数理統計 補助資料
～実証のための回帰モデル～

2024年度2学期：　月曜1限, 水曜3限
担当教員：　石垣　司

回帰分析による「実証」

仮説や理論をデータを用いて統計的に検証する

- ある説明変数が目的変数の変動へ影響を与えているかどうかをデータから検証したい
  - 例：最高気温は消費電力に影響を与えているか？
  - 例：広告費を上げると商品の売り上げは増えるか？
  - 例：リフレッシュ休暇をとると生産性は上がるか？

目標：表中の数値（仮説検定の結果）の意味を正しく解釈できる

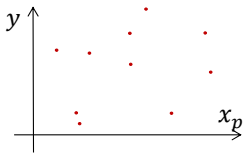
	Estimate (推定値)	Std.Error (標準誤差)	t value (t値)	Pr(>|t|) (p値)
切片 (b_0)	106146	24196	4.39	0.000***
年齢 (b_1)	841	382	2.21	0.028*
家族人数 (b_2)	23170	2602	8.91	0.000***
高齢者の有無 (b_3)	-1063	8202	-0.13	0.897
子供の有無 (b_4)	7941	7633	1.04	0.299
家からの時間(b_5)	-3208	598	-5.37	0.000**
Adjusted R-squared (自由度調整済み決定係数)		0.106		

回帰係数の検定の意義

説明変数が目的変数に影響を与えるか否かの検定

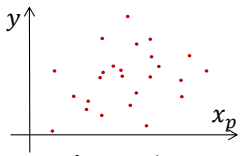
$$y = b_0 + b_1 x_1 +, \cdots, +b_P x_P$$

- 説明変数 x_p が y に影響を与える ⇔ 回帰係数 $b_p \neq 0$
- 説明変数 x_p が y に影響を与えない ⇔ 回帰係数 $b_p = 0$

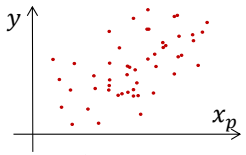


サンプルサイズ $N = 10$

$b_p \neq 0$ or $b_p = 0$ ？？？



サンプルサイズ $N = 25$



サンプルサイズ $N = 50$

$b_p \neq 0$ に思えるなぁ

- 最小2乗推定量 $\hat{b}_p$ の分散は母集団からのサンプリングとサンプルサイズ (N) に依存

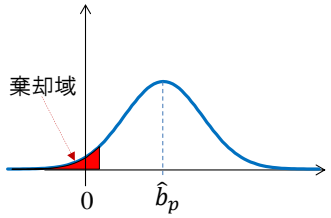
回帰係数の検定の手続きのイメージ

帰無仮説 $H_0 : b_p = 0$

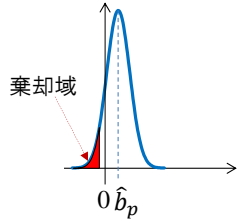
対立仮説 $H_1 : b_p \neq 0$

$\boldsymbol{b} = [b_0, \cdots, b_p, \cdots, b_P]^T$ を真の係数の値
$\hat{\boldsymbol{b}} = [\hat{b}_0, \cdots, \hat{b}_p, \cdots, \hat{b}_P]^T$ を最小2乗推定量として表記する

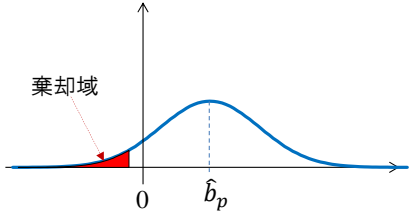
- 検定統計量を適切に定めて，検定統計量の実現値が棄却域に入るかどうかを検定する
- 直感的なイメージ：最小2乗推定量 $\hat{b}_p$ の分布を推定して，$\hat{b}_p = 0$ の点が棄却域にあるか調べることと同様



