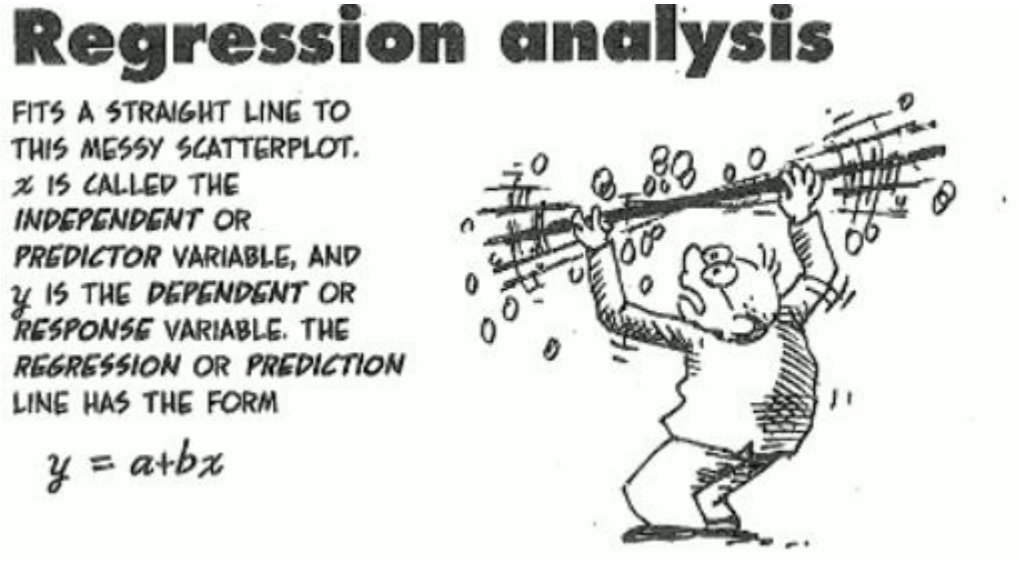
# 計量分析演習

第5回

岡島　成治

# 今週の予定

- 最小二乗法を学ぶ
- 詳しい式の展開が知りたい人は連絡を



# 政策の「効果」とは？

- 因果関係　「朝ごはんと成績の関係」

文部科学省が全国学力・学習状況調査を行っている。

このアンケート調査を見ると「朝ごはんを食べている生徒はテストの点が高い」と言うことが分かった。

では「朝ごはんを食べると学力向上する」から「学力向上のための朝給食を提供」という政策を支持できるか？

# 政策の「効果」とは？

- 「朝ごはんを毎日食べている生徒はテストの点が高い」というのは、朝ごはんを食べている生徒のほうが、そうでない生徒に比べてテストの点が高い「傾向」（相関関係）があると言っているに過ぎない。
- 家庭環境がテストの点に影響している可能性がある

朝ごはんと食べさせている親は子供の教育に熱心

なので、いままで朝ごはんを食べなかった家庭の生徒に朝ごはんを食べさせてもテストの点はきっと変わらない。

よってきちんと因果関係を推定しなければいけない。

# 単回帰モデル

- 相関関係から因果関係へ

条件付き期待値

E[テストの点数｜朝ごはんを食べてるか]

E[Y|X=x] ただし

朝ごはんを食べてたら　　　x=1

朝ごはんを食べていないなら　　x=0

# 単回帰モデル

$$\mathrm{E}[Y|X=1]-E[Y|X=0]$$

となる。

一方、家庭環境等の外的条件がそろっていない場合には、外的条件に関する情報を制御する確率変数Cのとりうる値ｃを使い外的条件をそろえる。(重回帰分析)

