授業計画

Prologue: データを読み解くとは何なのか
(多変量解析とデータサイエンスと統計学とパターン認識と機械学習とデータマイニング)

DAY-1 6/16 (01)(02) **単回帰: 点群への直線当てはめを”真剣に”考える**
(見えない世界へようこそ)
DAY-2 6/23 (03)(04) **重回帰と線形代数: 回帰の行列計算とその意味**
(データの計算とデータの解釈)
DAY-3 6/30 (05)(06) **重回帰と確率統計: なぜ回帰に確率が必要?**
(推測統計入門:データの向こう側について語るための代償)
DAY-4 7/07 (07)(08) **多変量正規分布: 多次元の正規分布と線形代数**
(ゼロから理解する正規分布)
DAY-5 7/14 (09)(10) **マハラノビス距離と判別分析: 線形代数を使う1**
(最適な判別とは)
DAY-6 7/28 (11)(12) **固有値分解と主成分分析: 線形代数を使う2**
(高次元データがかかえる大問題)
DAY-7 8/04 (13)(14) **特異値分解と数量化: 線形代数を使う3**
(数値じゃない対象に統計を効かすには)

Epilogue: 基礎の上に在る世界(話したことと話さなかったこと)

今日の話：回帰と判別

[午前]
- 回帰分析の多変量版(重回帰)
- 決定係数なども含めて具体的に計算をフォローしてみる
- 多変量版のまとめ

[午後]
- 判別と回帰
- 判別分析
- マハラノビス距離

多変量の回帰分析

表 1.3　中古マンションのデータ

サンプル No.	広さ x_1 (m^2)	築年数 x_2 (年数)	価格 y (千万円)
1	51	16	3.0
2	38	4	3.2
3	57	16	3.3
4	51	11	3.9
5	53	4	4.4
6	77	22	4.5
7	63	5	4.5
8	69	5	5.4
9	72	2	5.4
10	73	1	6.0

説明変数　　目的変数

多変量の回帰分析

x1	x2	y
3.830	−1.487	37.537
1.207	3.920	13.162
−1.302	4.093	4.851
3.291	2.659	23.068
0.109	0.284	16.770
−2.147	5.206	−1.665
−1.067	−1.053	17.103
1.798	−2.292	29.806
3.311	−2.420	32.913
−3.856	0.860	−0.098
−2.968	4.926	−5.243
3.775	4.328	20.182
−0.701	−3.429	22.761
−2.690	−1.836	11.897
−0.427	−3.381	24.219
2.863	−1.056	30.548
1.571	−3.674	33.147
2.254	−0.160	25.374
−4.490	−9.077	23.642
0.702	−3.250	27.714

説明変数　　目的変数

- x1とx2を用いてyを説明したい
- (x1,x2)が分かればyの妥当な値を予測したい

左のデータが与えられた時
どんな傾向がわかる？

多変量の回帰分析(当てるだけ)

x1	x2	y
3.830	−1.487	37.537
1.207	3.920	13.162
−1.302	4.093	4.851
3.291	2.659	23.068
0.109	0.284	16.770
−2.147	5.206	−1.665
−1.067	−1.053	17.103
1.798	−2.292	29.806
3.311	−2.420	32.913
−3.856	0.860	−0.098
−2.968	4.926	−5.243
3.775	4.328	20.182
−0.701	−3.429	22.761
−2.690	−1.836	11.897
−0.427	−3.381	24.219
2.863	−1.056	30.548
1.571	−3.674	33.147
2.254	−0.160	25.374
−4.490	−9.077	23.642
0.702	−3.250	27.714

説明変数　　目的変数

$$\boldsymbol{X} = \begin{bmatrix} 1 & 3.830 & -1.487 \\ 1 & 1.207 & 3.920 \\ 1 & -1.302 & 4.093 \\ 1 & 3.291 & 2.659 \\ \vdots & \vdots & \vdots \\ 1 & 0.702 & -3.250 \end{bmatrix}, \boldsymbol{y} = \begin{bmatrix} 37.537 \\ 13.162 \\ 4.851 \\ 23.068 \\ \vdots \\ 27.714 \end{bmatrix}$$

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} 17.654 \\ 3.619 \\ -2.381 \end{bmatrix}$$

y = 17.654 + 3.619 x1 - 2.381 x2

多変量の回帰分析(さらに深掘り)

x1	x2	y
3.830	−1.487	37.537
1.207	3.920	13.162
−1.302	4.093	4.851
3.291	2.659	23.068
0.109	0.284	16.770
−2.147	5.206	−1.665
−1.067	−1.053	17.103
1.798	−2.292	29.806
3.311	−2.420	32.913
−3.856	0.860	−0.098
−2.968	4.926	−5.243
3.775	4.328	20.182
−0.701	−3.429	22.761
−2.690	−1.836	11.897
−0.427	−3.381	24.219
2.863	−1.056	30.548
1.571	−3.674	33.147
2.254	−0.160	25.374
−4.490	−9.077	23.642
0.702	−3.250	27.714

説明変数　　目的変数

仮定：yは以下のルールで生成されたものとしてみる。(正規線形モデル)

$$y = b0 + b1\ x1 + b2\ x2 + z$$

b0, b1, b2　→　未知だが固定値

z　→　乱数

ただし正規分布 N(0,σ)に従うと仮定

σ　→　未知だが固定値

多変量の回帰分析(さらに深掘り)

n = 20

x1	x2	y
3.830	−1.487	37.537
1.207	3.920	13.162
−1.302	4.093	4.851
3.291	2.659	23.068
0.109	0.284	16.770
−2.147	5.206	−1.665
−1.067	−1.053	17.103
1.798	−2.292	29.806
3.311	−2.420	32.913
−3.856	0.860	−0.098
−2.968	4.926	−5.243
3.775	4.328	20.182
−0.701	−3.429	22.761
−2.690	−1.836	11.897
−0.427	−3.381	24.219
2.863	−1.056	30.548
1.571	−3.674	33.147
2.254	−0.160	25.374
−4.490	−9.077	23.642
0.702	−3.250	27.714

説明変数　　目的変数

実は左のデータは以下のルールで生成

$y = b0 + b1\ x1 + b2\ x2 + z$

z → $N(0,\sigma^2)$に従う乱数

真の答え
(だが未知)

b0 = 17.5
b1 = 3.2
b2 = -2.4
σ^2 = 4.0

データからどれくらい当たる？

→ 数学的に議論可能

多変量の回帰分析(さらに深掘り)

n = 20

x1	x2	y	$\hat{y}$	ε
3.830	−1.487	37.537	34.162	3.375
1.207	3.920	13.162	12.643	0.520
−1.302	4.093	4.851	4.160	0.691
3.291	2.659	23.068	22.392	0.676
0.109	0.284	16.770	17.916	−1.146
−2.147	5.206	−1.665	−1.252	−0.413
−1.067	−1.053	17.103	17.373	−0.270
1.798	−2.292	29.806	29.579	0.227
3.311	−2.420	32.913	34.752	−1.839
−3.856	0.860	−0.098	3.776	−3.874
−2.968	4.926	−5.243	−3.211	−2.032
3.775	4.328	20.182	19.909	0.274
−0.701	−3.429	22.761	24.301	−1.540
−2.690	−1.836	11.897	14.050	−2.153
−0.427	−3.381	24.219	25.066	−0.846
2.863	−1.056	30.548	30.011	0.537
1.571	−3.674	33.147	32.196	0.951
2.254	−0.160	25.374	25.886	−0.512
−4.490	−9.077	23.642	25.794	−2.152
0.702	−3.250	27.714	28.378	−0.664

説明変数　　目的変数　　予測　残差

真の答え
(だが未知)

b0 = 17.5
b1 = 3.2
b2 = -2.4
σ^2 = 4.0

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} 17.654 \\ 3.619 \\ -2.381 \end{bmatrix} \quad \begin{matrix} \text{(b0)} \\ \text{(b1)} \\ \text{(b2)} \end{matrix}$$

$$\hat{y} = 17.654 + 3.619x_1 - 2.381x_2$$

$$\varepsilon = y - \hat{y}$$

$$\frac{1}{n-3}\sum_i \varepsilon_i^2 = 3.00 \qquad (\sigma^2)$$

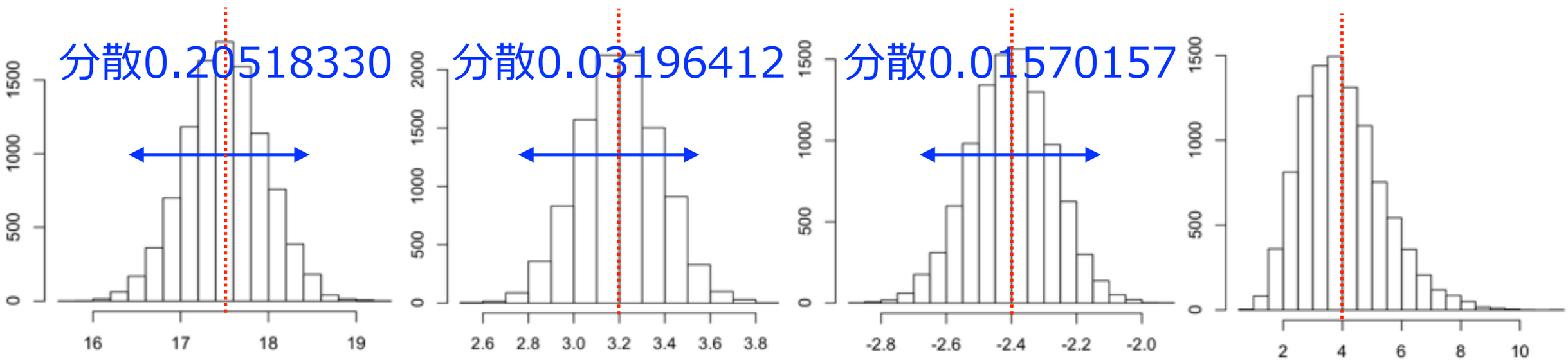
シミュレーションしてみた

$y = b0 + b1\ x1 + b2\ x2 + z$ (ただし$z \sim N(0,\sigma^2)$)