$b_p = 0$ である確率は統計的に有意に小さい



$b_p = 0$ である確率は統計的に小さいとは言えない



$b_p = 0$ である確率は統計的に小さいとは言えない

ベクトルの期待値の表記法

$\hat{\boldsymbol{b}} = [\hat{b}_0, \cdots, \hat{b}_p, \cdots, \hat{b}_P]^T$ のとき,

$$E[\hat{\boldsymbol{b}}] = \left[E[\hat{b}_0], \cdots, E[\hat{b}_p], \cdots, E[\hat{b}_P]\right]^T$$

$$V[\hat{\boldsymbol{b}}] = E\left[(\hat{\boldsymbol{b}} - \boldsymbol{E}[\hat{\boldsymbol{b}}])(\hat{\boldsymbol{b}} - \boldsymbol{E}[\hat{\boldsymbol{b}}])^T\right]$$

$$= \begin{bmatrix} V[\hat{b}_0] & \cdots & Cov[\hat{b}_0\hat{b}_p] & \cdots & Cov[\hat{b}_0\hat{b}_P] \\ \vdots & \ddots & \vdots & & \vdots \\ Cov[\hat{b}_0\hat{b}_p] & \cdots & V[\hat{b}_p] & \cdots & Cov[\hat{b}_p\hat{b}_P] \\ \vdots & & \vdots & \ddots & \vdots \\ Cov[\hat{b}_0\hat{b}_P] & \cdots & Cov[\hat{b}_p\hat{b}_P] & \cdots & V[\hat{b}_P] \end{bmatrix}$$

定数行列 A に対する公式: $V[A\hat{\boldsymbol{b}}] = AV[\hat{\boldsymbol{b}}]A^T$

– **これ以降は, ベクトルと行列で回帰係数の検定を考える**

ベクトルと行列に慣れるよい機会ととらえてほしい

回帰係数の検定の準備 #1

仮定1: 説明変数は確率変数ではなく定数である
仮定2: 説明変数間に多重共線性はない
仮定3: 誤差項 e は平均0, 分散 σ^2 に従う確率変数であり,
$\{e_1, \cdots, e_N\}$ は互いに独立である($V[\boldsymbol{e}] = \sigma^2 I$)

仮定1, 2, 3 の下で成り立つ命題 check!

1. **最小2乗推定量 $\hat{\boldsymbol{b}}$ は不偏推定量となる** ($E[\hat{\boldsymbol{b}}] = \boldsymbol{b}$)
2. **最小2乗推定量 $\hat{\boldsymbol{b}}$ の分散(条件付き分散)は** $V[\hat{\boldsymbol{b}}] = \sigma^2(X^TX)^{-1}$
3. Gauss-Markov **の定理:**
   最小2乗推定量 $\hat{\boldsymbol{b}}$ は最小分散線形不偏推定量となる

多変量正規分布

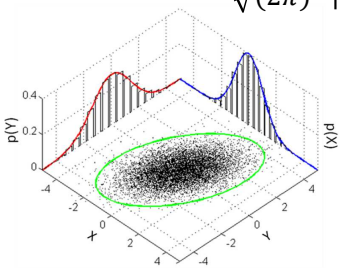
正規分布を高次元化した確率分布 $N(\boldsymbol{\mu}, V)$

$$\boldsymbol{y} \sim N(\boldsymbol{\mu}, V)$$

$\boldsymbol{y}$: M 次元確率変数ベクトル, $\boldsymbol{\mu}$: M 次元の平均ベクトル
V: $M \times M$ 次元の分散共分散行列

– **確率密度関数**

$$f(\boldsymbol{y}; \boldsymbol{\mu}, V) = \frac{1}{\sqrt{(2\pi)^M |V|}} \exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^T V^{-1} (\boldsymbol{y} - \boldsymbol{\mu})\right\}$$



Wikipedia 多変量正規分布

例: $\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 3/5 \\ 3/5 & 2 \end{bmatrix}\right)$