$$\mathrm{E}[Y|X=1,C=c]-E[Y|X=0,C=c]$$

# 単回帰モデル

- 関数のモデル化

すでに外的条件が制御されているとする。（共変量がいらない）

因果関係をみるには、政策変数X=xのみで条件付けした成果変数Yの期待値を考えればよい。

$$E[Y|X=x] = \beta_0 + \beta_1 X$$

# 単回帰モデル

$$E[Y|X = x] = \beta_0 + \beta_1 X$$

- 朝ごはんを毎日食べている場合の期待されるテストの点数

$$E[Y|X = 1] = \beta_0 + \beta_1 \times 1 \ = \beta_0 + \beta_1$$

- 朝ごはんを食べていない場合の期待されるテストの点数

$$E[Y|X = 0] = \beta_0 + \beta_1 \times 0 = \beta_0$$

- 朝ごはんを毎日食べていることのテストの点数の因果関係

$$E[Y|X = 1] - E[Y|X = 0] = \beta_1$$

# 単回帰モデル

この式はあくまでも平均的な値

$$E[Y|X=x] = \beta_0 + \beta_1 X$$

しかしたとえ朝ごはんを毎日食べていてもテストの日の調子によって実際のテストの点数は上下する。この「揺らぎ」を誤差項として書き直す。

$$Y = \beta_0 + \beta_1 X + U$$

# 単回帰モデル

$$Y = \beta_0 + \beta_1 X + U$$

被説明変数（従属変数）　説明変数（独立変数）　誤差項

切片パラメーター（定数項）　係数（傾き）パラメター

# 単回帰モデル

- 因果関係を示すための条件

外的条件が制御されていればパラメーター$\beta_1$は因果関係

外的条件が制御されていないならパラメーター$\beta_1$は相関関係

この前提条件が正しいという条件は、誤差項$\boldsymbol{U}$が以下の性質をもたなければいけない。

# 単回帰モデル　因果関係のための仮定

- 因果関係のための仮定１

説明変数Xと誤差項$U$は平均独立

$$E[U|X] = E[U]$$

説明変数Xの値がわかったとしても誤差項$U$の平均に関して一切わからない。

# 単回帰モデル　因果関係のための仮定

- 因果関係のための仮定２

誤差項$U$の母平均は0

$$E[U] = 0$$

# 最小２乗法

単回帰モデル

$$Y = \beta_0 + \beta_1 X + U$$

のパラメーター（$\beta_0, \beta_1$)を推定する。

このパラメーターを推定するためには標本（データ）が必要

n個の観測値からなる標本$X_{\mathrm{i}}, Y_{\mathrm{i}}$を使うことが出来るとする。

説明変数と被説明変数のペアを一つの観測値と呼ぶのでｎ人分の$X_{\mathrm{i}}, Y_{\mathrm{i}}$が含まれているデータは「標本サイズはｎ」になる。

# 最小２乗法

- 説明変数と被説明変数のペアを一つの観測値と呼ぶのでｎ人分の$X_{\mathrm{i}}, Y_{\mathrm{i}}$が含まれているデータは「標本サイズはｎ」になる。

| | 点数(y) | 朝食と取ったか(x) |
|---|---|---|
| れいさん | ８３ | 0 |
| なっちゃん | ２１ | 1 |
| ひなさん | ７１ | 1 |
| ゆうとくん | ９３ | 1 |
| かいとくん | ５３ | 0 |

# 最小２乗法 パラメータの計算方法

- モーメント法(因果関係のための仮定を使う)
- 誤差項の平均独立と誤差項の期待値が0から

$$\mathrm{E}[\mathrm{U}|\mathrm{X}] = \mathrm{E}[\mathrm{U}] = 0$$

さらに$\mathrm{E}[\mathrm{U}] = 0$という仮定の下では$\mathrm{E}[X\mathrm{U}] = 0$

# 最小２乗法　パラメータの計算方法

$$\mathrm{E[U]} = 0$$
$$\mathrm{E}[X\mathrm{U}] = 0$$

から

$$U = Y - \beta_0 - \beta_1 X$$

を代入すると

$$\mathrm{E}[Y - \beta_0 - \beta_1 X\ ] = 0$$
$$\mathrm{E}[X(Y - \beta_0 - \beta_1 X)] = 0$$