$b0 = 17.5,\ b1 = 3.2,\ b2 = -2.4,\ \sigma^2 = 4.0$

n = 20の事例を生成 ⟶ 予測(b0) (b1) (b2) (σ^2)を得る

を10000回やってみたときの、予測値10000個のヒストグラム



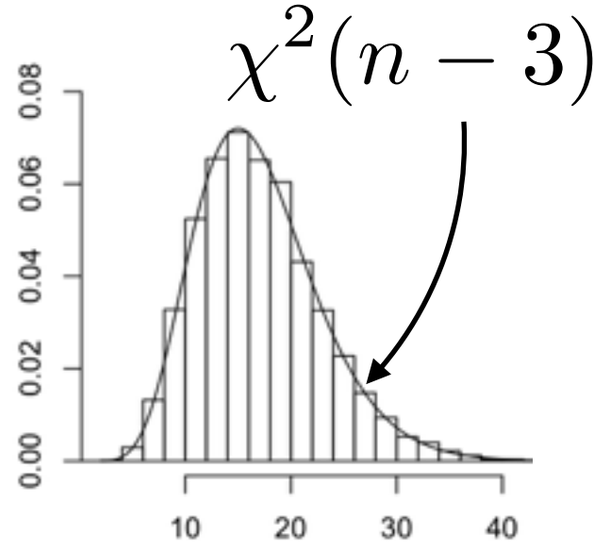
平均17.502188　平均3.201595　平均-2.401380　平均4.001294

理論値との整合

理論値(後で復習)

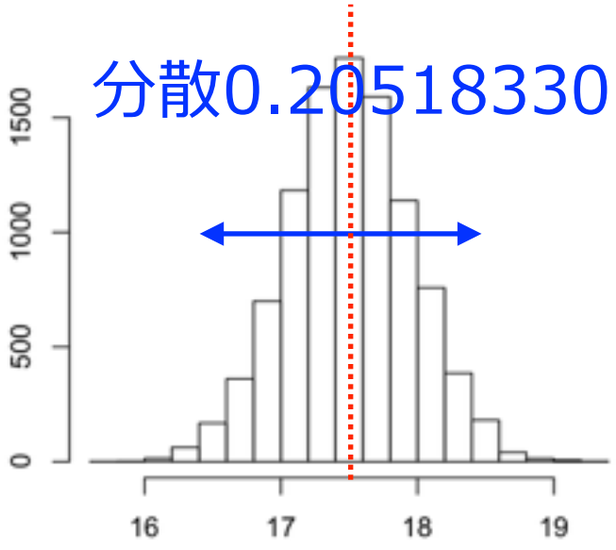
$$\mathbb{E}\{\hat{\boldsymbol{\beta}}\} = \boldsymbol{\beta} = \begin{bmatrix} 17.502188 \\ 3.201595 \\ -2.401380 \end{bmatrix}$$

$$\mathrm{var}\{\hat{\boldsymbol{\beta}}\} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 0.204234242 & -0.008670133 & 0.005962479 \\ -0.008670133 & 0.031197609 & -0.002255615 \\ 0.005962479 & -0.002255615 & 0.015764810 \end{bmatrix}$$



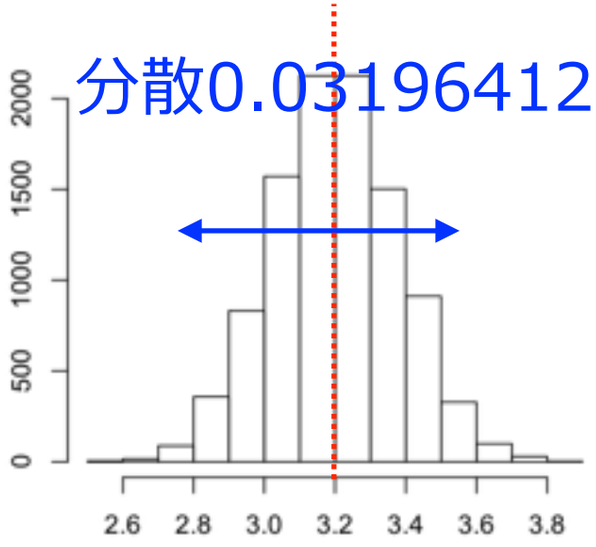
$(\sigma^2)\times 4.25$

$=(n-3)/\sigma^2$



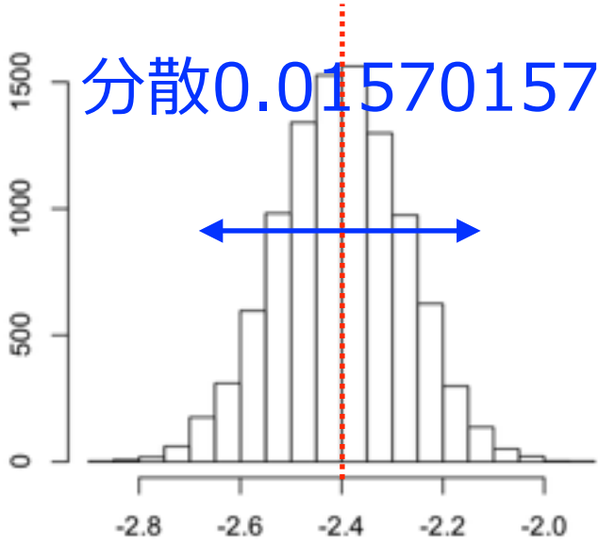
(b0)

平均17.502188



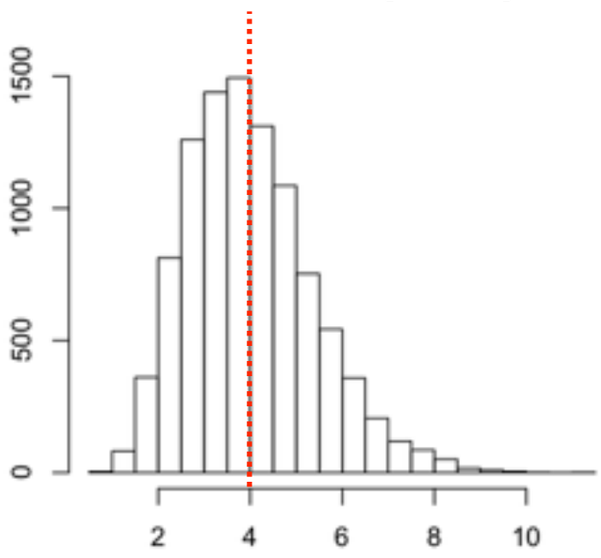
(b1)

平均3.201595



(b2)

平均-2.401380



(σ^2)

平均4.001294

理論が言っていること

本当は未知量だが真値 $\boldsymbol{\beta}$ と σ^2 がわかっているとする。

$$\longrightarrow \quad \hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

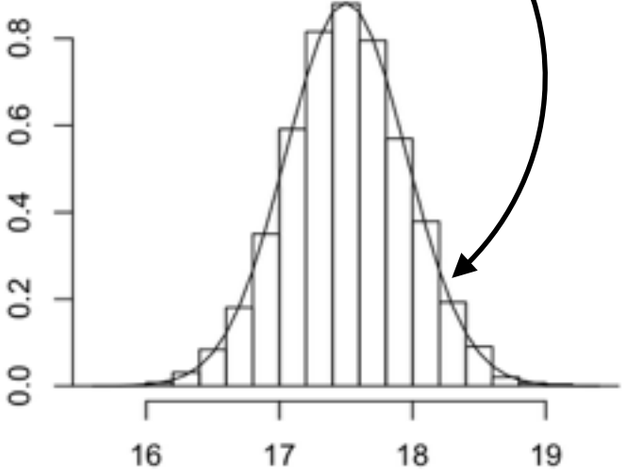
$$\hat{\sigma}^2 = \frac{1}{n-p-1}\|\boldsymbol{y}-\hat{\boldsymbol{y}}\|^2, \hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

の分布(どういう値を取りやすいか)が分かる

記法

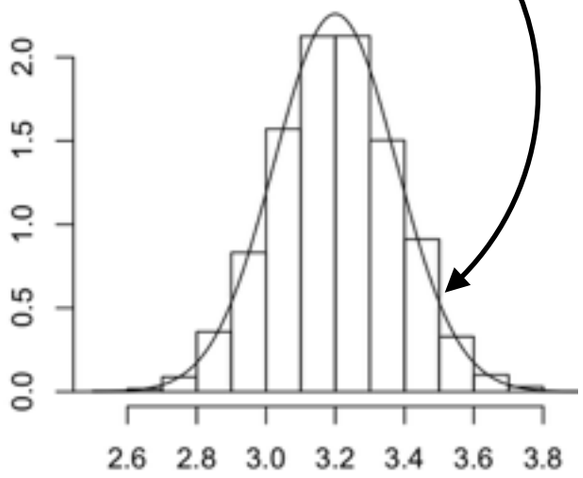
$$\boldsymbol{\beta} = \begin{bmatrix}\beta_0\\ \beta_1\\ \beta_2\end{bmatrix} \quad \hat{\boldsymbol{\beta}} = \begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix}$$

