2022年3月26日
第24回春の合宿セミナー（日本行動計量学会）
（統計的因果推論入門）

講義4
重回帰分析による交絡因子の統制の意味

長崎大学 情報データ科学部 准教授
高橋 将宜
博士（理工学）
m-takahashi@nagasaki-u.ac.jp

概要

- 回帰分析の復習
- 講義1の復習
- 三変数のバレンティン・ベン図
- 三変数の重回帰モデル
- 共分散分析（再考）
- 重回帰モデルによる分析

教科書Ch.5

教科書Ch.6

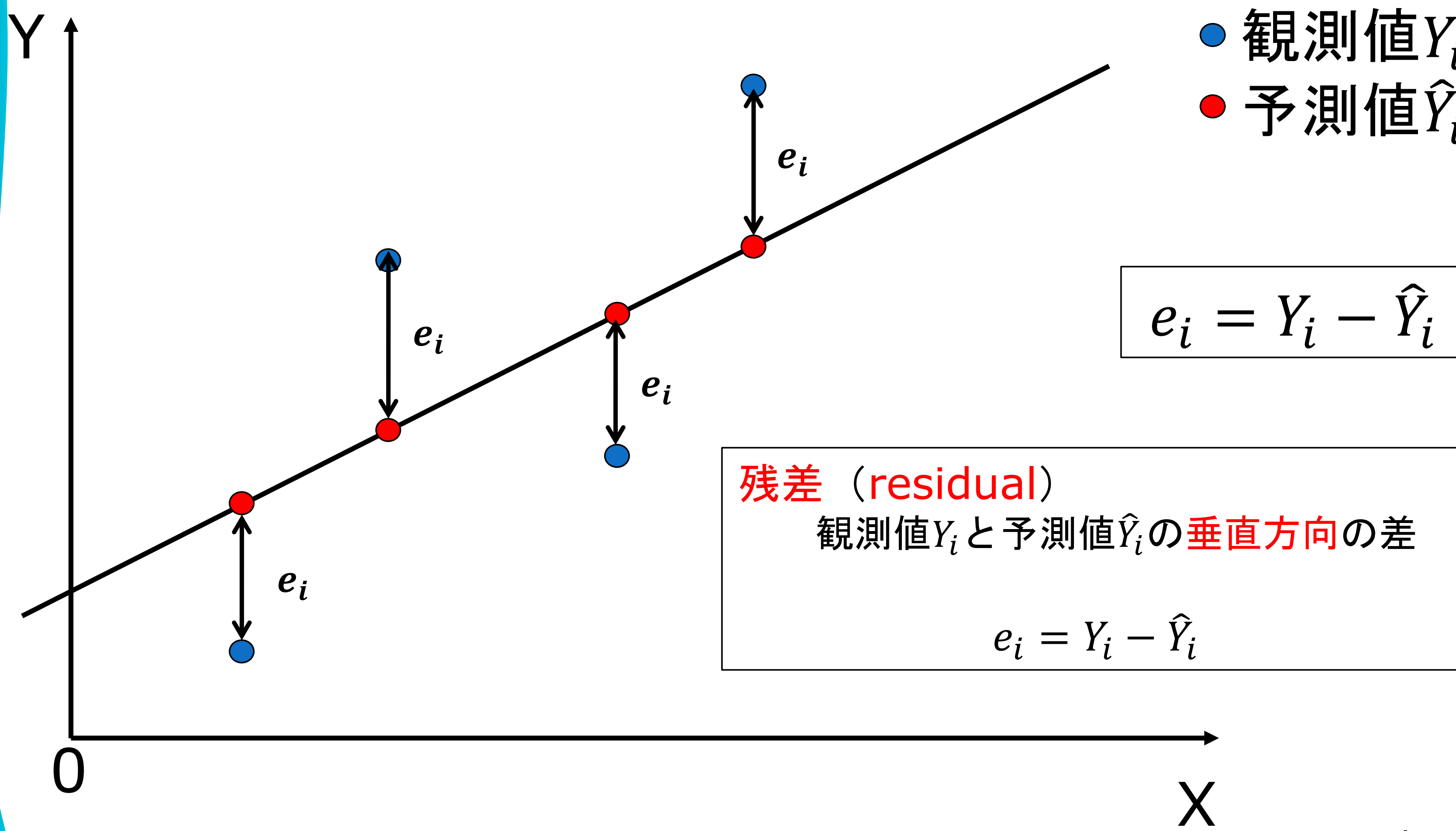
回帰分析の復習

母集団と標本における回帰式

母集団	標本
$Y_i = \alpha + \beta X_i + \varepsilon_i$	$Y_i = \hat{\alpha} + \hat{\beta} X_i + e_i$
$E[Y_i \mid X_i] = \alpha + \beta X_i$	$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$
$\varepsilon_i = Y_i - E[Y_i \mid X_i]$	$e_i = Y_i - \hat{Y}_i$

誤差項εは偶然によって生じる確率変数と考える。

残差e_i



$$e_i = Y_i - \hat{Y}_i$$

残差（residual）

観測値Y_iと予測値$\hat{Y}_i$の垂直方向の差

$$e_i = Y_i - \hat{Y}_i$$

通常の最小二乗法（OLS: Ordinary Least Squares）

- 残差平方和を最小化する切片と傾きを用いて線を引く方法

$$\hat{\beta} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

lmの構文
lm(変数1~変数2)

数式の左辺に来る変数が左側、右辺に来る変数が右側

回帰係数に関する仮説

- 帰無仮説

$$H_0\colon \beta = 0$$

- 対立仮説

$$H_A\colon \beta \neq 0$$

- 回帰式$Y_i = \alpha + \beta X_i + \varepsilon_i$において、$\beta = 0$は、$X_i$から$Y_i$への直線的な影響がないことを意味する。
  - つまり、X_iが増加しても、それにつれてY_iが増減する傾向がない。

