

回歸分析

重回歸(1)

項目

- 重回帰モデル
- 最小二乗推定量の性質*
 - 仮説検定（単一の制約）
 - 決定係数
- 回帰分析の実際
- 非線形効果
- ダミー変数
 - 定数項ダミー
 - 傾きのダミー
 - 3つ以上のカテゴリー

* 詳細は「回帰分析（重回帰）」reg2.pdf を参照してください

重回帰モデル

multiple regression model

- 重回帰 → 説明変数が2個以上の場合
 - 単回帰は説明変数が1つ
- モデル

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

- 係数の意味： $\beta_i = \frac{\partial y}{\partial x_i}$
 - 他の説明変数の値が一定だとして、 x_i だけを1単位増加させたときに y が何単位増えるかを表す

重回帰モデル

前提

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$$

1. 線型モデル（パラメータに関し）
2. 誤差項の期待値は0
3. 誤差項は互いに独立
4. 誤差項の分散は一定（分散均一性）
5. 誤差項は正規分布に従う
 - BLUEの成立のためには5番目の条件は不要

最小二乗法

残差平方和を最小にするようにパラメータを決定

- $a, b_1, b_2, \dots, b_k$: 未知パラメータ $\alpha, \beta_1, \beta_2, \dots, \beta_k$ の推定値
- e_i : 残差

次の式を最小にするように $a, b_1, b_2, \dots, b_k$ を決める

$$\begin{aligned} S(a, b_1, b_2, \dots, b_k) &= \sum_{i=1}^n e_i^2 \\ &= \sum_{i=1}^n (y_i - a - b_1 x_{1i} - b_2 x_{2i} - \dots - b_k x_{ki})^2 \end{aligned}$$

最小二乗推定量

- b_j の期待値と分散

$$E(b_j) = \beta_j$$

$$\text{var}(b_j) = \frac{\sigma^2}{S_{xx}^j}$$

係数の期待値は真の値に等しい（不偏性という性質を持つ）

また、 S_{xx}^j は説明変数 x_j 固有の平方和（ x_j を他の説明変数に回帰したときの残差平方和=他の説明変数の影響を除いた x_j 固有の平方和）

- 誤差項の分散の推定量

$$s^2 = \frac{1}{n - (k + 1)} RSS = \frac{1}{n - (k + 1)} \sum_{i=1}^n e_i^2$$

$n - (k + 1)$ は残差の自由度

$k + 1$ は説明変数の个数（定数項+説明変数の数）

s : SER (standard error of the regression: 回帰の標準誤差)

s は s^2 （誤差項の分散の推定値）の平方根

仮説の検定

次の仮説を考える $H_0: \beta_j = \beta_{j0}$

H_0 が正しければ、 b_j は次の分布に従う

$$\frac{b_j - \beta_{j0}}{\sqrt{\sigma^2 / S_{xx}^j}} \sim N(0,1)$$

σ^2 は未知なので、残差の平方和から計算された s^2 を用いると

$$\frac{b_j - \beta_{j0}}{\sqrt{s^2 / S_{xx}^j}} = \frac{b_j - \beta_{j0}}{\text{s.e.}(b_j)} \sim t(n - (k + 1))$$

$\frac{b_j - \beta_{j0}}{\text{s.e.}(b_j)}$ は自由度 $n - (k + 1)$ の t 分布に従う。

仮説の検定(2)

- R等の統計ソフトでのoutputでは β_{j0} の値が0だった場合の t 値, p 値が出力されます。つまり, 次の t が t 値として出力されます

$$t = \frac{b_j}{s.e.(b_j)}$$

- この t は自由度 $(n-(k+1))$ の t 分布に従います
- p 値とは自由度 $(n-(k+1))$ の t 分布にしたがう確率変数 x が, $\Pr(|x| > |t|)$ となる確率のことです。ここで, t は上の式で求めた t 値で, $| \ |$ は絶対値を表します。
- p 値が0.02であるとは, 係数の真の値が0という仮説が正しい場合, 推定された係数以上の値 (絶対値でみて) が推計される確率は0.02であることを意味します。したがって, このようなことはほぼ確率的に起こらないと考え, 係数の真の値が0だという仮説を棄却します。
- 通常は, p 値が0.05以下の場合, 係数が0だという仮説を棄却します。そして, 係数は0と有意に異なる(significantly different from 0)と言います,
- β_{j0} の値が0でない場合は, $(b_j - \beta_j^0)/s.e.(b_j)$ を計算して仮説の検定を行います。