$N(\beta_0, (\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})_{11})$



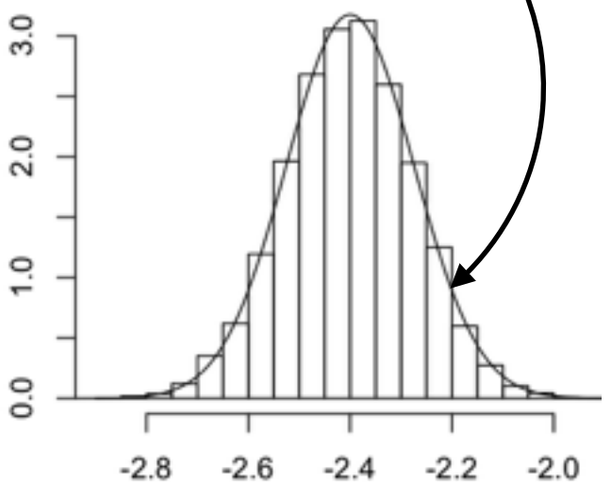
$\hat{\beta}_0$

$N(\beta_1, (\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})_{22})$



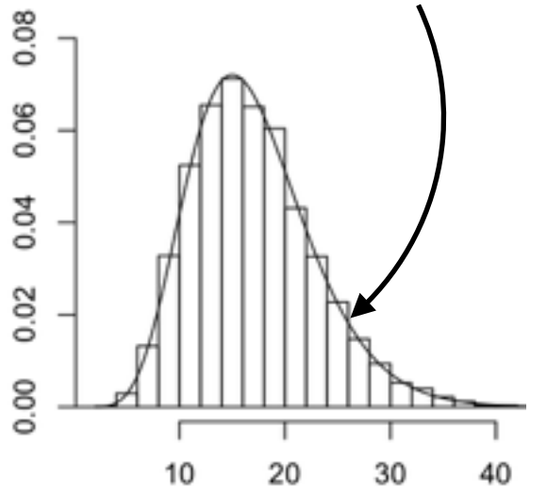
$\hat{\beta}_1$

$N(\beta_2, (\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})_{33})$



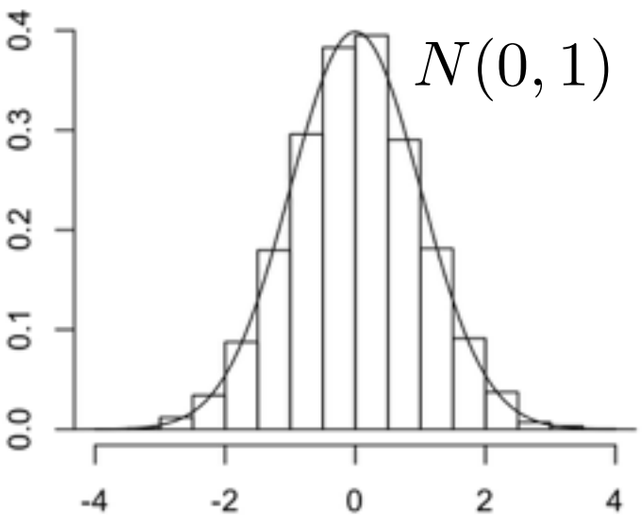
$\hat{\beta}_2$

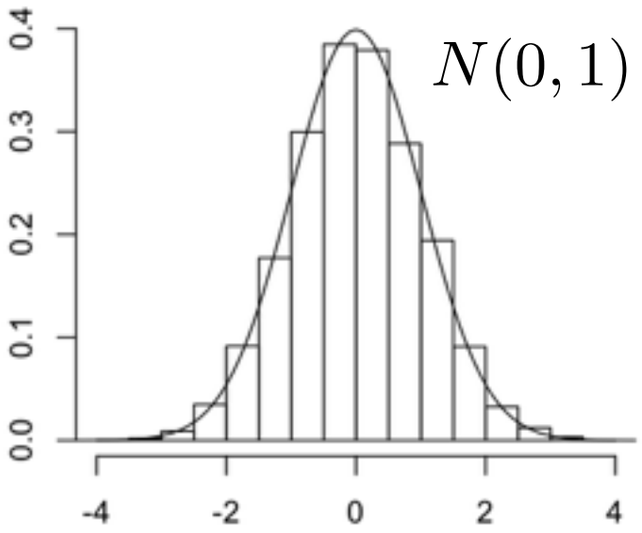
$\chi^2(n-3)$

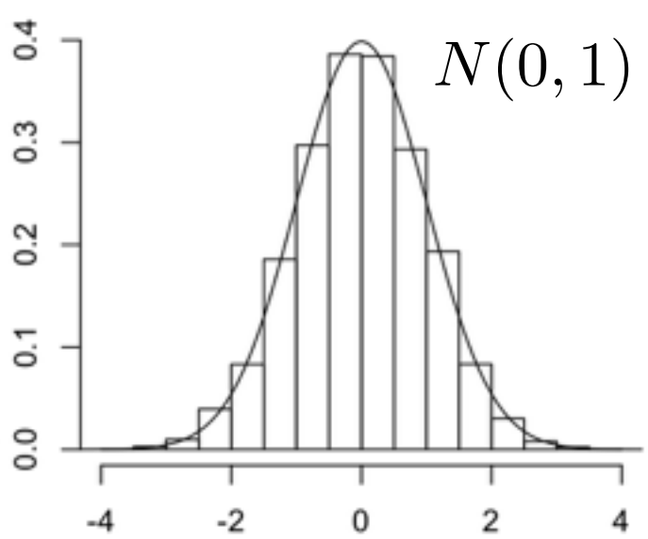


$\hat{\sigma}^2 \times \frac{n-3}{\sigma^2}$

統一的にいけるように標準化して考える







←同じ分布に

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{(\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})_{11}}}$$

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{(\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})_{22}}}$$

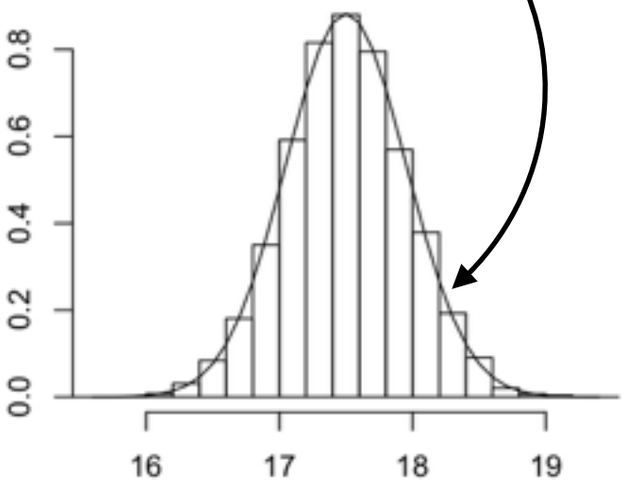
$$\frac{\hat{\beta}_2 - \beta_2}{\sqrt{(\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})_{33}}}$$

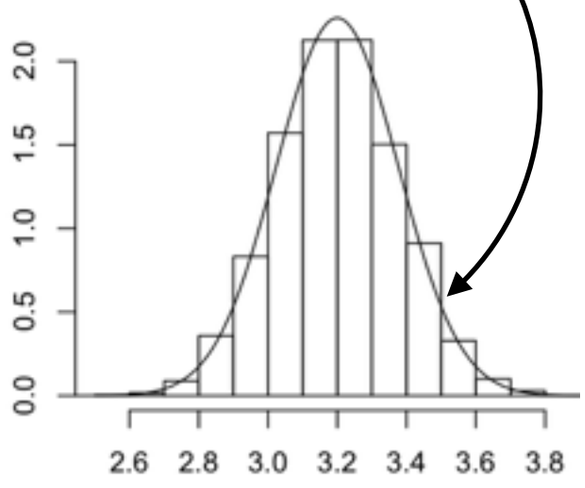
$N(\beta_0, (\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})_{11})$

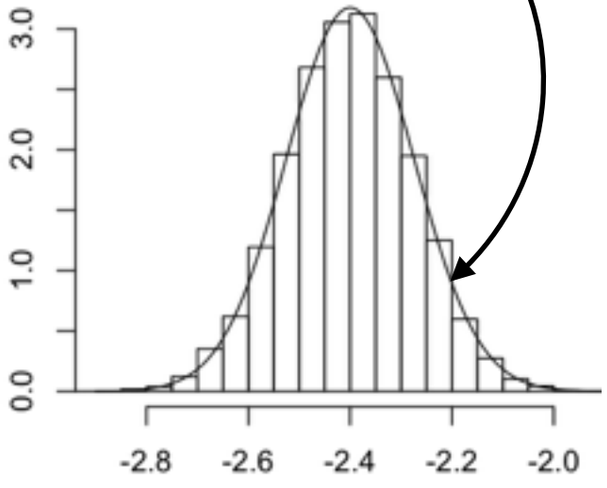
$N(\beta_1, (\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})_{22})$

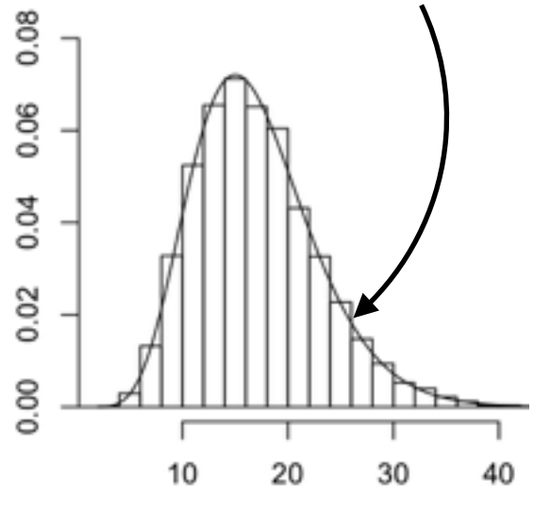
$N(\beta_2, (\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})_{33})$

$\chi^2(n-3)$









$\hat{\beta}_0$

$\hat{\beta}_1$

$\hat{\beta}_2$

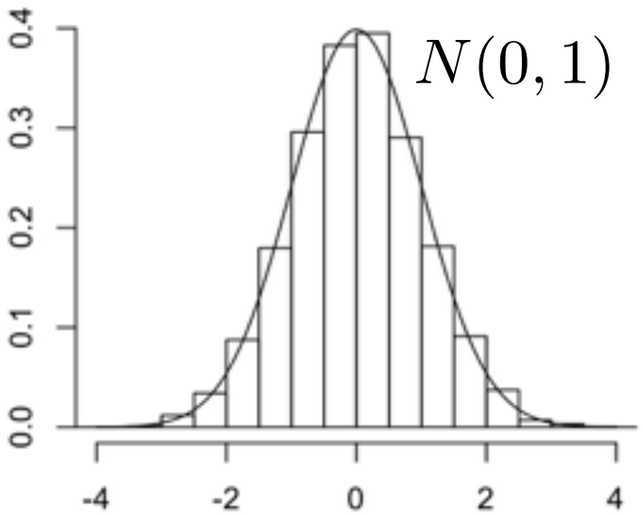
$$\hat{\sigma}^2 \times \frac{n-3}{\sigma^2}$$

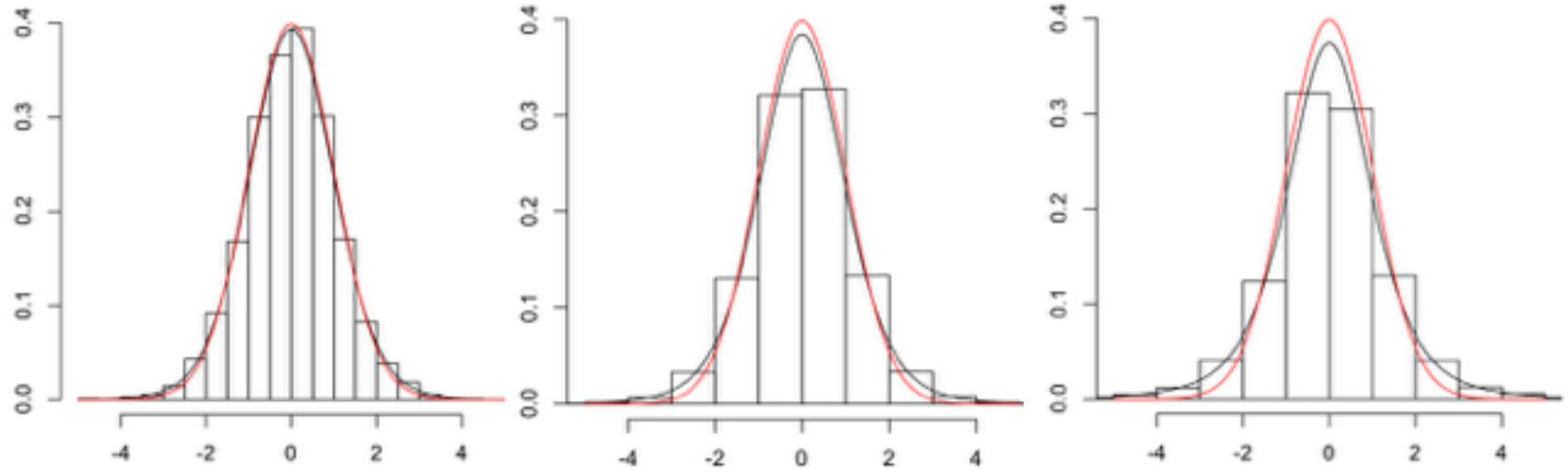
$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{(\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})_{11}}}$

