数理統計 補助資料
～実証のための線形回帰モデル～

2023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大きくなるか？
  - 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			

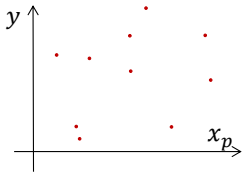
回帰係数の検定の意義

説明変数が目的変数に影響を与えるか否かの検定

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

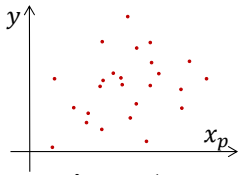
- 説明変数 x_p が y に影響を与える ⇔ 回帰係数 $b_p \neq 0$
- 説明変数 x_p が y に影響を与えない ⇔ 回帰係数 $b_p = 0$

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

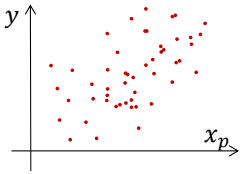


サンプルサイズ $N = 10$

$b_p \neq 0$ 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

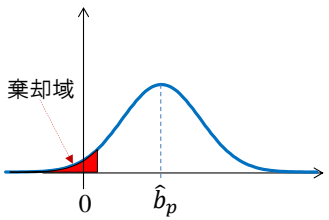
回帰係数の検定の手続きのイメージ

帰無仮説 $H_0: b_p = 0$

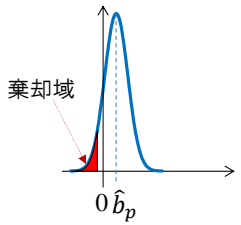
対立仮説 $H_1: b_p \neq 0$

$\boldsymbol{b} = [b_0, b_1, \cdots, b_p, \cdots, b_P]^T$ を真の係数の値
$\hat{\boldsymbol{b}} = [\hat{b}_0, \hat{b}_1, \cdots, \hat{b}_p, \cdots, \hat{b}_P]^T$ を最小2乗推定量として表記する

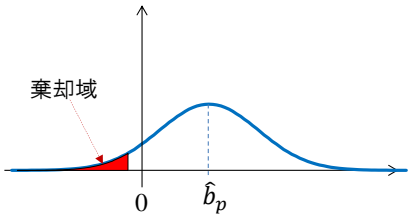
- 検定統計量を適切に定めて, 検定統計量の実現値が棄却域に入るかどうかを検定する
- 直感的なイメージ: 最小2乗推定量 $\hat{b}_p$ の分布を推定して, $\hat{b}_p = 0$ の点が棄却域にあるか調べることと同様



$b_p = 0$ 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

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 はすべての説明変数と互いに独立である
仮定4：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

仮定1, 2, 3, 4 の下で成り立つ命題

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

回帰係数の検定の準備 #2

仮定5：誤差項 e は正規分布 $N(0, \sigma^2)$ に従う

仮定1, 2, 3, 4, 5 の下で成り立つ命題

1. $\hat{\boldsymbol{b}} \sim N(\boldsymbol{b}, \sigma^2(X^TX)^{-1})$
2. $(X^TX)^{-1} = \begin{bmatrix} \lambda_1^2 & & & \\ & \ddots & & \\ & & \lambda_p^2 & \\ & & & \ddots \\ & & & & \lambda_P^2 \end{bmatrix}$ **と書くと,** $\dfrac{\hat{b}_p - b_p}{\sqrt{\sigma^2\lambda_p^2}} \sim N(0,1)$
3. **誤差の標本分散を** $S^2 = \frac{1}{N-P-1}\sum_{i=1}^{N} e_i^2$ **とすると,**

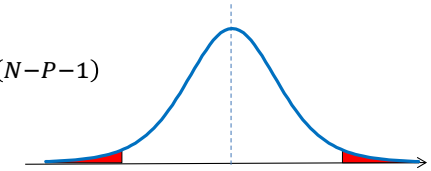
$$\frac{\hat{b}_p - b_p}{\sqrt{S^2\lambda_p^2}} \sim t^{(N-P-1)}$$

←これを検定統計量として採用する

回帰係数の検定

仮定1, 2, 3, 4, 5 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：** $T = \dfrac{\hat{b}_p - b_p}{\sqrt{S^2\lambda_p^2}} \sim t^{(N-P-1)}$