決定係数R^2の定義

- R^2は、回帰モデルによって説明できるYの変動の割合を表す。

Y_i ： 観測値
$\bar{Y}$ ： 平均値
$\hat{Y}_i$ ： 予測値

$$R^2 = 1 - \frac{USS}{TSS}$$

$$0 \leq R^2 \leq 1$$

TSS:（総和変動）
Total Sum of Squares

$$\sum (Y_i - \bar{Y})^2$$

ESS:（回帰モデルで説明できる変動）
Explained Sum of Squares

$$\sum \left(\hat{Y}_i - \bar{Y}\right)^2$$

USS:（回帰モデルで説明できない変動）
Unexplained Sum of Squares

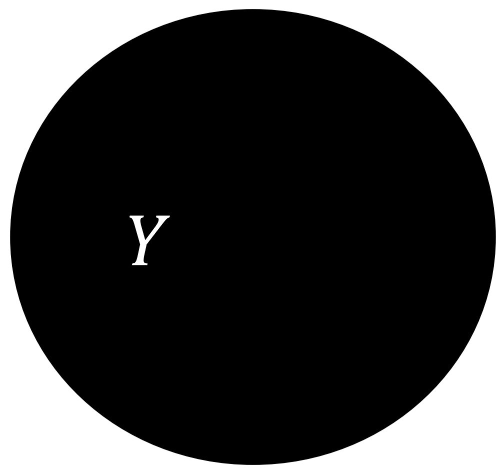
$$\sum \left(Y_i - \hat{Y}_i\right)^2$$

全変動（TSS）のバレンティン・ベン図

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_{1i}$$

$$TSS = \sum (Y_i - \bar{Y})^2$$

分散の分子と同じ

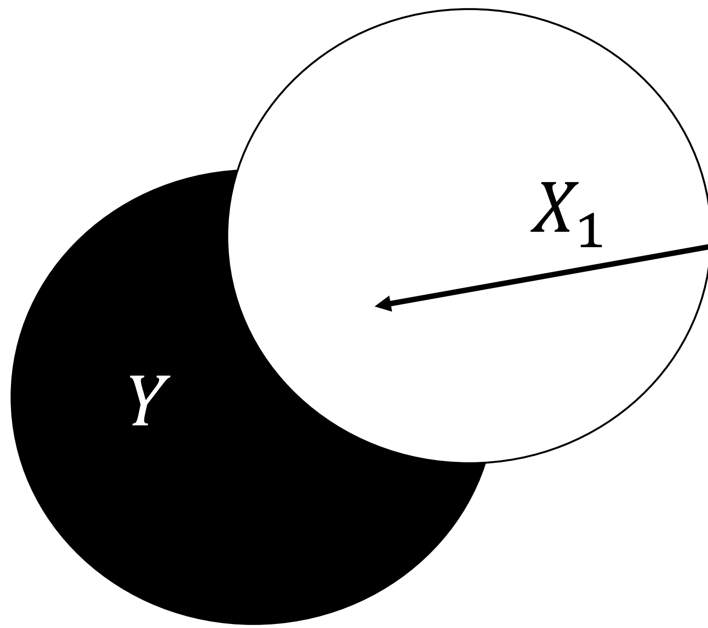


黒丸：Yの全変動（TSS）

説明できる変動（ESS）のバレンティン・ベン図

$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_{1i}$

$$ESS = \sum \left(\hat{Y}_i - \bar{Y}\right)^2$$



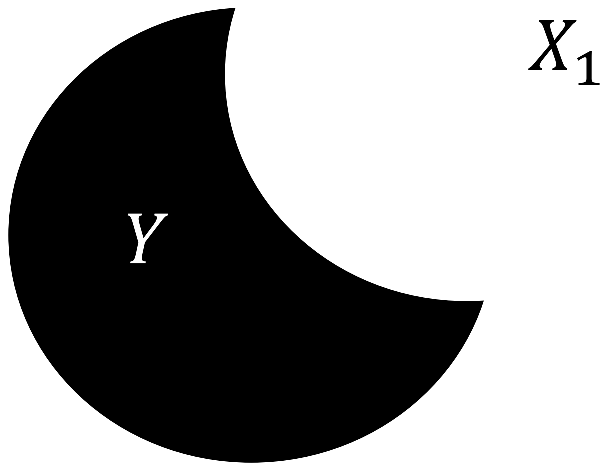
灰色の部分が、Yの説明できる変動（ESS）を表す。

説明できない変動（USS）のバレンティン・ベン図

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_{1i}$$

$$USS = \sum \left(Y_i - \hat{Y}_i\right)^2$$

残差：$e_i = Y_i - \hat{Y}_i$



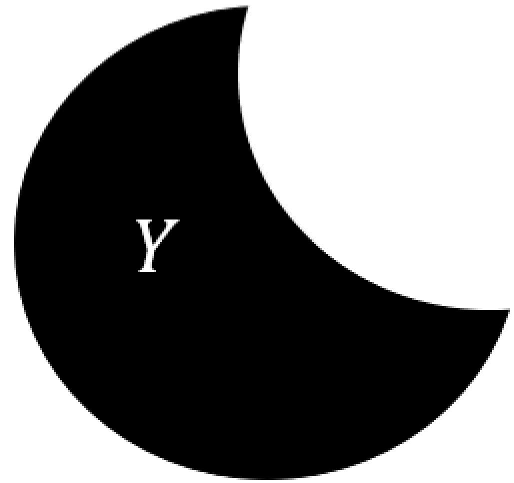
残った黒い部分が、Yの説明できない変動（USS）を表す。

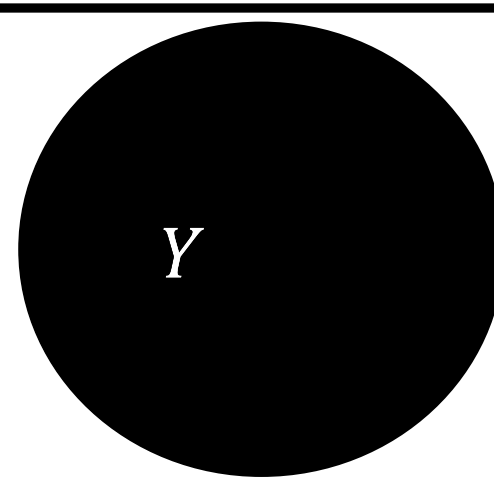
説明できる変動の割合のベン図

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_{1i}$$

$$\boxed{R^2 = 1 - \frac{USS}{TSS}}$$

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

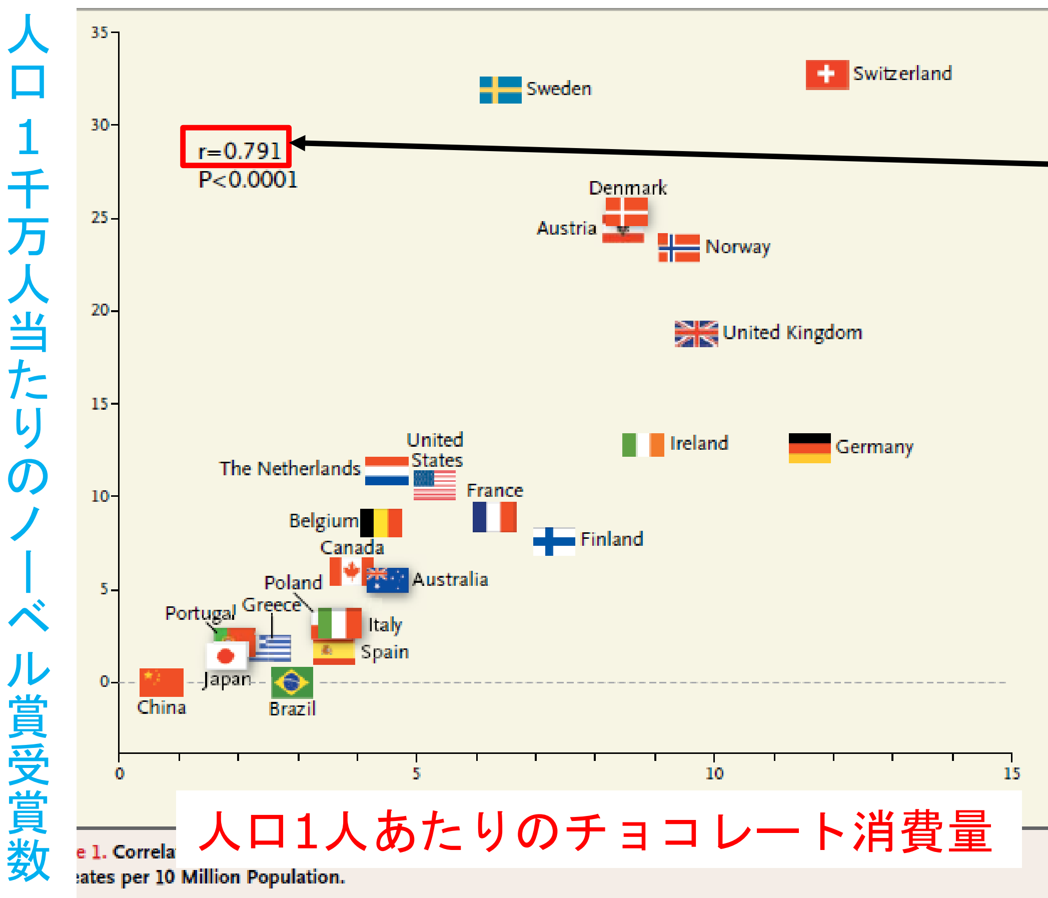




講義1の復習

散布図と相関係数

出典：Messerli, F. H. (2012) “Chocolate Consumption, Cognitive Function, and Nobel Laureates,” The New England Journal of Medicine, 367 (16), pp.1562-1564.