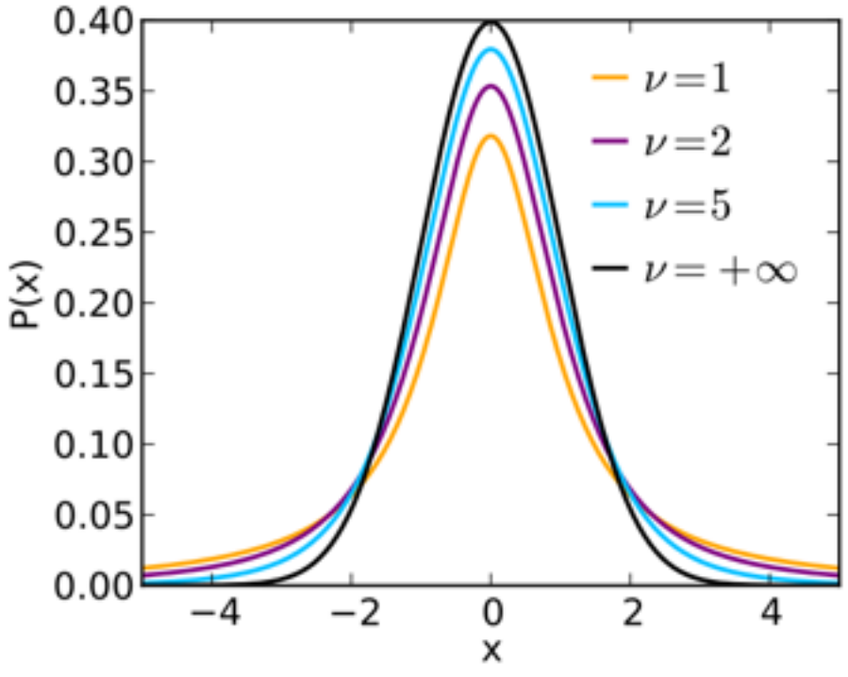
$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{(\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1})_{11}}}$

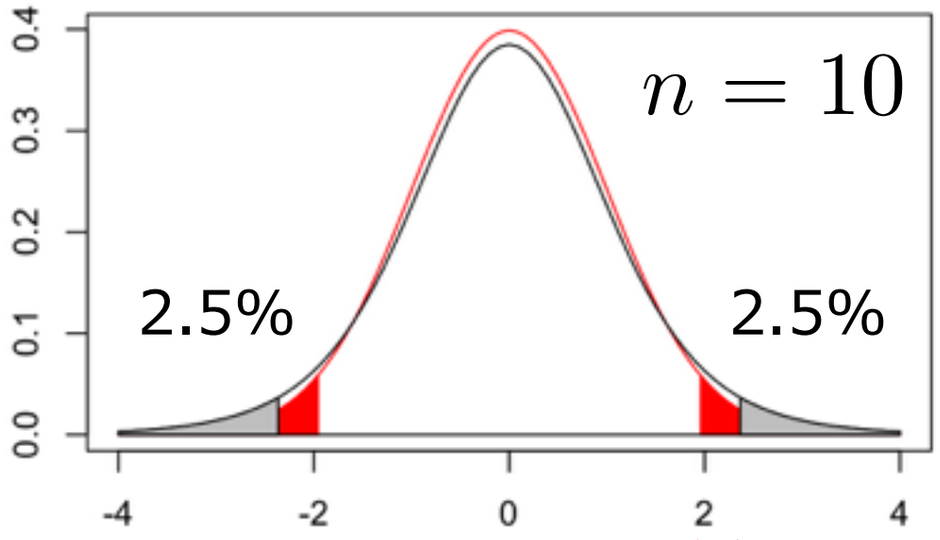
—— $t(n-3)$ —— $N(0,1)$

$n = 20$ $n = 10$ $n = 7$









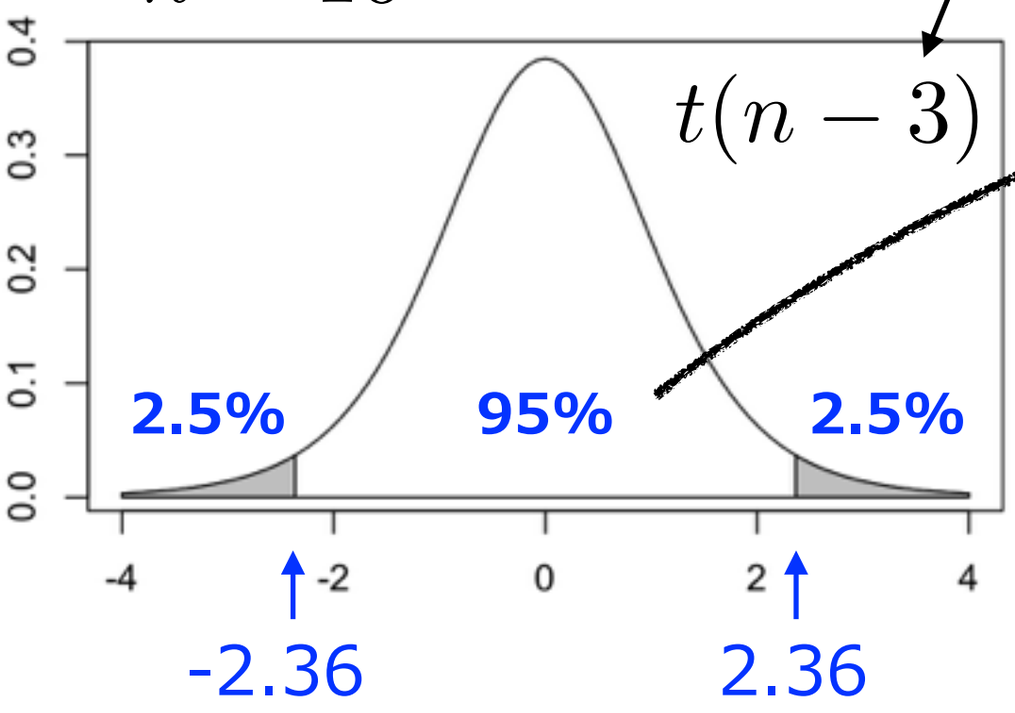
$N(0,1)$

$t(n-3)$

1.96

2.36

p変量のときn-p-1
(この例ではp=2変量)



信頼度	有意水準	境界値
90%	10%	1.89
95%	5%	2.36
99%	1%	3.50
99.9%	0.1%	5.41

95%で

$$-2.36 \leq \frac{\hat{\beta}_0 - \beta_0}{\sqrt{(\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1})_{11}}} \leq 2.36$$

【1】$\beta_0 = 0$ かどうかを検定

もし真とすると $\beta_0 \leftarrow 0$ と
代入すると上の範囲に95%で入るはず

【2】β_0 の95%信頼区間(CI)を推定

式変形すると

$$\hat{\beta}_0 - 2.36v \leq \beta_0 \leq \hat{\beta}_0 + 2.36v$$

$$v = \sqrt{(\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1})_{11}}$$

実際に計算した値と公式

$$n = 20, p = 2$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} 17.654 \\ 3.619 \\ -2.381 \end{bmatrix}$$

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

$$\bar{y} = 19.384$$

回帰統計	
重相関 R	0.995846
重決定 R2	0.991709
補正 R2	0.990734
標準誤差	1.179935
観測数	20

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	2	2831.067	1415.533	1016.726151	2.0328E-18
残差	17	23.66819	1.392247		
合計	19	2854.735			

Excel回帰分析では
「分散分析表」に出る数字

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = 2831.067 \quad \text{(回帰変動)}$$

$$\sum_{i=1}^{n}(y_i - \hat{y})^2 = 23.66819 \quad \text{(残差変動)}$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = 2854.735 \quad \text{(全変動)}$$

$$\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{p} = 1415.533$$

$$\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{(n-p-1)} = 1.392247$$

$$\frac{1415.533}{1.392247} = 1016.726 \sim F(p, n-p-1)$$

$$R^2 = \frac{2831.067}{2854.735} = 1 - \frac{23.66819}{2854.735} = 0.9917091$$

$$\sqrt{R^2} = 0.9958459 = \mathrm{cor}(\hat{y}, y)$$

$$R^{*2} = 1 - \frac{23.66819/(n-p-1)}{2854.735/(n-1)} = 0.9907338$$

Excel回帰分析では
「回帰統計」に出る数字

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2 = 1.392 \quad \sqrt{\hat{\sigma}^2} = 1.179935$$

重相関係数　$\boldsymbol{y}$と$\hat{\boldsymbol{y}}$の相関係数

$$R = \frac{(\boldsymbol{y}-\bar{\boldsymbol{y}})'(\hat{\boldsymbol{y}}-\bar{\boldsymbol{y}})}{\|\boldsymbol{y}-\bar{\boldsymbol{y}}\| \cdot \|\hat{\boldsymbol{y}}-\bar{\boldsymbol{y}}\|}$$

$$= \frac{(\boldsymbol{y}-\bar{\boldsymbol{y}})'H(\boldsymbol{y}-\bar{\boldsymbol{y}})}{\|\boldsymbol{y}-\bar{\boldsymbol{y}}\| \cdot \|\hat{\boldsymbol{y}}-\bar{\boldsymbol{y}}\|} = \frac{(\boldsymbol{y}-\bar{\boldsymbol{y}})'H^2(\boldsymbol{y}-\bar{\boldsymbol{y}})}{\|\boldsymbol{y}-\bar{\boldsymbol{y}}\| \cdot \|\hat{\boldsymbol{y}}-\bar{\boldsymbol{y}}\|} = \frac{\|H(\boldsymbol{y}-\bar{\boldsymbol{y}})\|^2}{\|\boldsymbol{y}-\bar{\boldsymbol{y}}\| \cdot \|\hat{\boldsymbol{y}}-\bar{\boldsymbol{y}}\|}$$

$$= \frac{\|\hat{\boldsymbol{y}}-\bar{\boldsymbol{y}}\|^2}{\|\boldsymbol{y}-\bar{\boldsymbol{y}}\| \cdot \|\hat{\boldsymbol{y}}-\bar{\boldsymbol{y}}\|} = \frac{\|\hat{\boldsymbol{y}}-\bar{\boldsymbol{y}}\|}{\|\boldsymbol{y}-\bar{\boldsymbol{y}}\|}$$

$$\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}} = \underline{X(X'X)^{-1}X'\boldsymbol{y}}$$
$$= H \text{ (射影行列)}$$