$T = 0$
自由度 $N - P - 1$ の t 分布

仮定1, 2, 3, 4, 5 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

スーパーマーケットデータでの仮定の検証 #1

仮定1: 説明変数は確率変数ではなく定数である → OKとする

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 はすべての説明変数と互いに独立である

→ 説明変数との相関は低く, 多重共線性は小さい

```
> cor(Reg$residuals,Data[,2:6])
              Age          Family             Old           Child            Time
[1,] 2.09557e-16 -9.306995e-17 -4.885401e-17 -4.178021e-17 -1.328542e-16
```

仮定4: 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

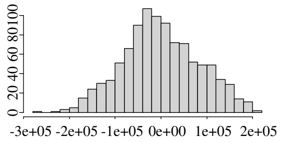
(Durbin-Watson比 1.983。値が2に近いと誤差項の自己相関無し)

スーパーマーケットデータでの仮定の検証 #2

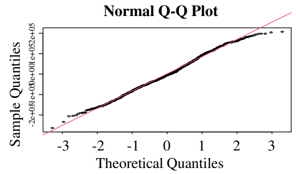
仮定5: 誤差項 e は正規分布 $N(0,\sigma^2)$ に従う

– **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5を強く否定する結果ではない

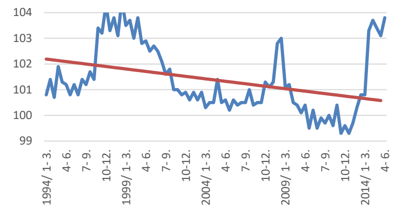
↑この文章の意味は授業中に説明する。誤差項の正規性を仮説検定する方法もあり。回帰係数の検定の仮定のチェックは「回帰診断」などとよばれる。

系列相関（仮定4の誤差項間の独立性を満たさない例）

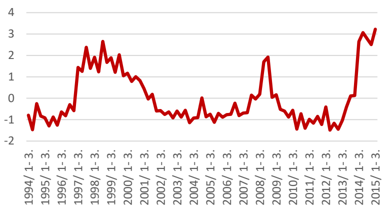
時系列データのほとんどは誤差項は非独立

– **過去の値が現在の値に影響を与える**

例：今期の政策の効果は次期以降に徐々に現れる



消費者物価指数（CPI）の時系列データと回帰直線



残差系列

– **系列相関の結果**

最小2乗推定量は不偏性と一致性を持つ

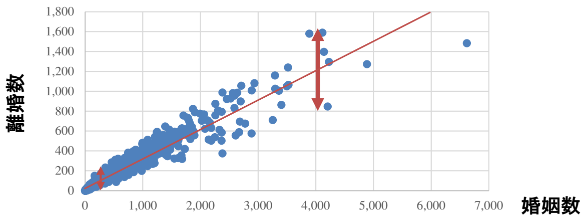
分散を過小評価の傾向 ⇒ 係数の検定は誤った結果を出す

自己相関を分析対象とする時系列分析を学習

不均一分散（仮定5の分散σ^2の均一性を満たさない例）

「所得と消費額」,「失業者数と犯罪発生率」に関係はあるか？

– **所得や都市の規模が大きいほど分散が大の傾向**



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

– **不均一分散の結果**

最小2乗推定量は不偏性と一致性を持つ

分散を過小評価の傾向 ⇒ 係数の検定は誤った結果を出す

加重最小2乗法や一般化最小2乗法などを学習

内生性（仮定3を満たさない例）

従軍経験とその後の賃金は関係あるのか？(Angrist 1990)

– **従軍経験者と非経験者の賃金を比べる ⇒ 不十分**

そもそも, 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？

?

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

?

– **内生性（同時方程式）バイアスの結果**

最小2乗推定量は不偏性も一致性も持たない

自然実験, 操作変数法などを学習

※(Angrist 1990)**では米国ベトナム戦争の影響を研究。単純な線形回帰分析では白人男性従軍経験者の賃金低下の推定値は** 2-3%**。一方, 操作変数法により内生性の影響を取り除いたときの賃金低下の推定値は** 15%

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)

演習問題

「仮定1：説明変数は確率変数ではなく定数である」が，なぜ必要となるのかをもう少し詳しく考えてみよう。

仮定1, 2, 3, 4 の下で最小2乗推定量の期待値が

$$E\left[\widehat{\boldsymbol{b}}\right] = \boldsymbol{b}$$

となることを示しなさい。

ヒント：

- **最小2乗推定量** $\widehat{\boldsymbol{b}} = (X^T X)^{-1} X^T \boldsymbol{y}$
- **線形回帰モデル** $\boldsymbol{y} = X\boldsymbol{b} + \boldsymbol{e}$