$r = 0.791$

単回帰モデル

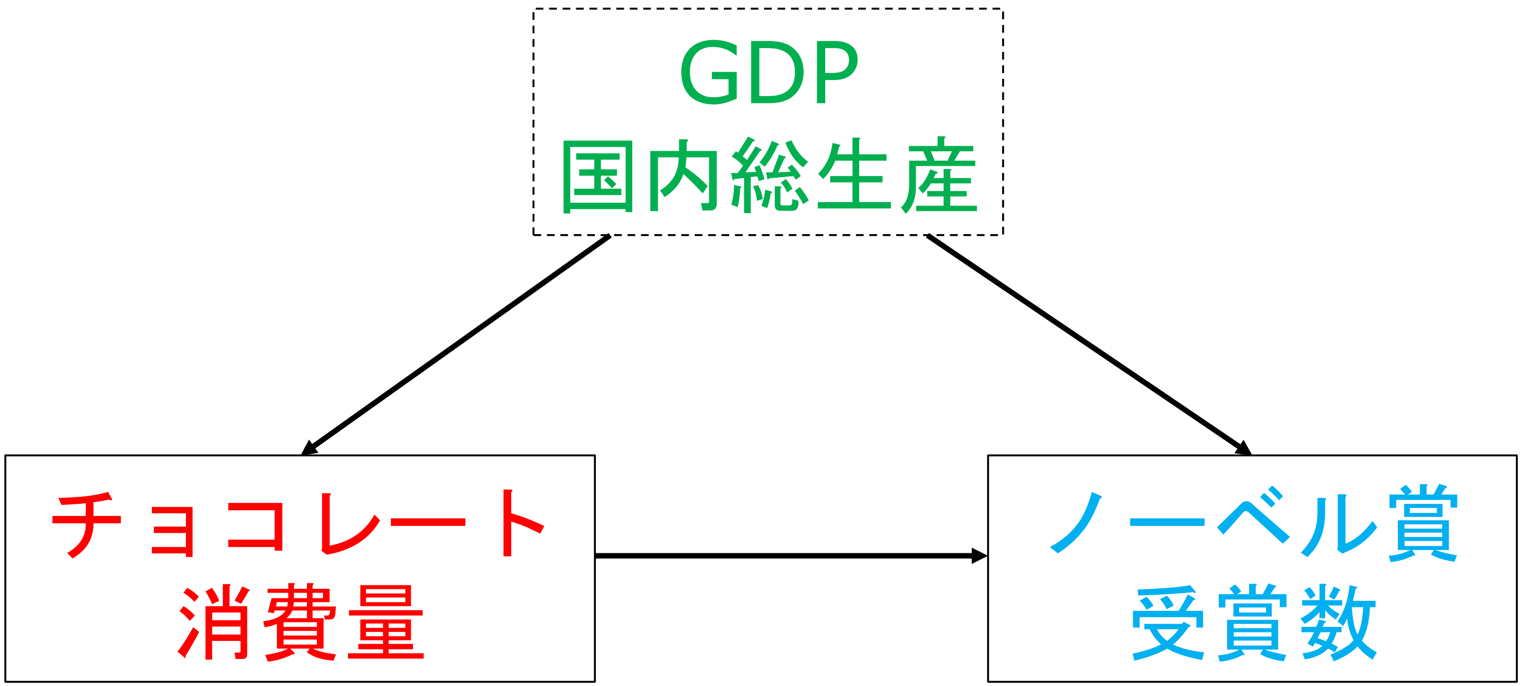
```
> summary(model1<-lm(Nobel~Choco))

Call:
lm(formula = Nobel ~ Choco)

Residuals:
     Min       1Q   Median       3Q      Max
-10.1603  -4.3915  -0.7202   2.5621  16.3355

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.4217     3.2274  -1.060 0.301096
Choco         2.7044     0.5985   4.519 0.000188 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.65 on 21 degrees of freedom
Multiple R-squared:  0.493,     Adjusted R-squared:  0.4689
F-statistic: 20.42 on 1 and 21 DF,  p-value: 0.000188

> confint(model1)
                 2.5 %   97.5 %
(Intercept) -10.133320 3.290018
Choco         1.459837 3.949020
```

疑似相関

- 相関係数の絶対値が1に近いにもかかわらず，実際には2つの現象に直接的な関係がないこと

第3の変数

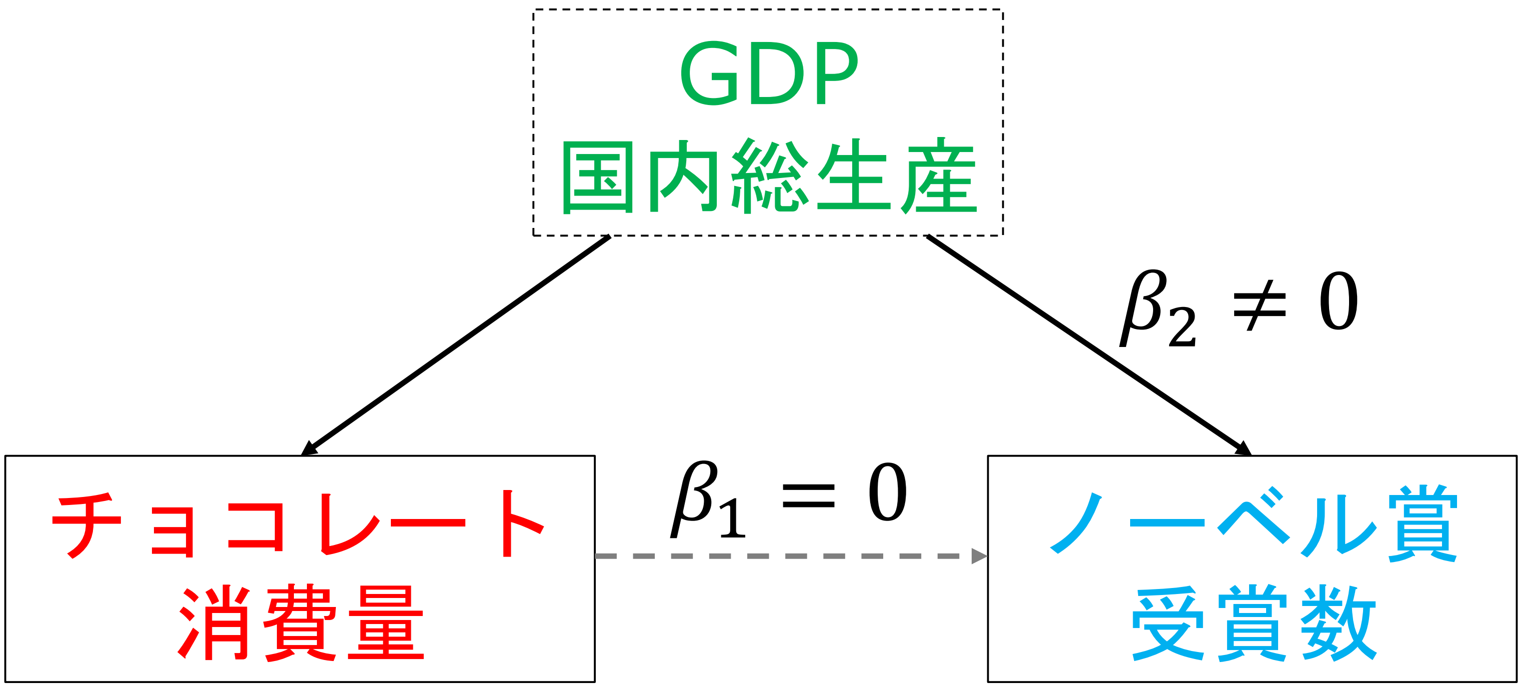
- ノーベル賞受賞数とチョコレート消費量の背後に共通の原因であるGDP（国内総生産）があると考えられる.