偏差平方和（全変動）　回帰による平方和　残差平方和

$$\|\boldsymbol{y}-\bar{\boldsymbol{y}}\|^2 = \|\hat{\boldsymbol{y}}-\bar{\boldsymbol{y}}\|^2 + \underline{\|\hat{\boldsymbol{y}}-\boldsymbol{y}\|^2}$$
$$= \|\boldsymbol{\epsilon}\|^2$$

※教科書p.48,p.69

決定係数(寄与率)

$$R^2 = \frac{\|\hat{\boldsymbol{y}}-\bar{\boldsymbol{y}}\|^2}{\|\boldsymbol{y}-\bar{\boldsymbol{y}}\|^2} = \frac{\|\boldsymbol{y}-\bar{\boldsymbol{y}}\|^2 - \|\boldsymbol{\epsilon}\|^2}{\|\boldsymbol{y}-\bar{\boldsymbol{y}}\|^2} = 1 - \frac{\|\boldsymbol{\epsilon}\|^2}{\|\boldsymbol{y}-\bar{\boldsymbol{y}}\|^2}$$

自由度調整済み決定係数(寄与率)

$$R^{*2} = 1 - \frac{\|\boldsymbol{\epsilon}\|^2/(n-p-1)}{\|\boldsymbol{y}-\bar{\boldsymbol{y}}\|^2/(n-1)}$$

分散分析とF値：回帰式が役に立つかどうかのF検定

回帰分散　$V_{\hat{\boldsymbol{y}}} := \dfrac{1}{p}\|\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}}\|^2$

誤差分散　$V_{\boldsymbol{\epsilon}} := \dfrac{1}{n-p-1}\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2 = \dfrac{1}{n-p-1}\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2$

分散比 $F := \dfrac{V_{\hat{\boldsymbol{y}}}}{V_{\boldsymbol{\epsilon}}}$ は、「仮説 $\boldsymbol{\beta} = \mathbf{0}$ が正しい」とき、

自由度 $p, n-p-1$ のF分布に従う。

※ $\boldsymbol{\beta} = \mathbf{0}$ なら投入した説明変数は目的変数の説明に何の役にも立っていないということなので、回帰式は無意味になる

実際に計算した値と公式(残り)

$$n = 20, p = 2$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} 17.654 \\ 3.619 \\ -2.381 \end{bmatrix}$$

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2 = 1.392$$

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 0.051058561 & -0.0021675332 & 0.0014906198 \\ -0.002167533 & 0.0077994023 & -0.0005639038 \\ 0.001490620 & -0.0005639038 & 0.0039412026 \end{bmatrix}$$

$$\sqrt{\hat{\sigma}^2 \cdot 0.051058561} = 0.2666198$$

$$\sqrt{\hat{\sigma}^2 \cdot 0.0077994023} = 0.104205$$

$$\sqrt{\hat{\sigma}^2 \cdot 0.0039412026} = 0.07407513$$

$$17.654/0.2666198 = 66.21322$$

$$3.619/0.104205 = 34.72737$$

$$-2.381/0.07407513 = -32.14857$$

$n = 20$	信頼度	有意水準	境界値
$t(n-3)$	95%	5%	2.11

17.654 + 2.11 * 0.2666198 = 18.21627
3.619 + 2.11 * 0.104205 = 3.838619
-2.381 + 2.11 * 0.07407513 = -2.225125

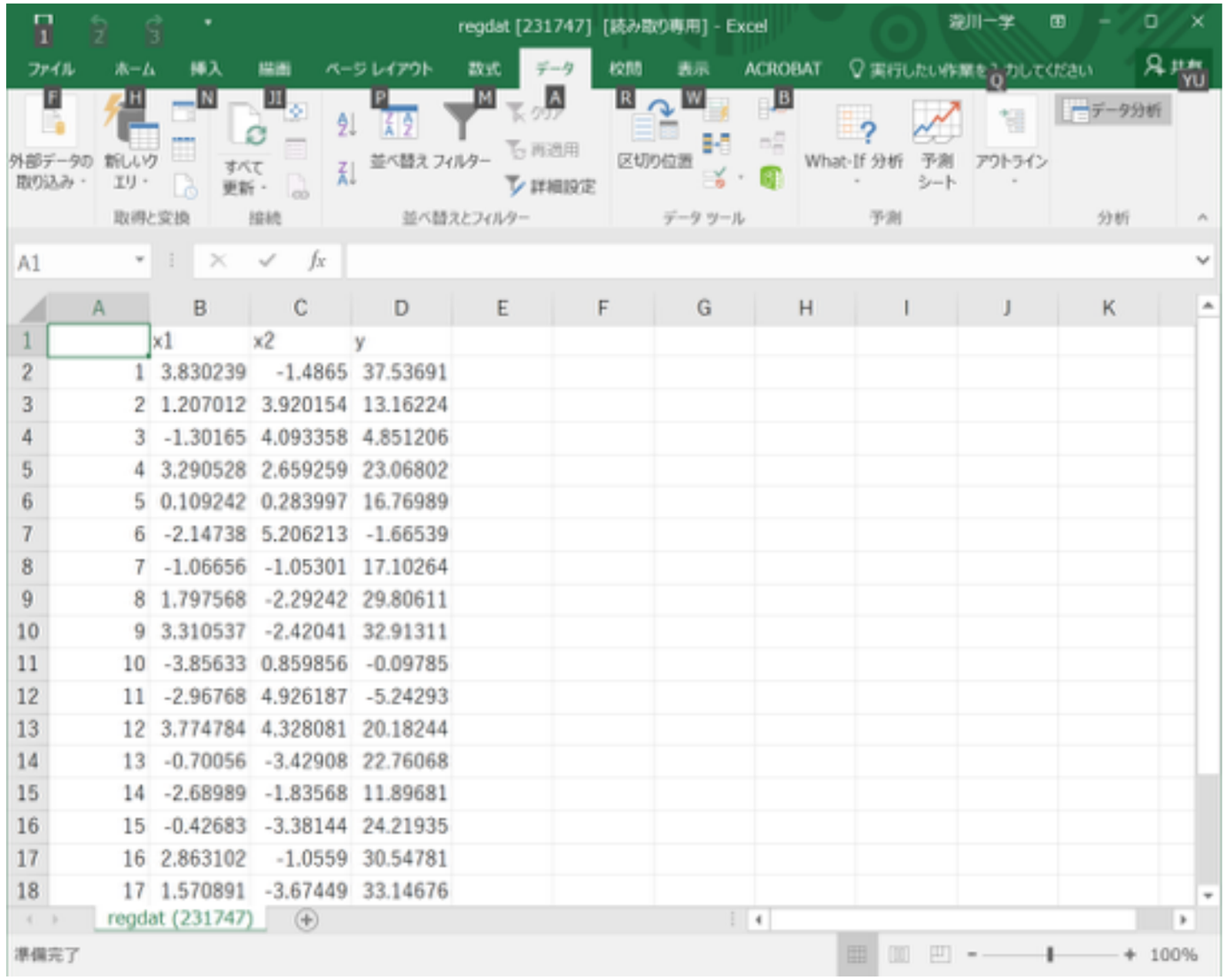
↑下限はここがマイナス

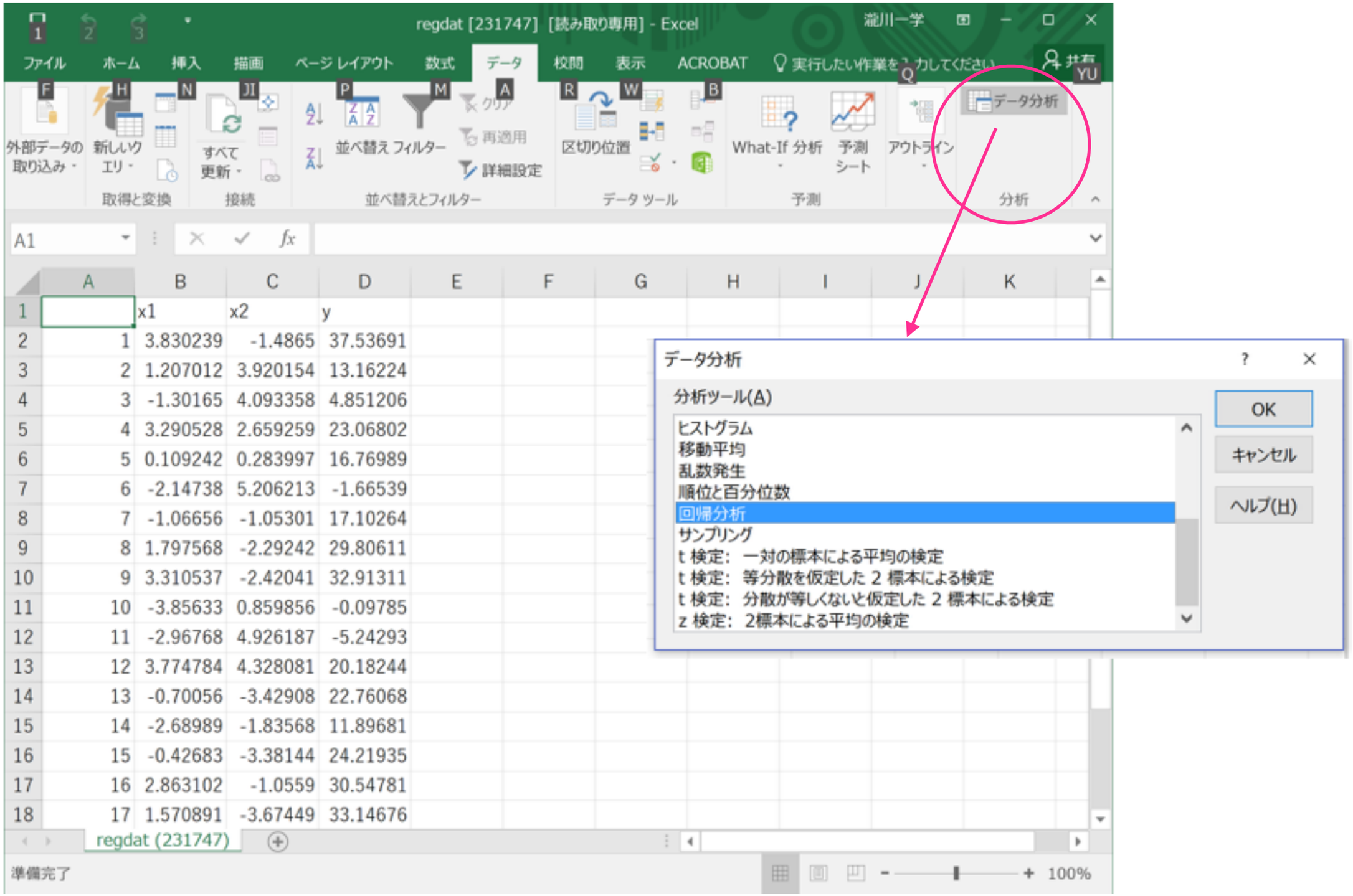
	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 90.0%	上限 90.0%
切片	17.65376	0.26662	66.21323	5.88238E-22	17.09123714	18.21627423	17.18994214	18.11756923
X 値 1	3.618766	0.104205	34.72736	3.15336E-17	3.398912451	3.83861929	3.437490079	3.800041661
X 値 2	-2.38141	0.074075	-32.1486	1.14906E-16	-2.537694528	-2.225124783	-2.510271258	-2.252548053