# 最小２乗法　パラメータの計算方法

- これを解くと

$$\beta_1 = \frac{COV[X,Y]}{Var[X]}$$

$$\beta_0 = E[Y] - \frac{COV[X,Y]}{Var[X]}E[X]$$

# 傾きパラメターをどう解釈するか？

単回帰モデル

$$Y = \beta_0 + \beta_1 X + U$$

傾きパラメター$\beta_1$の推定値の解釈は、説明変数と被説明変数それぞれの単位によって決まる。

# 傾きパラメーターをどう解釈するか？

単回帰モデル

$$Y = \beta_0 + \beta_1 X + U$$

例）

「修学年数が増えると、年収がどれだけ増えるのか」

修学年数が１年増えると、年収が$\beta_1$万円増える。

単位：説明変数の単位は「年」被説明変数の単位は「万円」

# 傾きパラメターをどう解釈するか？

単回帰モデル

$$Y = \beta_0 + \beta_1 X + U$$

単位：説明変数の単位は「年」被説明変数の単位は「万円」

もし被説明変数の単位が「千円」ならば、$\beta_1$ は10倍になる。

しかしもし修学年数が１年増えると、年収が「何％」増えるかがわかると、年収の単位に依存しなくてよい。

# 傾きパラメターをどう解釈するか？

単回帰モデル

$$lnY = \beta_0 + \beta_1 X + U$$

説明変数が１単位増えたときの被説明変数が何％増えるのかを調べる方法として、被説明変数の自然対数をとればよい。

# 傾きパラメターをどう解釈するか？

| 被説明変数 | 説明変数 | 解釈 |
| --- | --- | --- |
| Y(レベル) | X(レベル) | Xが１単位増えたとき、Yが$\beta_1$単位増える。 |
| lnY(ログ) | X(レベル) | Xが１単位増えたとき、Yが$100 \times \beta_1$単位増える。 |
| Y(レベル) | lnX(ログ) | Xが１％増えたとき、Yが$\beta_1$/100単位増える。 |
| lnY(ログ) | lnX(ログ) | Xが１%増えたとき、Yが$\beta_1$%増える。 |

# 決定係数

XがYをどの程度説明したかの指標

$$Y_i = \alpha + \beta X_i + U_i = \hat{\mathrm{Y}} + U_i$$

**Y** が完全に説明されるとき

$U_i$ = 0, i = 1,...,n

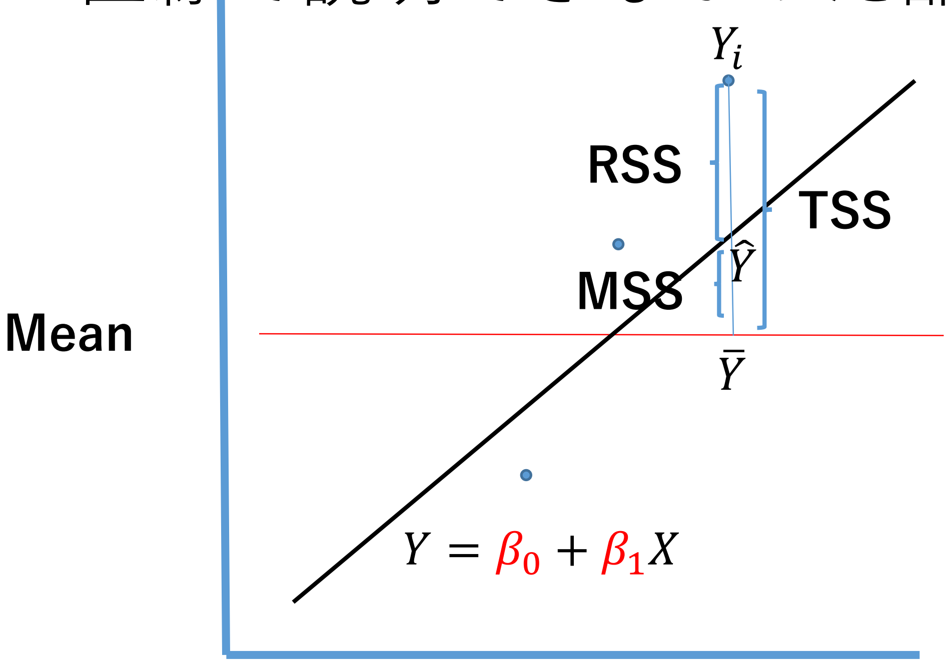
**Y** が全く説明されないとき

Y の予測($\hat{\mathrm{Y}}$)にX は全く役立たない

あてはめ値がX の値に関わらず一定 ($\bar{\mathrm{Y}}$)