重回帰モデル

$$\hat{Y}_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

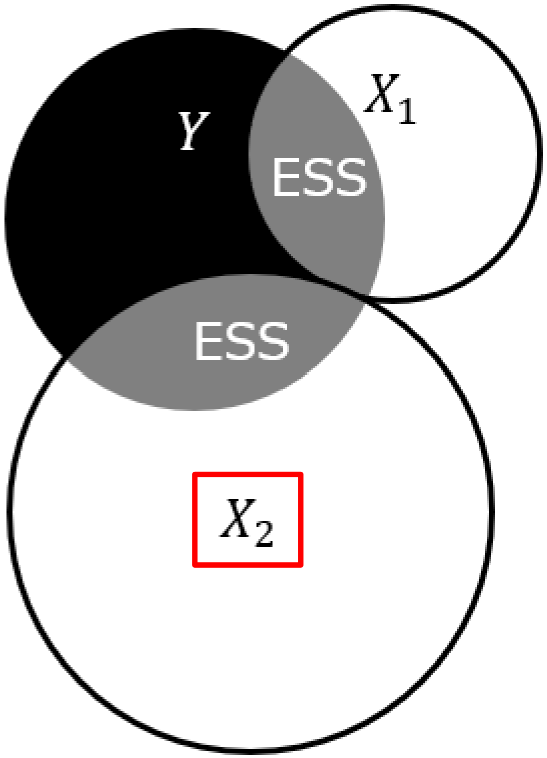
$$\widehat{\text{ノーベル賞受賞数}}_i = \beta_0 + \beta_1 \text{チョコレート消費量}_i + \beta_2 GDP_i$$

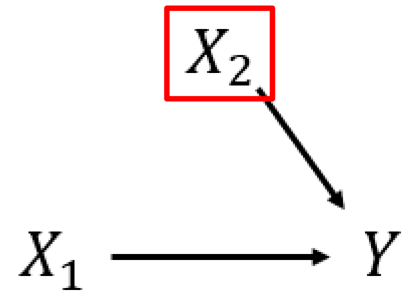


三変数のバレンティン・ベン図

X_2は共変量だが，
交絡因子ではない

説明変数の間に相関がない場合（1）





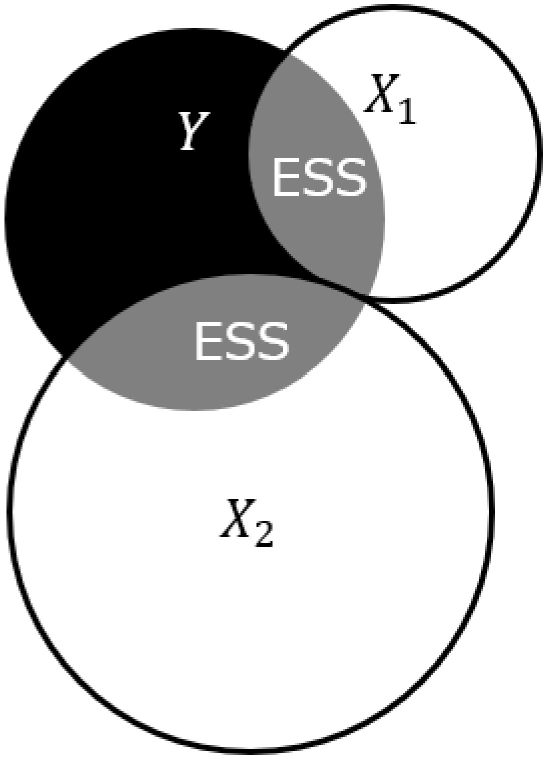
ESS:（回帰モデルで説明できる変動）
Explained Sum of Squares

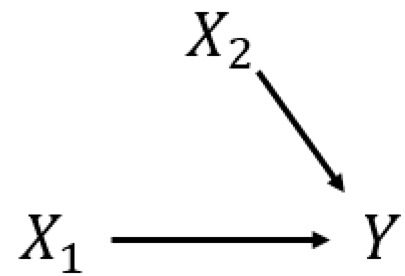
$$\sum(\hat{Y}_i - \bar{Y})^2$$

- これまでの2つの変数Yと変数X_1に加えて，3つ目の変数X_2があるとしよう．
- 変数Yと変数X_2の重なっている部分がESSである．
- 変数X_1と変数X_2には重なりがないことから，相関は0である．
- 方向付き非巡回グラフ（DAG）で表すと，X_1とX_2の間に矢印がない状態である．

X_2は共変量だが，
交絡因子ではない

説明変数の間に相関がない場合（2）

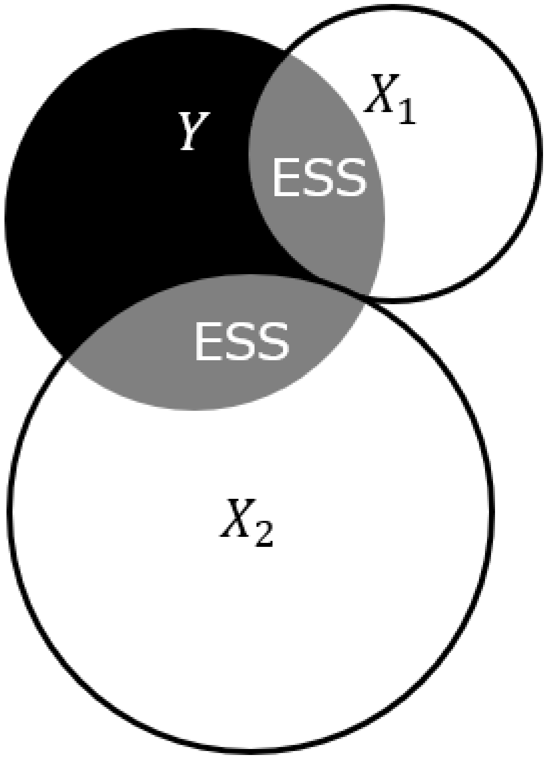


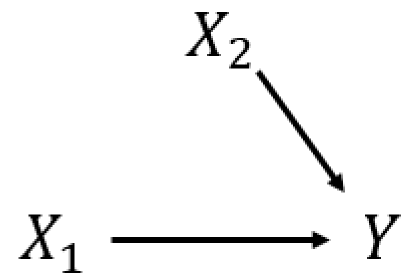


- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$における$\hat{\beta}_1$は，$X_2$からの影響を何も受けていないことから，$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i}$における$\hat{\beta}_1$と一致する．
- X_2をモデルに含めなくても，$\hat{\beta}_1$はβ_1の不偏推定量である．
- X_2は交絡因子ではない．

X_2は共変量だが，
交絡因子ではない

説明変数の間に相関がない場合（3）

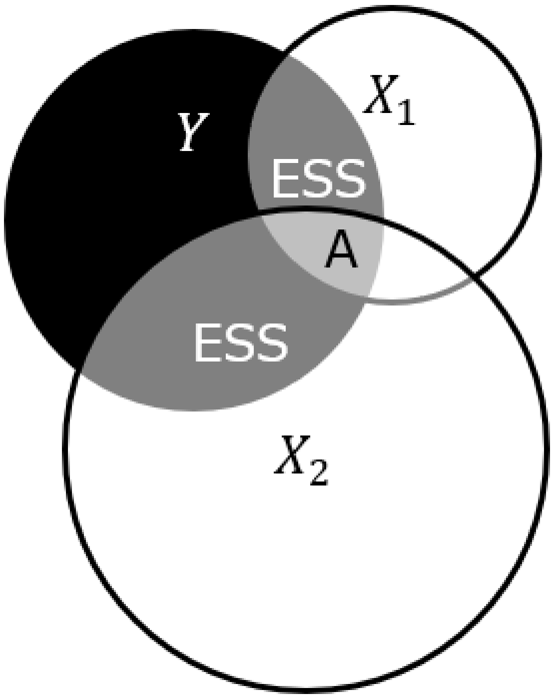


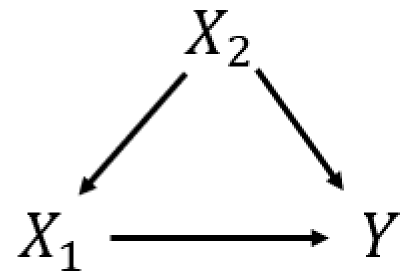


- 単回帰モデルと比べて，重回帰モデルの方が，ESSの部分が大きくなったため，Yの変動を説明する力が向上している．
- このような変数は，回帰係数の不偏性には影響を与えないが，標準誤差の大きさに影響を及ぼすため，データセット内に含まれていて利用できる状況ならば，モデルに取り込むことが望ましいが，モデルに入れなくても問題ではない．