Excel(2016)で回帰分析をする場合

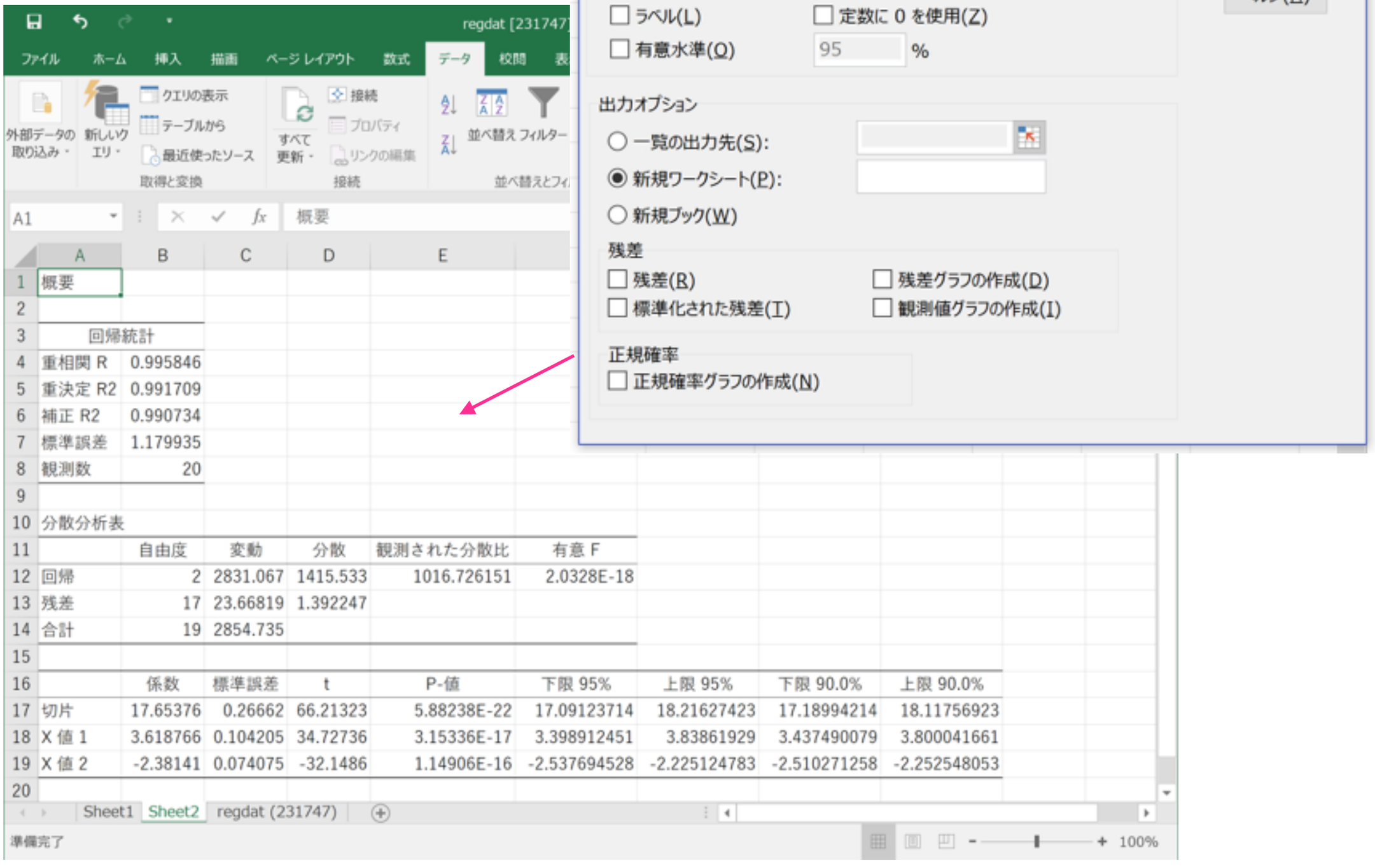
x1	x2	y
3.830	−1.487	37.537
1.207	3.920	13.162
−1.302	4.093	4.851
3.291	2.659	23.068
0.109	0.284	16.770
−2.147	5.206	−1.665
−1.067	−1.053	17.103
1.798	−2.292	29.806
3.311	−2.420	32.913
−3.856	0.860	−0.098
−2.968	4.926	−5.243
3.775	4.328	20.182
−0.701	−3.429	22.761
−2.690	−1.836	11.897
−0.427	−3.381	24.219
2.863	−1.056	30.548
1.571	−3.674	33.147
2.254	−0.160	25.374
−4.490	−9.077	23.642
0.702	−3.250	27.714

説明変数　　目的変数





	A	B	C	D
1		x1	x2	y
2	1	3.830239	-1.4865	37.53691
3	2	1.207012	3.920154	13.16224
4	3	-1.30165	4.093358	4.851206
5	4	3.290528	2.659259	23.06802
6	5	0.109242	0.283997	16.76989
7	6	-2.14738	5.206213	-1.66539
8	7	-1.06656	-1.05301	17.10264
9	8	1.797568	-2.29242	29.80611
10	9	3.310537	-2.42041	32.91311
11	10	-3.85633	0.859856	-0.09785
12	11	-2.96768	4.926187	-5.24293
13	12	3.774784	4.328081	20.18244
14	13	-0.70056	-3.42908	22.76068
15	14	-2.68989	-1.83568	11.89681
16	15	-0.42683	-3.38144	24.21935
17	16	2.863102	-1.0559	30.54781
18	17	1.570891	-3.67449	33.14676



	A	B	C	D	E
1	概要				
2					
3	回帰統計				
4	重相関 R	0.995846			
5	重決定 R2	0.991709			
6	補正 R2	0.990734			
7	標準誤差	1.179935			
8	観測数	20			
9					
10	分散分析表				
11		自由度	変動	分散	観測された分散比	有意 F
12	回帰	2	2831.067	1415.533	1016.726151	2.0328E-18
13	残差	17	23.66819	1.392247		
14	合計	19	2854.735			
15						
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 90.0%	上限 90.0%
17	切片	17.65376	0.26662	66.21323	5.88238E-22	17.09123714	18.21627423	17.18994214	18.11756923
18	X 値 1	3.618766	0.104205	34.72736	3.15336E-17	3.398912451	3.83861929	3.437490079	3.800041661
19	X 値 2	-2.38141	0.074075	-32.1486	1.14906E-16	-2.537694528	-2.225124783	-2.510271258	-2.252548053
20									

Rでの回帰分析

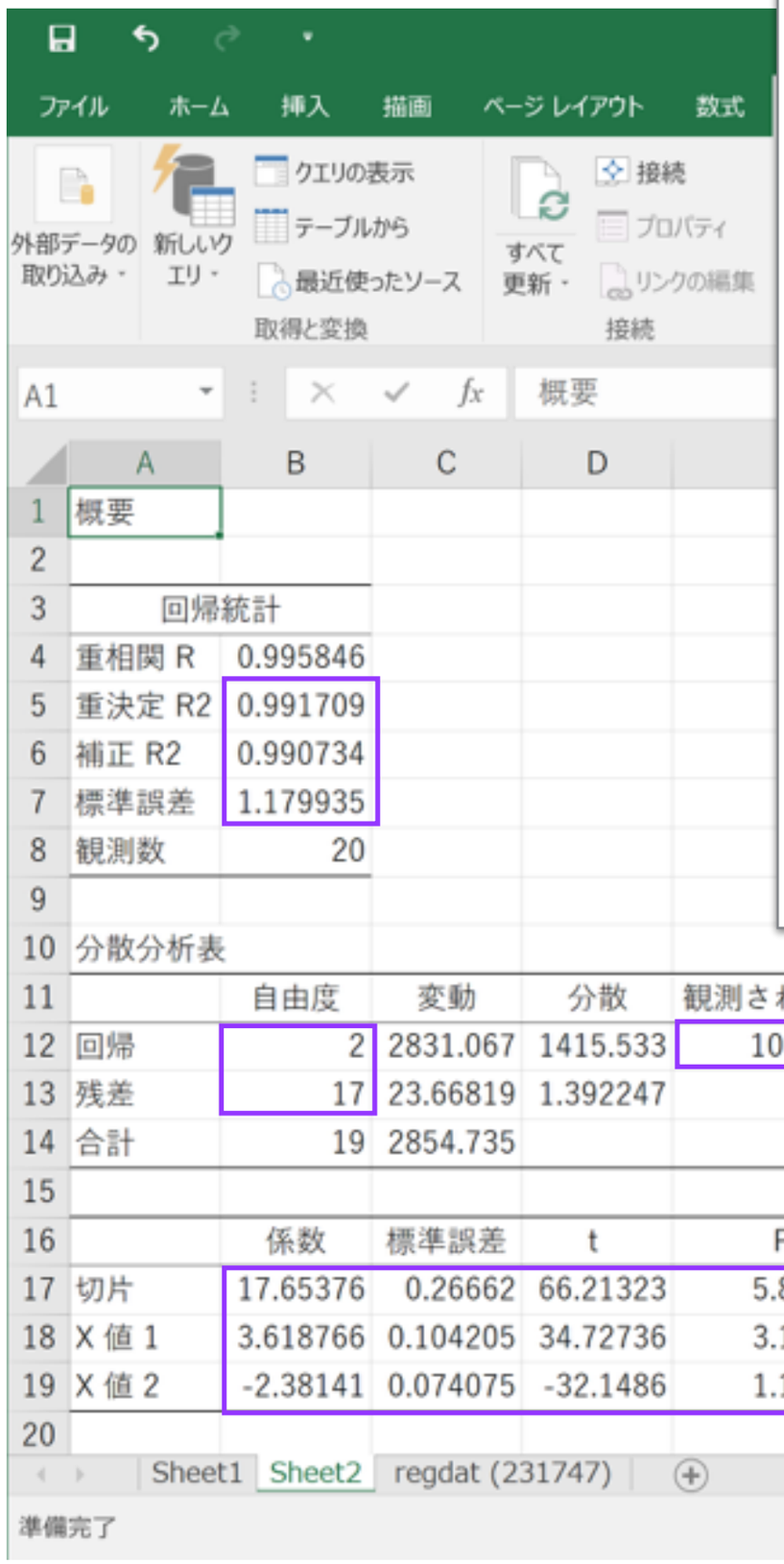
```
> res <- lm(y~X)
> summary(res)

Call:
lm(formula = y ~ as.matrix(X))

Residuals:
     Min       1Q   Median       3Q      Max
-2.48469 -0.54371 -0.07098  0.66454  2.48245

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.65376    0.26662   66.21   <2e-16 ***
x1            3.61877    0.10421   34.73   <2e-16 ***
x2           -2.38141    0.07408  -32.15   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.18 on 17 degrees of freedom
Multiple R-squared:  0.9917,    Adjusted R-squared:  0.9907
F-statistic:  1017 on 2 and 17 DF,  p-value: < 2.2e-16
```