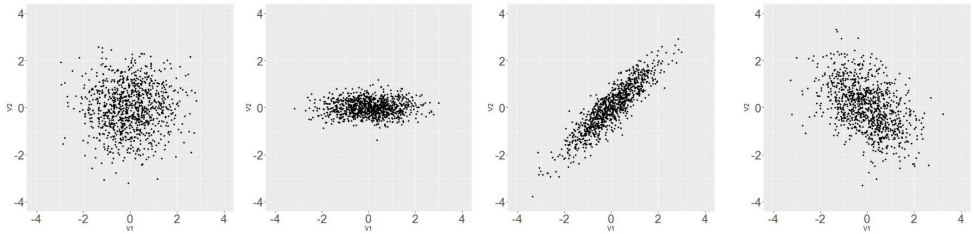
多変量正規分布と分散共分散行列

多変量正規分布の分散共分散行列 $V = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1M} \\ \vdots & \ddots & \vdots \\ \sigma_{1M} & \cdots & \sigma_M^2 \end{bmatrix}$

σ_i^2: **変数 i の分散**, σ_{ij}: **変数 i と j の共分散**

– **平均ベクトル** $\boldsymbol{\mu} = [0 \quad 0]^T$ **でそれぞれの分散共分散行列 Σ の多変量正規分布から, それぞれ1000個の乱数を発生**

$V = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ $V = \begin{bmatrix} 1 & 0 \\ 0 & 0.1 \end{bmatrix}$ $V = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$ $V = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 0.5 \end{bmatrix}$



回帰係数の検定の準備 #2

仮定4: 誤差項 e は正規分布 $N(0,\sigma^2)$ に従う

仮定1, 2, 3, 4 の下で成り立つ命題

1. $\hat{\boldsymbol{b}} \sim N(\boldsymbol{b}, \sigma^2 (X^T X)^{-1})$

2. $\sigma^2 (X^T X)^{-1} = \begin{bmatrix} \sigma^2 d_1^2 & & \\ & \ddots & \\ & & \sigma^2 d_p^2 \end{bmatrix}$ と書くと, $\dfrac{\hat{b}_p - b_p}{\sqrt{\sigma^2 d_p^2}} \sim N(0,1)$

3. 誤差の標本分散を $S^2 = \dfrac{1}{N-P-1}\sum_{i=1}^{N} e_i^2$ とすると,

$$\frac{\hat{b}_p - b_p}{\sqrt{S^2 d_p^2}} \sim t^{(N-P-1)}$$

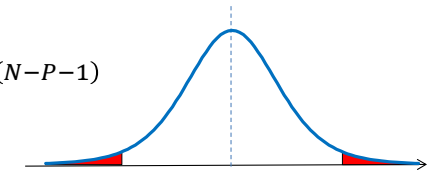
←これを検定統計量として採用する

#2,3が分からない場合、t 分布の導入と性質を復習してほしい

回帰係数の検定

仮定1, 2, 3, 4 の下での, 個別の回帰係数の検定

- **帰無仮説** H_0 : $b_p = 0$
- **対立仮説** H_1 : $b_p \neq 0$
- **検定統計量:** $T = \dfrac{\hat{b}_p}{\sqrt{S^2 \lambda_p^2}} \sim t^{(N-P-1)}$



自由度 $N-P-1$ の t 分布

仮定1, 2, 3, 4 の下での, 回帰係数全体の F 検定

- **帰無仮説** H_0 : $b_1 =, \ldots, = b_P = 0$
- **対立仮説** H_1 : b_1 から b_P のどれかがゼロではない

重回帰分析の結果の解釈 #1

	Estimate (推定値)	Std.Error (標準誤差)	t value (t値)	Pr(>|t|) (p値)	
切片 (b_0)	106146	24196	4.39	0.000	***
年齢 (b_1)	841	382	2.21	0.028	*
家族人数 (b_2)	23170	2602	8.91	0.000	***
高齢者の有無 (b_3)	-1063	8202	-0.13	0.897	
子供の有無 (b_4)	7941	7633	1.04	0.299	
家からの時間(b_5)	-3208	598	-5.37	0.000	**
Adjusted R-squared (自由度調整済み決定係数)		0.106			