説明変数の間に相関がある場合（1）

X_2 は共変量であり，交絡因子である．

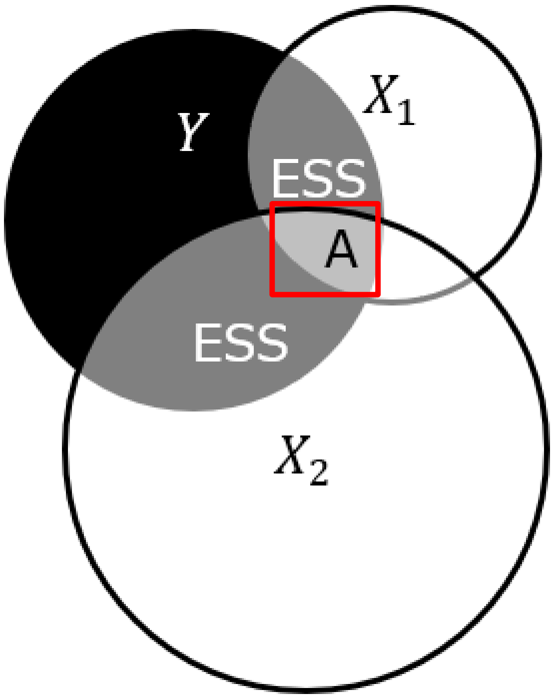


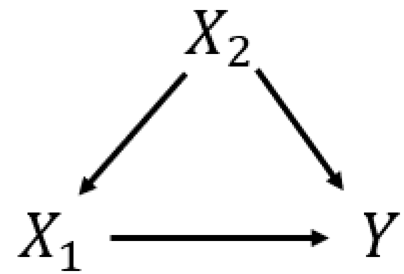


- 変数X_1と変数X_2に重なりがあるとしよう．
- DAGで表すと，X_2からX_1へ矢印がある状態である．

X_2 は共変量であり，
交絡因子である．

説明変数の間に相関がある場合（2）

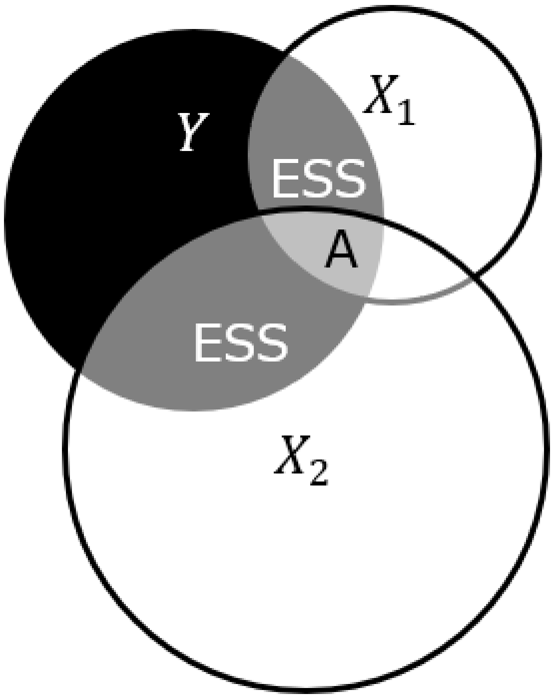


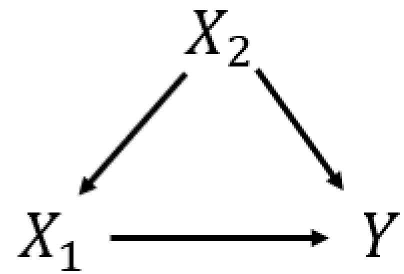


- X_1からYへの因果効果を$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i}$の単回帰モデルの $\hat{\beta}_1$ で測ろうとすると，その中には，Aで表されている部分が含まれている．
- このAの部分は，X_2からの間接的な効果であり，これが交絡である．
- X_1とX_2が重なっている部分Aを取り除いて分析する必要がある．

三変数の重回帰モデル

説明変数の間に相関がある場合の続き





- $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$における$\beta_1$の偏りのない推定を行うことを考える.
- Aの部分を取り除いて，X_1からYへの純粋な効果を測る方法を考察する.

三変数の重回帰モデル

- $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$
- $\hat{\beta}_1 = \frac{\sum(X_{1i}-\hat{X}_{1i})(Y_i-\bar{Y})}{\sum(X_{1i}-\hat{X}_{1i})^2}$
- $\hat{\beta}_2 = \frac{\sum(X_{2i}-\hat{X}_{2i})(Y_i-\bar{Y})}{\sum(X_{2i}-\hat{X}_{2i})^2}$
- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$

- β_1とβ_2は，特に偏回帰係数（partial regression slope）と呼ばれる．
- ここでは，$\hat{\beta}_1$を使って，具体的にどのようにして交絡を取り除くことができているのかを確認する．
- この考え方は，統計的因果推論において極めて重要である．

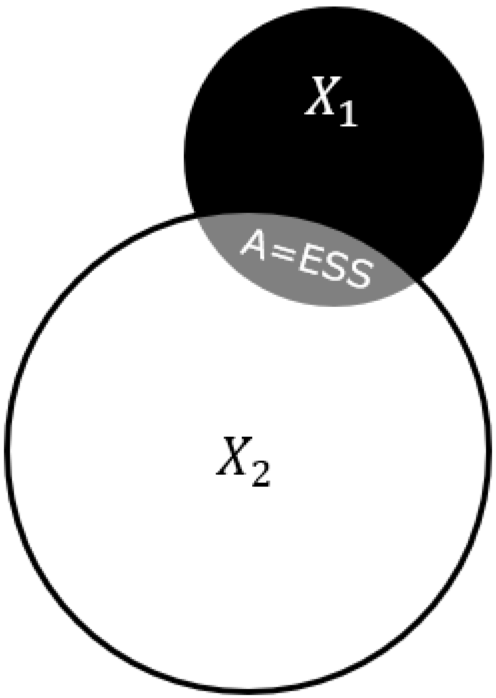
単回帰モデルと重回帰モデルの違い

- 単回帰モデル
- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta} X_{1i}$
- $\hat{\beta} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$

- 重回帰モデル
- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$
- $\hat{\beta}_1 = \frac{\sum(X_{1i} - \hat{X}_{1i})(Y_i - \bar{Y})}{\sum(X_{1i} - \hat{X}_{1i})^2}$

- 重回帰モデルの$\hat{\beta}_1$は，一見すると単回帰モデルの$\hat{\beta}$とよく似ている.
- $\hat{\beta}$はXの平均値からの偏差を計算している.
- $\hat{\beta}_1$はX_1の残差を計算している.
- 重回帰モデルでは，$\hat{X}_{1i} = a_1 + b_1 X_{2i}$である点に注意しよう.
- すなわち，三変数の重回帰モデルは，二段階で分析を行っている.
- まず，単回帰モデル$\hat{X}_{1i} = a_1 + b_1 X_{2i}$を作り，$X_{2i}$から予測値$\hat{X}_{1i}$を計算する.
- 次に，残差$X_{1i} - \hat{X}_{1i}$を計算して，この残差部分からYへの回帰を行う.