	A	B	C	D
1	概要			
2				
3	回帰統計			
4	重相関 R	0.995846		
5	重決定 R2	0.991709		
6	補正 R2	0.990734		
7	標準誤差	1.179935		
8	観測数	20		
9				
10	分散分析表			
11		自由度	変動	分散	観測された分散比	有意 F
12	回帰	2	2831.067	1415.533	1016.726151	2.0328E-18
13	残差	17	23.66819	1.392247		
14	合計	19	2854.735			
15						
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 90.0%	上限 90.0%
17	切片	17.65376	0.26662	66.21323	5.88238E-22	17.09123714	18.21627423	17.18994214	18.11756923
18	X 値 1	3.618766	0.104205	34.72736	3.15336E-17	3.398912451	3.83861929	3.437490079	3.800041661
19	X 値 2	-2.38141	0.074075	-32.1486	1.14906E-16	-2.537694528	-2.225124783	-2.510271258	-2.252548053
20									

空欄を埋めてみよう (p値は特殊なので後回しでOK)

x1	x2	y
-1.8	-5.5	13.4
0.4	4.2	-11.5
1.8	-0.2	0.1
0.1	-1.8	5
-5.4	7.7	-22

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 0.231 & 0.031 & -0.001 \\ 0.031 & 0.041 & 0.010 \\ -0.001 & 0.010 & 0.012 \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} -0.46 \\ 0.22 \\ -2.64 \end{bmatrix}$$

$$\hat{\sigma}^2 = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 = 0.337$$

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = 775.15$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = 775.82$$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 0.675$$

$t(2)$ 信頼度 95% 有意水準 5% 境界値 4.303

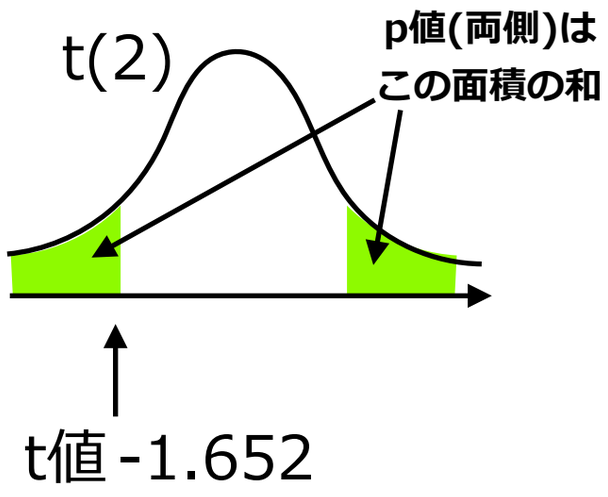
	係数	標準誤差	t値	(p値)	下限95%	上限95%
切片						
x1						
x2						

回帰統計	
重相関 R	
重決定 R^2	
補正 R^2	
標準誤差	
観測数	

https://www.wolframalpha.com/

簡易電卓がわり
(ルートや四則演算)

t値が-1.652の
ときのp値の算出



```
> x1 <- c(-1.8,0.4,1.8,0.1,-5.4)
> x2 <- c(-5.5,4.2,-0.2,-1.8,7.7)
> y <- c(13.4,-11.5,0.1,5,-22)
> res <- lm(y~cbind(x1,x2))
> summary(res)

Call:
lm(formula = y ~ cbind(x1, x2))

Residuals:
       1        2        3        4        5
-0.26798 -0.03541 -0.36156  0.68606 -0.02111

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.46128    0.27920  -1.652 0.240311
cbind(x1, x2)x1  0.21927    0.11723   1.870 0.202339
cbind(x1, x2)x2 -2.64072    0.06323 -41.763 0.000573 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5809 on 2 degrees of freedom
Multiple R-squared:  0.9991, Adjusted R-squared:  0.9983
F-statistic:  1149 on 2 and 2 DF,  p-value: 0.0008699
```

Rの実行結果
Excel版は各自やってみてね

```
> confint(res)
                               2.5 %     97.5 %
(Intercept)     -1.6625879  0.7400286
cbind(x1, x2)x1 -0.2851356  0.7236830
cbind(x1, x2)x2 -2.9127791 -2.3686577
```

多変量版の計算だけいちおう確認

単変量

期待値

$$\mathbb{E}\{X\} = \int x p(x) dx$$

分散

$$\begin{aligned}\mathrm{var}\{X\} &= \mathbb{E}\{(X - \mathbb{E}\{X\})^2\} \\ &= \int (x - \mathbb{E}\{X\})^2 p(x) dx\end{aligned}$$

共分散

$$\begin{aligned}\mathrm{cov}(X, Y) &= \mathbb{E}\{(X - \mathbb{E}\{X\})(Y - \mathbb{E}\{Y\})\} \\ &= \int\int (x - \mathbb{E}\{X\})(y - \mathbb{E}\{Y\}) p(x, y) dx dy\end{aligned}$$

多変量

期待値ベクトル

$$\mathbb{E}\{\boldsymbol{x}\} = \begin{bmatrix} \mathbb{E}\{X_1\} \\ \mathbb{E}\{X_2\} \\ \vdots \\ \mathbb{E}\{X_p\} \end{bmatrix}$$

分散共分散行列

$$\mathrm{var}\{\boldsymbol{x}\} = \mathbb{E}\{(\boldsymbol{x} - \mathbb{E}\{\boldsymbol{x}\})(\boldsymbol{x} - \mathbb{E}\{\boldsymbol{x}\})'\}$$

$$= \begin{bmatrix} \mathrm{var}\{X_1\} & \mathrm{cov}(X_1, X_2) & \cdots & \mathrm{cov}(X_1, X_p) \\ \mathrm{cov}(X_2, X_1) & \mathrm{var}\{X_2\} & \cdots & \mathrm{cov}(X_2, X_p) \\ \vdots & & & \vdots \\ \mathrm{cov}(X_p, X_1) & \mathrm{cov}(X_p, X_2) & \cdots & \mathrm{var}\{X_p\} \end{bmatrix}$$

正規線形モデル

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

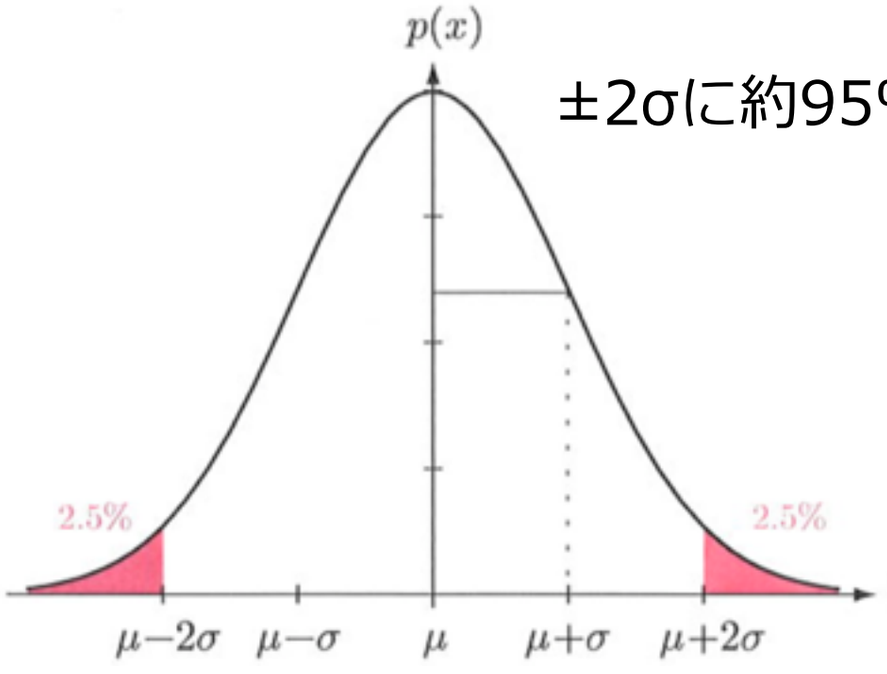
$\boldsymbol{X}$：計画行列

$$\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

$$\Leftrightarrow \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & & & \vdots \\ 1 & x_{p1} & \cdots & x_{pp} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \sim N(0, \sigma^2)$$

$$\Leftrightarrow \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \right)$$



±2σに約95%

定理 4.2　$\boldsymbol{X}$ が k 次元正規分布 $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ に従うとき

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{a},$$

$\boldsymbol{A}$ は $p \times k$ の定数行列，$\boldsymbol{a}$ は p 次元定数ベクトル

とおくと，$\boldsymbol{Y}$ は $N_p(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{a}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}')$ に従う．

定理 7.4

（i） $E(\hat{\beta})=\beta$，すなわち $\hat{\beta}$ は β の不偏推定量である．

（ii） $V(\hat{\beta})=\sigma^2(X'X)^{-1}$.

証明

（i） $$E(\hat{\beta})=E(X'X)^{-1}X'y)=(X'X)^{-1}X'\cdot E(y)$$
$$=(X'X)^{-1}X'\cdot X\beta=\beta$$

（ii） $$V(\hat{\beta})=V((X'X)^{-1}X'y)$$
$$=(X'X)^{-1}X'\cdot V(y)\cdot X(X'X)^{-1}$$
$$=(X'X)^{-1}X'(\sigma^2 I_n)X(X'X)^{-1}$$
$$=\sigma^2(X'X)^{-1}X'X(X'X)^{-1}$$
$$=\sigma^2(X'X)^{-1}. \quad \square$$

定理 7.5

$$s^2:=\frac{\|y-X\hat{\beta}\|^2}{n-m-1}$$

は σ^2 の不偏推定量である．

少しだけ一般化　www.is.titech.ac.jp/~shimo/class/dk2005/san07.pdf

- 回帰係数の線形結合を考え，その性質を調べる

$$v = w_0\beta_0 + w_1\beta_1 + \cdots + w_p\beta_p$$

$$\hat{v} = w_0\hat{\beta}_0 + w_1\hat{\beta}_1 + \cdots + w_p\hat{\beta}_p$$

- $\boldsymbol{w} = (w_0, w_1, \ldots, w_p)$ を設定することにより，

$$w_j = 1, w_k = 0, k \neq j \quad \Rightarrow \quad \hat{v} = \hat{\beta}_j$$

偏回帰係数の推定量

$$w_0 = x_{i0}, \ldots w_p = x_{ip} \quad \Rightarrow \quad \hat{v} = \hat{y}_i$$

予測値(母回帰)の推定量

などが表現できるので便利.

