	x_1	x_2	x_3	x_4
M1	×	×	×	×
M2	○	×	×	×
M3	×	○	×	×
M4	×	×	○	×
M5	×	×	×	○
M6	○	○	×	×
M7	○	×	○	×
M8	○	×	×	○
M9	×	○	○	×
M10	×	○	×	○
M11	×	×	○	○
M12	○	○	○	×
M13	○	○	×	○
M14	○	×	○	○
M15	×	○	○	○
M16	○	○	○	○

変数増加法

		x_1	x_2	x_3	x_4
M1	★	×	×	×	×
M2		○	×	×	×
M3		×	○	×	×
M4	★	×	×	○	×
M5		×	×	×	○
M7		○	×	○	×
M9		×	○	○	×
M11	★	×	×	○	○
M14	★	○	×	○	○
M15		×	○	○	○
M16	★	○	○	○	○

③ 実践上の手続き

```
> library(leaps)
> longley = data.frame(scale(longley))
> y = longley$Employed; x = longley[,1:6]
> r1 = regsubsets(x,y)
> l1 = leaps(x,y,nbest =1,method ="r2")
> plot(r1); plot(l1$size, l1$adjr2, type = "b")
```