# 決定係数

- TSS(全変動)、MSS（回帰直線で説明できた部分）RSS（回帰直線で説明できなかった部分）




TSS(Total sum of Squares)は説明変数を全く投入しない切片だけのモデルの基準値であり、OLS回帰分析では説明変数を投入することでTSSを「説明」していく。

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

MSS(Model Sum of Squares)は回帰直線で説明できた部分。

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

RSS(Residual Sum of Squares)は回帰直線で説明できなかった部分

RSS= $\sum(Y_i - \hat{Y})^2$

# 決定係数

- 最小２乗法によって求めた一次関数が、どの程度データを説明してくれているのかの指標

**TSS=MSS+RSSから**

$$R^2 = \frac{MSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n} \widehat{u_i^2}}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

もし被説明変数を完全に説明してくれる一次関数を見つけることができれば決定係数は１になる。

# Stataで単回帰分析

- Stata code

reg y x

例）Arctic9には毎年の9月の北極圏の海氷面積のデータが入ってある。そのデータを使い年が経つごとに北極圏の海氷面積がどう変わっているかを分析する。

回帰分析

regress area year

# Stataで単回帰分析

回帰分析

regress area year

```
reg area year

      Source |       SS           df       MS      Number of obs   =        33
-------------+----------------------------------   F(1, 31)        =     99.55
       Model |  17.4995295         1  17.4995295   Prob > F        =    0.0000
    Residual |  5.44916742        31  .175779594   R-squared       =    0.7626
-------------+----------------------------------   Adj R-squared   =    0.7549
       Total |   22.948697        32   .71714678   Root MSE        =    .41926

------------------------------------------------------------------------------
        area |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |  -.0764773   .0076648    -9.98   0.000    -.0921098   -.0608447
       _cons |   157.4225   15.29154    10.29   0.000     126.2352    188.6098
------------------------------------------------------------------------------
```

北極圏の海氷面積の単位が100万km2なので、北極圏の海氷面積は年間、約76000km2減っている。

# 決定係数

MSS

回帰の自由度

観測値の数

RSS

TSS

決定係数＝MSS/TSS

| Source | SS | df | MS | | | |
|---|---|---|---|---|---|---|
| Model | 8191296.88 | 1 | 8191296.88 | Number of obs | = | 4,327 |
| Residual | 126657674 | 4,325 | 29285.0113 | F(1, 4325) | = | 279.71 |
| | | | | Prob > F | = | 0.0000 |
| Total | 134848971 | 4,326 | 31171.7455 | R-squared | = | 0.0607 |
| | | | | Adj R-squared | = | 0.0605 |
| | | | | Root MSE | = | 171.13 |

| income | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| yeduc | 23.151 | 1.384255 | 16.72 | 0.000 | 20.43715 | 25.86485 |
| _cons | -56.89278 | 19.35684 | -2.94 | 0.003 | -94.84211 | -18.94344 |

# 決定係数

残差の自由度
全データ数ー回帰の自由度ー１

回帰の自由度：回帰式中の説明変数の数

観測値の数

| Source | SS | df | MS |
|---|---|---|---|
| Model | 8191296.88 | 1 | 8191296.88 |
| Residual | 126657674 | 4,325 | 29285.0113 |
| Total | 134848971 | 4,326 | 31171.7455 |

| | | |
|---|---|---|
| Number of obs | = | 4,327 |
| F(1, 4325) | = | 279.71 |
| Prob > F | = | 0.0000 |
| R-squared | = | 0.0607 |
| Adj R-squared | = | 0.0605 |
| Root MSE | = | 171.13 |

決定係数＝MSS/TSS

| income | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| yeduc | 23.151 | 1.384255 | 16.72 | 0.000 | 20.43715 | 25.86485 |
| _cons | -56.89278 | 19.35684 | -2.94 | 0.003 | -94.84211 | -18.94344 |