- ベクトル表現

$$v = \boldsymbol{w}'\boldsymbol{\beta}, \quad \hat{v} = \boldsymbol{w}'\hat{\boldsymbol{\beta}}$$

- 回帰係数は多変量正規分布に従う

$$\hat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$$

従って，$\hat{v}$ は正規分布に従う

$$\hat{v} \sim N(\boldsymbol{w}'\boldsymbol{\beta}, \sigma^2\boldsymbol{w}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{w})$$

- 結局

$$\hat{v} \sim N(v, \sigma_v^2)$$

ただし

$$v = \boldsymbol{w}'\boldsymbol{\beta}, \quad \sigma_v^2 = \sigma^2 \boldsymbol{w}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{w}$$

- σ^2 を S_ϵ^2 で推定すると，

$$\hat{\sigma}_v^2 = S_\epsilon^2 \boldsymbol{w}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{w} = S_\epsilon^2 \frac{\sigma_v^2}{\sigma^2} = \frac{\|\boldsymbol{e}\|^2/\sigma^2}{n-p-1}\sigma_v^2$$

つまり

$$(n-p-1)\frac{\hat{\sigma}_v^2}{\sigma_v^2} \sim \chi_{n-p-1}^2$$

- 以上より，

$$\frac{\hat{v}-v}{\sigma_v} \Big/ \sqrt{\frac{\hat{\sigma}_v^2}{\sigma_v^2}} \sim t\text{-分布}_{n-p-1}$$

式を整理すると，

$$\frac{\hat{v}-v}{\hat{\sigma}_v} \sim t\text{-分布}_{n-p-1}$$

- 参考) 正規分布する変数を線形変換しても正規分布する

証明「数理統計学ー基礎から学ぶデータ解析ー(鈴木武・山田作太郎著)」p.121 定理4.2

定理 4.2 $\boldsymbol{X}$ が k 次元正規分布 $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ に従うとき

$$\boldsymbol{Y}=\boldsymbol{A}\boldsymbol{X}+\boldsymbol{a},$$

$\boldsymbol{A}$ は $p\times k$ の定数行列，$\boldsymbol{a}$ は p 次元定数ベクトル

とおくと，$\boldsymbol{Y}$ は $N_p(\boldsymbol{A}\boldsymbol{\mu}+\boldsymbol{a}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}')$ に従う．

証明 $\boldsymbol{t}=(t_1, \cdots, t_p)'$ に対して

$$M_Y(\boldsymbol{t})=E(e^{\boldsymbol{t}'(\boldsymbol{A}\boldsymbol{X}+\boldsymbol{a})})=e^{\boldsymbol{t}'\boldsymbol{a}}E(e^{\boldsymbol{t}'\boldsymbol{A}\boldsymbol{X}})$$

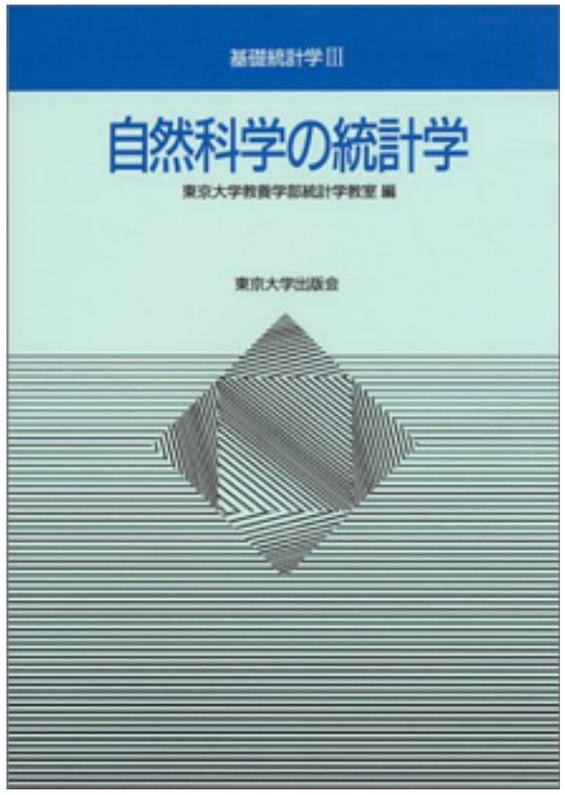
$$=e^{\boldsymbol{t}'\boldsymbol{a}}\exp\left\{\boldsymbol{t}'\boldsymbol{A}\boldsymbol{\mu}+\frac{1}{2}\boldsymbol{t}'\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'\boldsymbol{t}\right\}$$

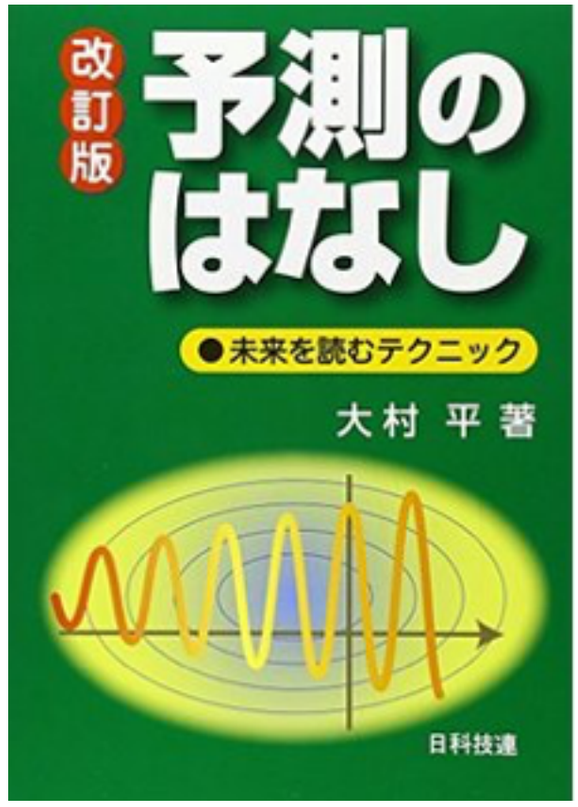
$$=\exp\left\{\boldsymbol{t}'(\boldsymbol{A}\boldsymbol{\mu}+\boldsymbol{a})+\frac{1}{2}\boldsymbol{t}'\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'\boldsymbol{t}\right\}$$

を得る．積率母関数は確率分布と1対1に対応するので，(4.51)より $\boldsymbol{Y}$ は p 次元正規分布 $N_p(\boldsymbol{A}\boldsymbol{\mu}+\boldsymbol{a}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}')$ に従う． □

定理 4.2 では $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'$ は対称行列ではあるが，必ずしも正値定符号ではない．$p\leqq k$ のとき $\boldsymbol{Y}$ の分布が(4.45)の意味での p 次元正規分布であるためには，$\boldsymbol{A}$ のランク(rank)が p であればよい．

さらに勉強するための参考文献





+

「データ解析」(下平英寿) 講義資料

http://www.is.titech.ac.jp/~shimo/class/gakubu200503.html

【見直しシート】

x1	x2	y
3.830	−1.487	37.537
1.207	3.920	13.162
−1.302	4.093	4.851
3.291	2.659	23.068
0.109	0.284	16.770
−2.147	5.206	−1.665
−1.067	−1.053	17.103
1.798	−2.292	29.806
3.311	−2.420	32.913
−3.856	0.860	−0.098
−2.968	4.926	−5.243
3.775	4.328	20.182
−0.701	−3.429	22.761
−2.690	−1.836	11.897
−0.427	−3.381	24.219
2.863	−1.056	30.548
1.571	−3.674	33.147
2.254	−0.160	25.374
−4.490	−9.077	23.642
0.702	−3.250	27.714

説明変数　　目的変数

(1) 不偏推定量をつくる

$$\mathbb{E}\{\hat{\theta}\} = \theta$$ ←真の値！

(2) 標準化から次が言える

$$\frac{\hat{\theta} - \mathbb{E}\{\hat{\theta}\}}{\sqrt{\mathrm{var}(\hat{\theta})}} \sim N(0,1)$$

(3) 不偏標本分散を代入

$$\frac{\hat{\theta} - \mathbb{E}\{\hat{\theta}\}}{\sqrt{\widehat{\mathrm{var}}(\hat{\theta})}} \sim t(\nu)$$

ν : $\widehat{\mathrm{var}}(\hat{\theta})$ の自由度

【見直しシート】

$$\frac{\hat{\theta} - \mathbb{E}\{\hat{\theta}\}}{\sqrt{\widehat{\mathrm{var}}(\hat{\theta})}} \sim t(\nu)$$

← 真の値を含み他は構成可能な量について分布が同定されている！

ν : $\widehat{\mathrm{var}}(\hat{\theta})$ の自由度

□ は検定やるにも区間推定やるにも計算する必要あり

$\hat{\theta}$ ↓　$\sqrt{\widehat{\mathrm{var}}(\hat{\theta})}$ ↓

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 90.0%	上限 90.0%
切片	17.65376	0.26662	66.21323	5.88238E-22	17.09123714	18.21627423	17.18994214	18.11756923
X 値 1	3.618766	0.104205	34.72736	3.15336E-17	3.398912451	3.83861929	3.437490079	3.800041661
X 値 2	-2.38141	0.074075	-32.1486	1.14906E-16	-2.537694528	-2.225124783	-2.510271258	-2.252548053

定理 7.4

（i） $E(\hat{\beta})=\beta$，すなわち $\hat{\beta}$ は β の不偏推定量である．

（ii） $V(\hat{\beta})=\sigma^2(X'X)^{-1}$．

証明

（i）
$$E(\hat{\beta})=E(X'X)^{-1}X'y)=(X'X)^{-1}X'\cdot E(y)$$
$$=(X'X)^{-1}X'\cdot X\beta=\beta$$

（ii）
$$V(\hat{\beta})=V((X'X)^{-1}X'y)$$
$$=(X'X)^{-1}X'\cdot V(y)\cdot X(X'X)^{-1}$$
$$=(X'X)^{-1}X'(\sigma^2 I_n)X(X'X)^{-1}$$
$$=\sigma^2(X'X)^{-1}X'X(X'X)^{-1}$$
$$=\sigma^2(X'X)^{-1}. \quad \square$$

定理 7.5

$$s^2:=\frac{\|y-X\hat{\beta}\|^2}{n-m-1}$$

は σ^2 の不偏推定量である．