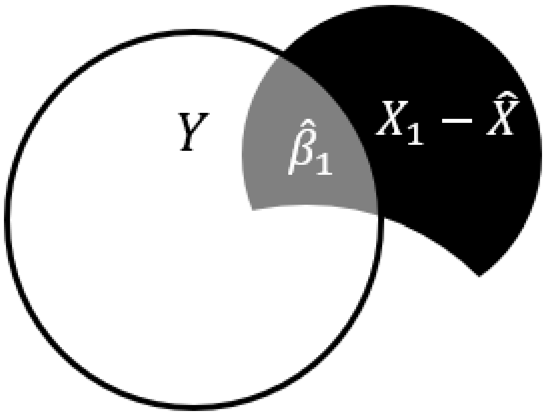
重回帰モデル：第1段階目

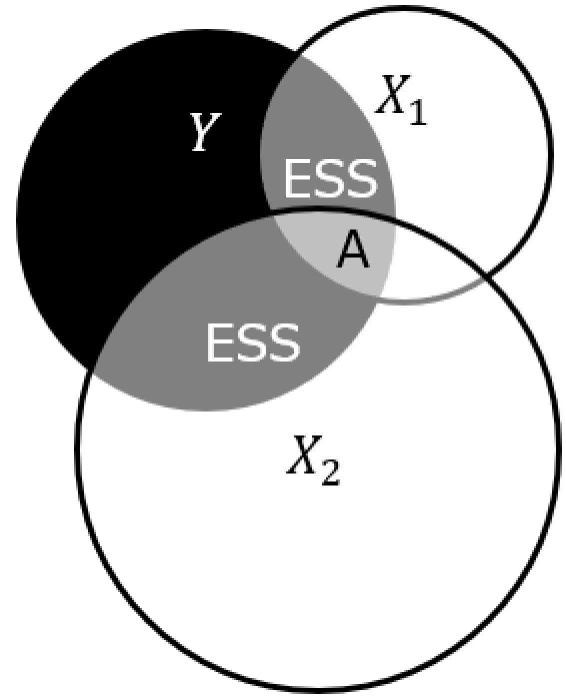


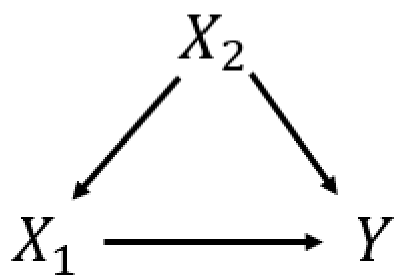
- まず，図から一時的に変数 Y を取り除く．
- すると，変数X_1と変数X_2の単回帰モデル$\hat{X}_{1i} = a_1 + b_1 X_{2i}$と同じ状況である．
- Aの部分は，ESSである．
- そして，図の黒い部分がUSSである．
- このとき，USSは以下の式である．

$$USS = \sum_{i=1}^{n} \left(X_{1i} - \hat{X}_{1i}\right)^2$$

重回帰モデル：第2段階目



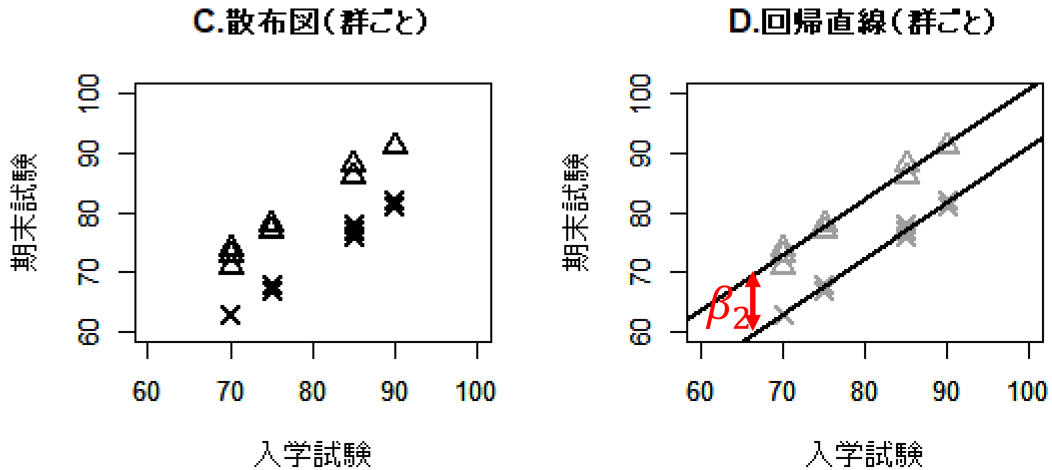




図に変数Yを戻すと，一段階目の手順により，変数X_2からの交絡Aを取り除くことができている様子が分かる．

共分散分析(再考)

復習：共分散分析（ANCOVA）



$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 T_i + \varepsilon_i$$

Y_i：結果変数（期末試験）

β_0，β_1，β_2：回帰の母数（パラメータ）

X_i：共変量（入学試験）

T_i：処置を表す二値変数（補習授業）

ε_i：誤差項（error term）

共分散分析（ANCOVA）

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 T_i + \varepsilon_i$$

- ダミー変数を説明変数として持つ重回帰モデルである.
  - ダミー変数とは，0または1の二値をとる変数のことである.

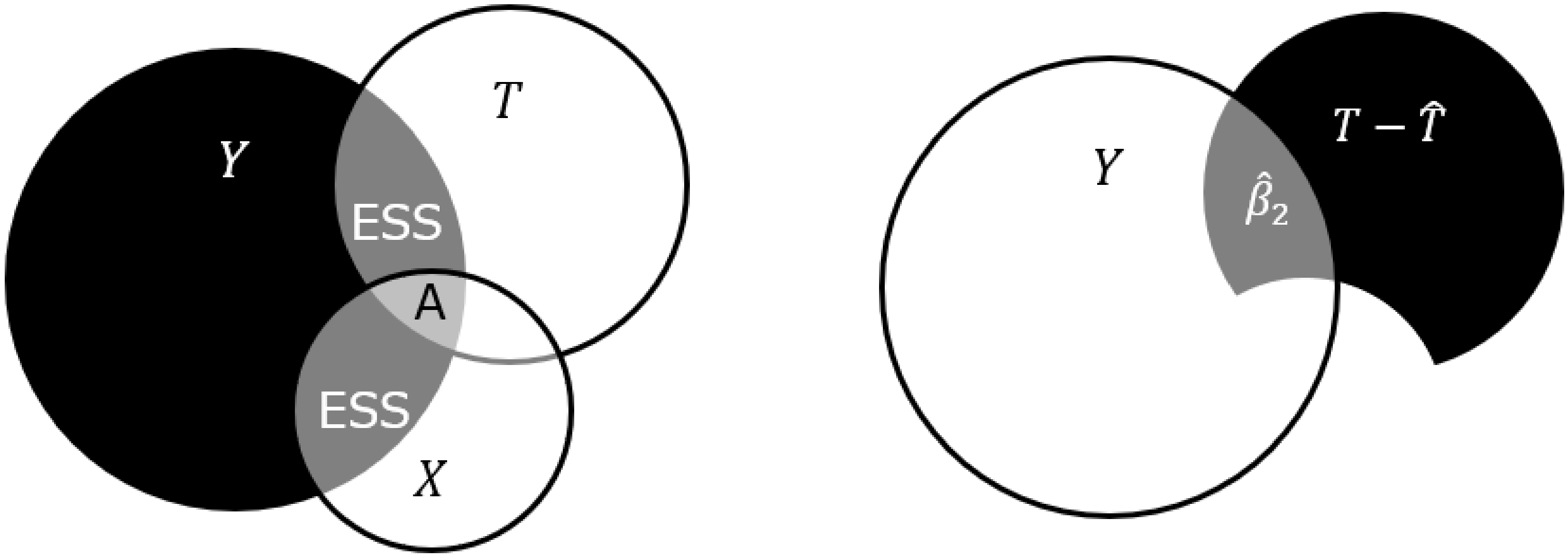
復習：処置の割付け変数

- T_i
  - 個体iが処置（treatment）に割付られたかどうかを表す二値変数
  - $T_i \in \{0, 1\}$
- $T_i = 0$
  - 個体iが処置に割付けられていないこと
  - 統制群：$T_i = 0$となる集団
- $T_i = 1$
  - 個体iが処置に割付られたこと
  - 処置群：$T_i = 1$となる集団