赤字部分の解釈

「年齢(b_1), 家族人数(b_2), 家からの距離(b_5)は, 有意水準5%で統計的に有意に購買金額に影響を与えている」

- **高齢者の有無(b_3), 子供の有無(b_4)は購買金額に影響を与えているかどうかは分からない**

重回帰分析の結果の解釈 #2

	Estimate (推定値)	Std.Error (標準誤差)	t value (t値)	Pr(>|t|) (p値)	
切片 (b_0)	106146	24196	4.39	0.000	***
年齢 (b_1)	841	382	2.21	0.028	*
家族人数 (b_2)	23170	2602	8.91	0.000	***
高齢者の有無 (b_3)	-1063	8202	-0.13	0.897	
子供の有無 (b_4)	7941	7633	1.04	0.299	
家からの時間(b_5)	-3208	598	-5.37	0.000	**
Adjusted R-squared (自由度調整済み決定係数)		0.106			

誤った赤字部分の解釈

- **高齢者の有無(b_3), 子供の有無(b_4)は購買金額に影響を与えていない**
- **家族人数(b_2)のP値が一番低いので, 家族人数が最も強く購買金額に影響を与えている**
- **有意水準を1%に設定すると, 年齢(b_1)が有意ではないので有意水準5%の方が良い分析結果である**

重回帰分析の結果の解釈 #3

	Estimate (推定値)	Std.Error (標準誤差)	t value (t値)	Pr(>|t|) (p値)
切片 (b_0)	106146	24196	4.39	0.000***
年齢 (b_1)	841	382	2.21	0.028*
家族人数 (b_2)	23170	2602	8.91	0.000***
高齢者の有無 (b_3)	-1063	8202	-0.13	0.897
子供の有無 (b_4)	7941	7633	1.04	0.299
家からの時間(b_5)	-3208	598	-5.37	0.000**
Adjusted R-squared (自由度調整済み決定係数)		0.106		

赤字部分の年齢の解釈

「$b_2, \ldots, b_5$ の影響を取り除いた場合，年齢(b_1)が1歳上がるごとに購買金額が841円大きくなる」

「b_1, b_3, b_4, b_5 の影響を取り除いた場合，家族の人数が1人増えるごとに購買金額が23,170円大きくなる」

「$b_1, \ldots, b_4$ の影響を取り除いた場合，家からの所要時間(b_5)が1分増えるごとに購買金額が3,208円小さくなる」

重回帰分析の結果の解釈（総合）

	Estimate (推定値)	Std.Error (標準誤差)	t value (t値)	Pr(>|t|) (p値)
切片 (b_0)	106146	24196	4.39	0.000***
年齢 (b_1)	841	382	2.21	0.028*
家族人数 (b_2)	23170	2602	8.91	0.000***
高齢者の有無 (b_3)	-1063	8202	-0.13	0.897
子供の有無 (b_4)	7941	7633	1.04	0.299
家からの時間(b_5)	-3208	598	-5.37	0.000**
Adjusted R-squared (自由度調整済み決定係数)		0.106		

- 購買金額には年齢，家族の人数，家からの所要時間が影響を与えている
- 関数の当てはまりには改善の余地がある
  - 当然，購買は今回利用した顧客属性以外の要因に基づいて変化しうる。購買金額を予測するためには，よりよい説明変数の追加などのモデルの改善が必要である
  - 一方，予測が目的ではない場合，決定係数は重要視する必要はない