# 決定係数

平均平方：平方和/自由度
何で自由度で割るかというと
母集団から取り出した標本に対する
不偏分散だから。

例えば
大きさｎの標本から不偏分散
を計算するときはn-1で
割るのだけどこのn-1が
自由度。

| Source | SS | df | MS |
|---|---|---|---|
| Model | 8191296.88 | 1 | 8191296.88 |
| Residual | 126657674 | 4,325 | 29285.0113 |
| Total | 134848971 | 4,326 | 31171.7455 |

| | | |
|---|---|---|
| Number of obs | = | 4,327 |
| F(1, 4325) | = | 279.71 |
| Prob > F | = | 0.0000 |
| R-squared | = | 0.0607 |
| Adj R-squared | = | 0.0605 |
| Root MSE | = | 171.13 |

Root mean squared error
＝自由度あたりのRSSの
平方根
予測値が正解からどの程度
乖離しているのか？
０に近いほどよい。

| income | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| yeduc | 23.151 | 1.384255 | 16.72 | 0.000 | 20.43715 | 25.86485 |
| _cons | -56.89278 | 19.35684 | -2.94 | 0.003 | -94.84211 | -18.94344 |

# 予想値の残差のStata Code

- 予想値
- predict areahat
- label variable areahat “Area predicted from Year”
- graph twoway connect area areahat year, msymbol(o +)
- 残差
- predict areares, resid
- label variable areares “Residualss, area predicted from year”
- summarize area areahat areares

# 問題１

- Sleepのデータの説明
- 通勤時間(commute、単位は分)
- 睡眠時間(sleep、単位は分)

のデータを使って通勤時間が長いと、睡眠時間が短くなるかを調べてください。

通勤時間が一分長くなると、睡眠時間がどれくらい減りますか？

# 問題 2

Incomeのデータの説明

4327人分の年収(income　万円)と修学年数(yeduc)が含まれている。

- 被説明変数をincome説明変数をyeducにして散布図と書いてレベルーレベル、ログーレベル、レベルーログ、ログーログすべてを回帰分析してください。そしてその解釈を述べなさい。
- またincomeの変数を千円単位income1000を作り、万単位のレベルレベルモデルの推定値$\beta$を比較してください。
- さらにincome1000のログレベルの推定値$\beta$とincomeのログレベルの推定値$\beta$を比較してください。

# 問題 3

Carsのデータの説明

- Prefecture:県名
- Cars:人口千人当たり自動車数
- Stations:鉄道駅数

被説明変数をCars説明変数をStationsにして散布図と書いて回帰分析してください。さらにその式の残差を計算してその期待値が０になっていることを確かめてください。

# 問題 4

Icecreamのデータの説明

- Icecream:世帯当たりのアイスクリーム年間消費額(単位:100)
- U15:世帯当たり１５歳以下の子供の平均人数
- 被説明変数をIcecream説明変数をU15にして散布図と書いて回帰分析してください。さらにその予想値を求め予想値を横軸、被説明変数を縦軸にとった図を書いてください。

# 問題 5

定数項のない回帰モデルの決定係数は当てはまりの尺度として適切ではない理由として$\sum_{i=1} U_i = 0$が保証されないからである。

そこでデータArctic9を使って被説明変数をarea説明変数をyearとして定数項のあるモデルは$\sum_{i=1} U_i = 0$が成立し、定数項のないモデルは$\sum_{i=1} U_i = 0$が成立しないことを確かめよ。

ヒント

定数項のないモデルのstataｺｰﾄﾞ

```
reg area year, noconstant
```

列の合計の計算コードは

```
total()
```

# エキストラ問題１

- 単回帰モデルの傾きパラメータの推定量が不偏性を持つ、つまり$E[\widehat{\beta_1}] = \beta_1$になることを確認してください。

- 基本の仮定が成り立ってるとする

(1) 説明変数$X_i$は非確率変数である。

(2) $E(U_i) = 0$

(3) $Var(U_i) = E\left(U_i^2\right) = \sigma^2$

(4) $Cov\left(U_i, U_j\right) = 0 \ (i \neq j)$

# エキストラ問題２

仮定(3),(4)の

(3) $Var(U_i) = E\left(U_i^2\right) = \sigma^2$

(4) $Cov\left(U_i, U_j\right) = 0 \ \ (i \neq j)$

の例を述べてください。

# エキストラ問題3

- 単回帰モデルの傾きパラメータ$\beta_1$を導出してください。