表2.2

ID	入学試験 x1	処置 t1	期末試験 0 y0	期末試験 1 y1	潜在的結果 0 y0t	潜在的結果 1 y1t	潜在的結果の差 y1t − y0t
1	74	1		76	68	76	8
2	82	0	75		75	84	9
3	72	1		75	65	75	10
4	96	0	84		84	97	13
5	83	0	75		75	84	9
6	72	1		74	65	74	9
7	85	0	76		76	87	11
8	87	0	77		77	89	12
9	86	0	77		77	87	10
10	77	1		80	70	80	10
11	95	0	87		87	96	9
12	84	0	75		75	85	10
13	74	1		77	67	77	10
14	58	1		61	52	61	9
15	91	0	81		81	93	12
16	80	0	72		72	84	12
17	80	0	72		72	82	10
18	89	0	70		70	89	19
19	88	0	70		70	90	20
20	86	0	78		78	87	9

二段階の推定により交絡を取り除く

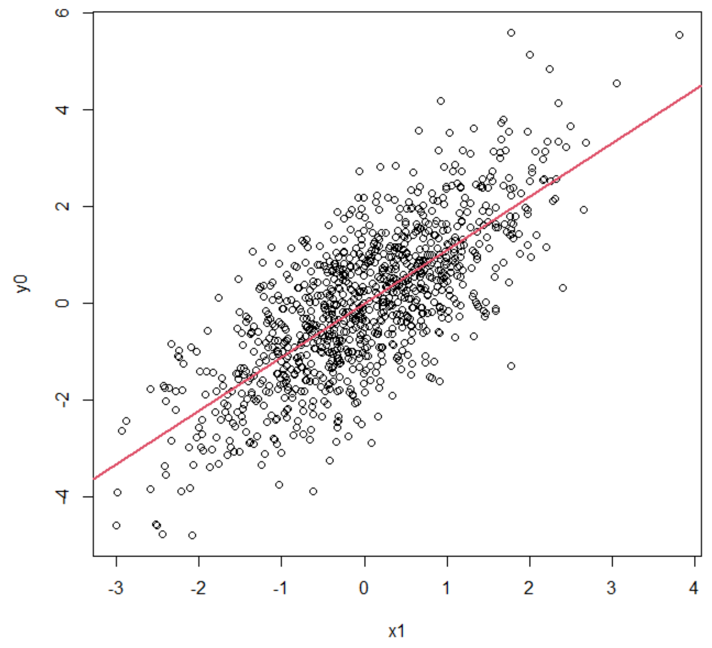


- 共分散分析のメカニズムも，これまでと同様に考えることができる．
- β_2は結果変数Yの変動を説明する際に，共変量に対して，ダミー変数Tの純粋な貢献度合いを表していると考えることができる

回帰分析（regression）

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

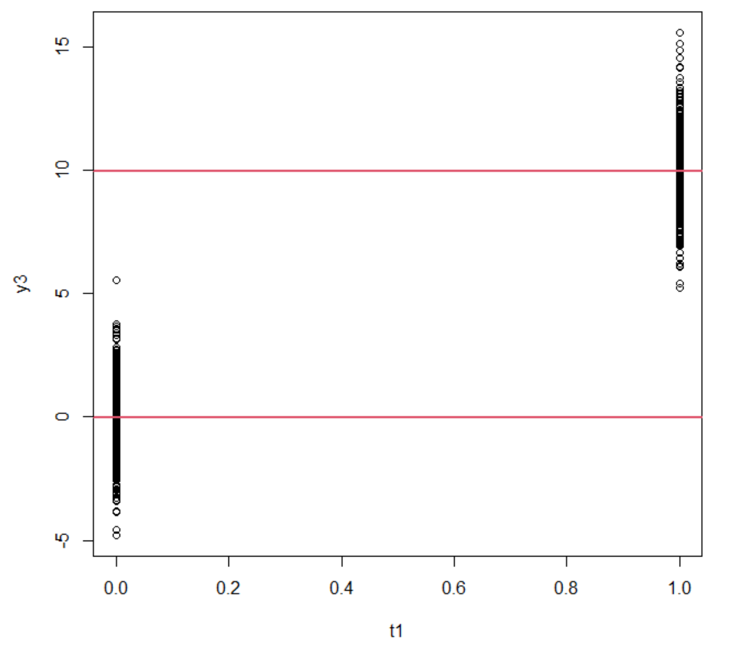
- すべての説明変数が連続変数のみから構成されている



分散分析（ANOVA: analysis of variance）

- $Y_i = \beta_0 + \beta_2 T_i + \varepsilon_i$
- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_2 T_i$

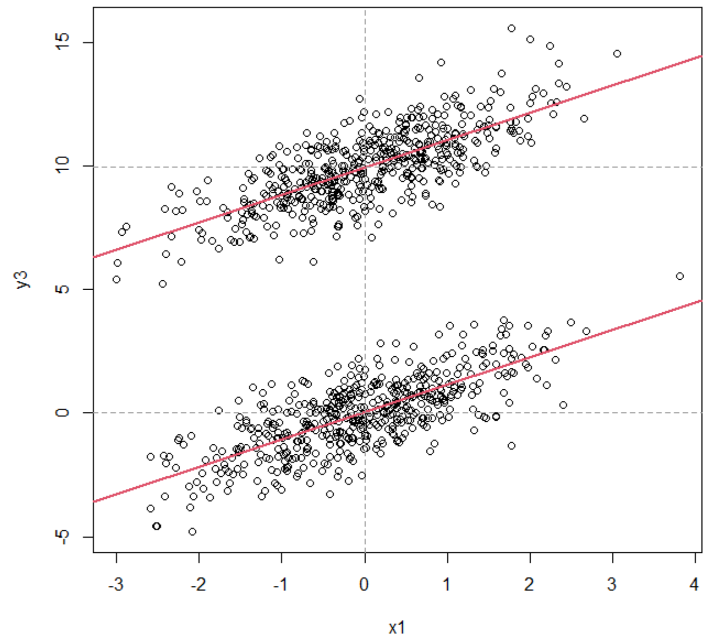
- すべての説明変数が**ダミー変数**のみから構成されている回帰分析の特殊版である.
  - T_iが二値の場合，t検定，分散分析，回帰分析はすべて同じである.
  - 説明変数にダミー変数のみを用いた回帰分析は，2標本t検定と同じである.



共分散分析（ANCOVA: analysis of covariance）

- $Y_i = \beta_0 + \beta_1 X_i + \beta_2 T_i + \varepsilon_i$
- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 T_i$

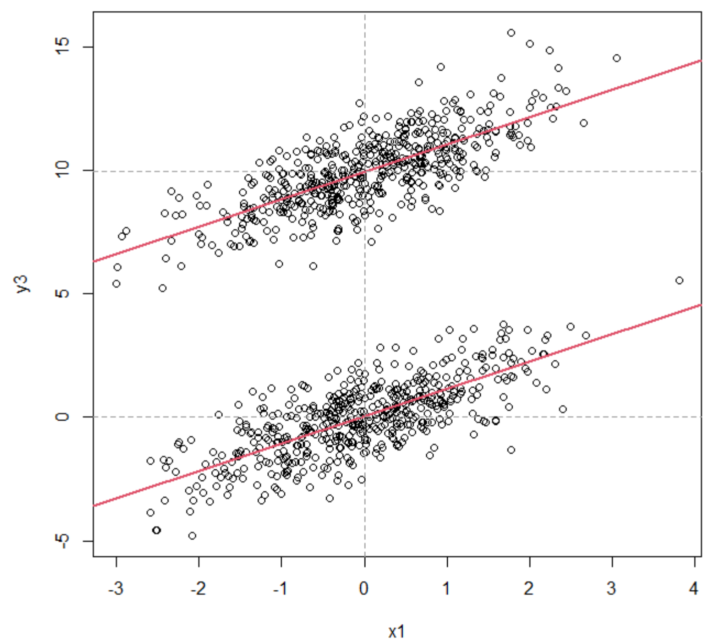
- 回帰分析と分散分析の長所を併せ持つように工夫されたものである.