仮定の検証の例（スーパーマーケットデータ）#1

仮定1：説明変数は確率変数ではなく定数である

- とりあえず OK として話を進める
  （ただし，後述のように実データの多くはこの仮定を満たさない）

仮定2：説明変数間に多重共線性はない

- VIF値は目安より低い

```
> vif(Reg)
     Age   Family      Old    Child     Time
2.325950 1.292024 1.669825 1.441859 1.019022
```

仮定3：誤差項 e は平均0，分散 σ^2 に従う確率変数であり $\{e_1, \cdots, e_N\}$ は互いに独立である

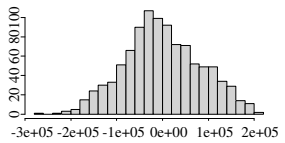
- 顧客 i の購買行動が顧客 j に影響を与えるとは考え難い
  (Durbin-Watson比 1.983。値が2に近いと誤差項の自己相関無し)

仮定の検証の例（スーパーマーケットデータ）#2

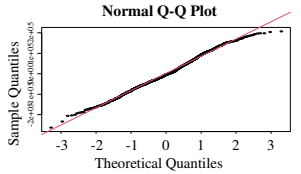
仮定4：誤差項 e は正規分布 $N(0, \sigma^2)$ に従う

- 正規性のチェック

```
> qqnorm(Reg$residuals)
> qqline(Reg$residuals, col=2)
```



誤差項のヒストグラム



誤差項のQ-Qプロット

- 均一分散のチェック

White の方法
（不均一分散頑健推定量）
→最小2乗推定とほぼ同じ結果

```
> coeftest(Reg, df=Inf, vcov=vcovHC(Reg,type="HC3"))

z test of coefficients:

             Estimate Std. Error z value  Pr(>|z|)
(Intercept) 106146.90   24834.74  4.2741 1.919e-05 ***
Age            841.78     385.69  2.1825   0.02907 *
Family       23170.59    2709.92  8.5503 < 2.2e-16 ***
Old          -1063.13    8508.23 -0.1250   0.90056
Child         7941.49    8079.54  0.9829   0.32565
Time         -3208.10     597.96 -5.3651 8.092e-08 ***
```

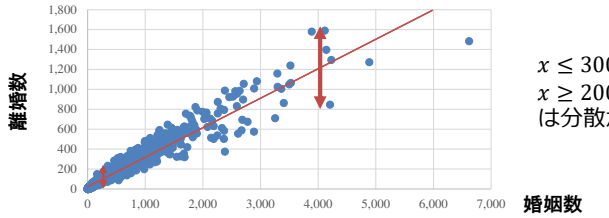
→ 仮定4を強く否定する結果ではない

↑この文章の意味は授業中に説明する。誤差項の正規性を仮説検定する方法もあり。回帰係数の検定の仮定のチェックは「回帰診断」などとよばれる。

不均一分散（仮定3の分散σ^2の均一性を満たさない例）

サンプル i 毎に誤差項 e_i の分散が異なる事象

- **例：所得と消費額，失業者数と犯罪発生率などの関係では，所得や都市の規模が大きいほど分散が大きくなる傾向**



$x \leq 300$ 付近と
$x \geq 2000$ 付近で
は分散が異なる

総務省統計局：統計でみる市区町村のすがた2015より1877市町村の婚姻と離婚数

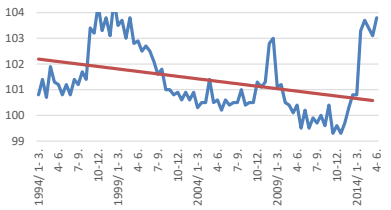
- **不均一分散のあるデータの最小2乗推定量の性質**
  最小2乗推定量の不偏性や検定結果は？ check!
- **対応方法：加重最小2乗推定や一般化最小2乗推定を学習**