$T_i = 0$のとき

- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 T_i$
- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

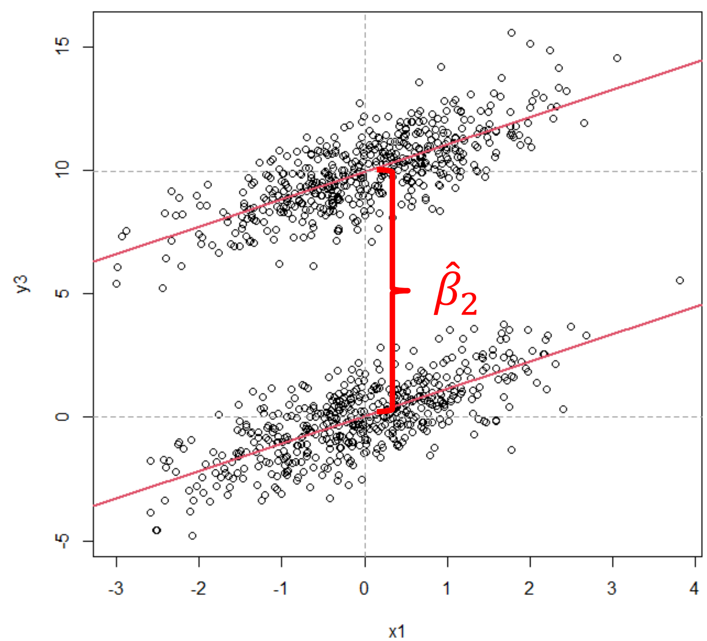
- 統制群におけるY-切片は，$\hat{\beta}_0$である．
- 統制群では，X_iの値が0のとき，Y_iの値は$\hat{\beta}_0$である．



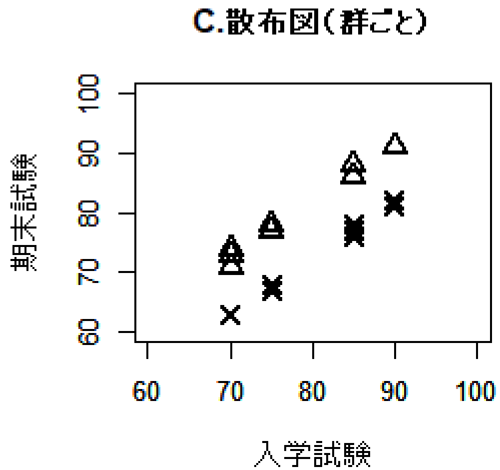
$T_i = 1$のとき

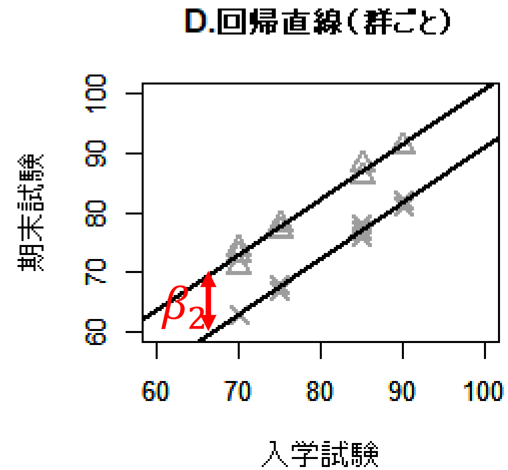
- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 T_i$
- $\hat{Y}_i = \left(\hat{\beta}_0 + \hat{\beta}_2\right) + \hat{\beta}_1 X_i$

- 処置群におけるY-切片は, $\hat{\beta}_0 + \hat{\beta}_2$である.
- 処置群では, X_iの値が0のとき, Y_iの値は$\hat{\beta}_0 + \hat{\beta}_2$である. したがって, $\hat{\beta}_2$は2つの集団における切片の変化を表している



復習：共分散分析（ANCOVA）





$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 T_i + \varepsilon_i$$

Y_i：結果変数（期末試験）
β_0，β_1，β_2：回帰の母数（パラメータ）
X_i：共変量（入学試験）
T_i：処置を表す二値変数（補習授業）
ε_i：誤差項（error term）

重回帰モデルによる分析

データの読み込み（教科書pp.77-78）

- data06 <- read.csv(file.choose( ))
- attach(data06)
- summary(data06)

y1：ノーベル賞受賞数
x1：チョコレート消費量
x2：国内総生産（GDP）

```
> summary(data06)
   country                 y1               x1             x2
 Length:23          Min.   : 0.000   Min.   :0.10   Min.   : 8.755
 Class :character   1st Qu.: 1.961   1st Qu.:4.00   1st Qu.:31.453
 Mode  :character   Median : 8.644   Median :4.90   Median :46.232
                    Mean   : 9.748   Mean   :4.87   Mean   :44.704
                    3rd Qu.:13.857   3rd Qu.:6.20   3rd Qu.:53.908
                    Max.   :30.763   Max.   :8.80   Max.   :85.135
```

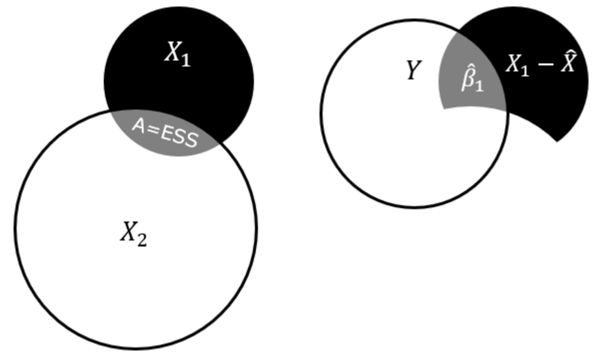
単回帰モデル（教科書p.80）

- model1 <- lm(y1 ~ x1)
- summary(model1)
- confint(model1, level=0.95)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.4217     3.2274  -1.060 0.301096
x1            2.7044     0.5985   4.519 0.000188 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.65 on 21 degrees of freedom
Multiple R-squared:  0.493,     Adjusted R-squared:  0.4689
F-statistic: 20.42 on 1 and 21 DF,  p-value: 0.000188

> confint(model1, level=0.95)
                 2.5 %   97.5 %
(Intercept) -10.133320 3.290018
x1            1.459837 3.949020
```

二段階推定による重回帰モデル（教科書p.87）

- model2a <- lm(x1 ~ x2)
- ex1 <- resid(model2a)
- model2b <- lm(y1 ~ ex1)
- summary(model2b)




```
Coefficients:
            Estimate
(Intercept)    9.748
ex1            1.505
```

重回帰モデル（教科書p.87）

- model3<-lm(y1~x1+x2)
- summary(model3)
- confint(model3, level=0.95)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.32035    3.20331  -1.973   0.0625 .
x1           1.50477    0.75619   1.990   0.0604 .
x2           0.19552    0.08531   2.292   0.0329 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.064 on 20 degrees of freedom
Multiple R-squared:  0.5985,    Adjusted R-squared:  0.5583
F-statistic:  14.9 on 2 and 20 DF,  p-value: 0.0001089

> confint(model3, level=0.95)
                   2.5 %    97.5 %
(Intercept) -13.00233992 0.3616436
x1           -0.07260349 3.0821496
x2            0.01757188 0.3734626
```