系列相関（仮定3の誤差項間の独立性を満たさない例）

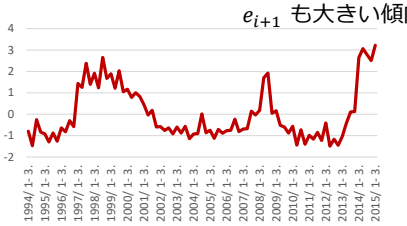
異なる点の誤差項 e_i, e_j に相関がある事象

- **例：時系列など過去の値が現在の値に影響を与えるデータ**
  今期の政策の効果は次期以降に徐々に現れる

e_i が大きければ
e_{i+1} も大きい傾向



消費者物価指数（CPI）の時系列データと回帰直線



残差系列

- **系列相関のあるデータの最小2乗推定量の性質**
  最小2乗推定量の不偏性や検定結果は？ check!
- **対応方法：時系列分析を学習**

補足：仮定1が成り立たないときの検定

計量経済学の学部上級生向けの内容を含むため、雰囲気だけを紹介。興味のある学生は条件付き期待値や繰り返し期待値の法則などから学習を始めてほしい

説明変数が確率変数の場合の最小2乗推定量 $\hat{b}$ の性質

- **仮定N1：誤差項 e はすべての説明変数と互いに独立**
- **説明変数 X と目的変数 y，最小2乗推定量 $\hat{b}$ 誤差項 e の関係は，それぞれ条件付き期待値 $E[y|X], E[\hat{b}|X], E[e|X]$ で表現する**

性質：条件付き期待値に関して仮定 2, 3, 4, N1 を満たすとき，最小2乗推定量 $\hat{b}$ はGauss-Markov の定理を満たす

- **仮定 1 が成り立たない場合でも，仮定 N1 が成り立てば最小2乗推定量はよい性質をもった推定量といえる**

内生性（仮定N1を満たさない例）#1

説明変数と誤差項の間に相関がある事象

内生性の例1：観測できない説明変数

- **v_i：サンプル i に対して観測できない属性変数**
- **例：学歴 x_i から給与 y_i を説明する回帰モデル。v_i は個人 i の観測できない能力で学歴と給与の両方と相関がある**

より正確なモデル
$y_i = b_0 + b_1 x_i + b_2 v_i + e_i$

我々が作成できるモデル
$y_i = b_0 + b_1 x_i + v_i e_i$

- **x_i が大きいと $v_i e_i$ も大きい傾向**

これを誤差項 e_i として観測してしまう

内生性（仮定N1を満たさない例）#2

内生性の例2：目的変数が説明変数に影響を与えている

例：従軍経験とその後の賃金は関係あるのか？ (Angrist 1990)

満足できる職に就けなかった人，つまり，元々賃金が低くなる傾向にある人たちが軍隊に入る傾向があるのでは？

例：警察予算を増額すると犯罪件数は減るか？ (Levitt 1997)

前年度の犯罪発生数に基づき予算を決める

$$賃金 = a + b_1 軍隊経験 + \cdots + b_p x_p + e$$

– 内生性（同時方程式）バイアスの結果

最小2乗推定量の不偏性や検定結果は？ check!

自然実験，操作変数法などを学習

J. D. Angrist, Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records, The American Economic Review, Vol. 80, No. 3, pp. 313-336, (1990)
S.D. Levitt, Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime, Ameri. Econ. Review, 87(3), 270–90 (1997)

演習問題

1. 不均一分散の誤差項は $e_i \sim N(0, \sigma_i^2)$ と表現できる。不均一分散があるデータに対して $E[\hat{b}]$ と $V[\hat{b}]$ を求めなさい。ただし，説明変数は非確率変数とする
   – $\hat{b} = b + (X^T X)^{-1} X^T e$ から $E[\hat{b}]$ と $V[\hat{b}]$ を計算
2. 系列相関があるデータに対して $E[\hat{b}]$ と $V[\hat{b}]$ を求めなさい。ただし，説明変数は非確率変数とする
3. 説明変数を確率変数とする。このとき，内生性のある現象から生じたデータに対して $E[\hat{b}]$ を求めて最小2乗推定量が不偏性をもつか確かめなさい