当てはまりの良さ

- $TSS = ESS + RSS$

全平方和 = 回帰式で説明できる部分の平方和 + 残差平方和

- 決定係数 R^2

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

問題点：説明変数の数 k を増やしていけば、 R^2 は単調に増加する

- 自由度修正済み決定係数 adjusted R^2

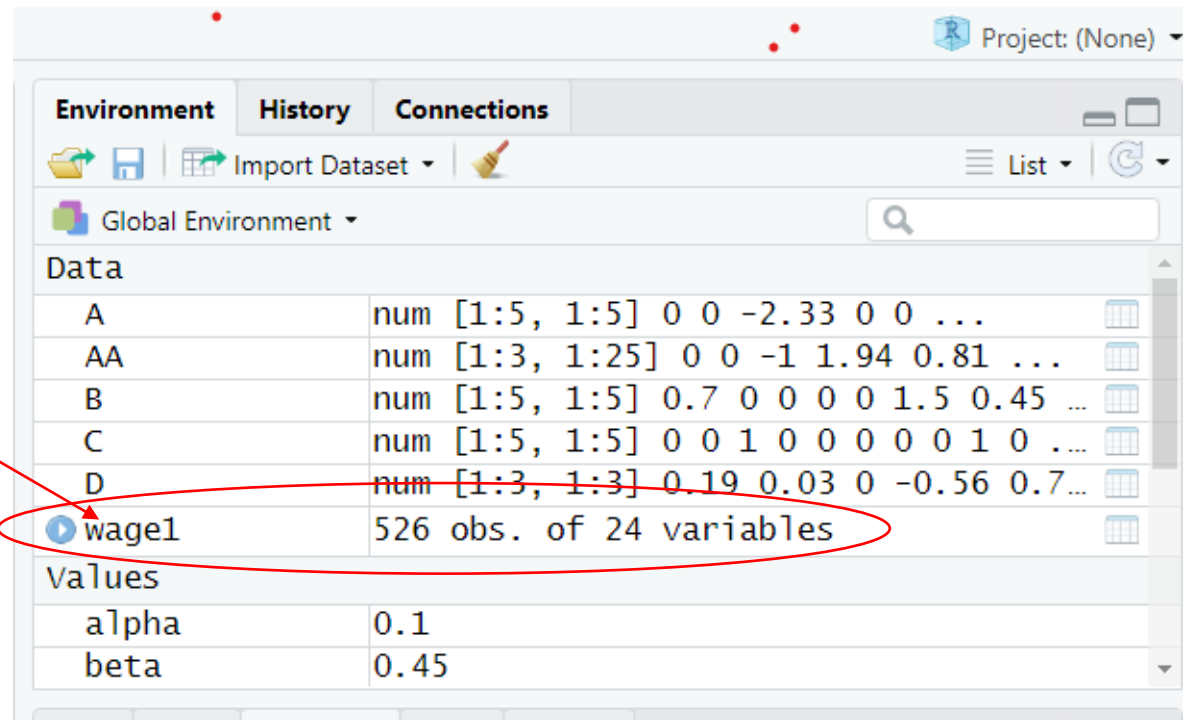
説明変数の増加にペナルティーを課すように修正した R^2

$$\bar{R}^2 = 1 - \frac{RSS/(n - k - 1)}{TSS/(n - 1)}$$

- 通常は、adjusted R^2 を報告する

回帰分析の実際

- wage1.xls または wage1.rawをRにimportしてあり, wage1というデータフレームができているとします。
- Rstudio の右上のwindowのEnvironmentでwage1が表示されていれば, importできています。
- 見当たらない場合はimportする
- 正しくimportできたかどうかは, コマンドラインで `summary(wage1)` とする→データフレーム wage1に含まれている変数の要約統計量が出力される→この結果で判断



回帰分析 R

- 回帰分析 → lm関数を用いる(分析に使用するデータフレームをattachしておくこと)

```
lm(y ~ x1 + x2 + x3 + x4)
```

- その結果をobjectに代入

例) `wage1.lm <- lm(wage ~ educ + exper + tenure)`

object名はどんな名前でもいいが、何を行ったかわかるようにしておく
(`wage1.lm`としたのは `lm()`の結果だとわかるようにするため)

- `summary(object名)`で結果の概要を出力
- `plot(object名)`で残差の診断
- 次のコマンドを実行

```
wage1_1.lm<- lm(wage~ educ + exper + tenure)
```

```
summary(wage1_1.lm)
```

```
plot(wage1_1.lm)
```

Rの出力

係数の推定値, 標準誤差, t 値, p 値が出力される

```
lm(formula = wage ~ educ + exper + tenure)
```

2.93×10^{-14} という意味

Coefficients:	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-2.87273	0.72896	-3.941	9.22e-05 ***
educ	0.59897	0.05128	11.679	< 2e-16 ***
exper	0.02234	0.01206	1.853	0.0645
tenure	0.16927	0.02164	7.820	2.93e-14 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.084 on 522 degrees of freedom

Multiple R-squared: 0.3064, Adjusted R-squared: 0.3024

F-statistic: 76.87 on 3 and 522 DF, p-value: $\leq 2.2e-16$

回帰の標準誤差 : standard error of the regression

自由度修正済み決定係数

問題(1)

- データセット `wage1.xls(wage1.raw)`
1. 次の回帰分析を行い、係数の意味を説明せよ。また、各係数は0と有意に異なるか？
 - 被説明変数：wage
 - 説明変数：educ, exper, tenure
 2. 次の回帰分析を行い、係数の意味を説明せよ。また、各係数は0と有意に異なるか？
 - 被説明変数：lwage (wageの対数値)
 - 説明変数：educ, exper, tenure

(被説明変数を対数値にした場合の解釈は「回帰分析 単回帰」ecnmtrcs03.pdfを参照のこと)

非線形効果

- 説明変数 x の2次の項を説明変数として加えることを考える

$$y = a + b_1x + b_2x^2 + b_3z + e$$

- z は他の説明変数
- 係数の意味

$$\frac{\partial y}{\partial x} = b_1 + 2b_2x$$

- 他の変数は一定だとして、 x を1単位増加させた場合、 y は b_1+2b_2x 増加する
 - x 増加の効果→ x の水準に依存
 - 例えば、 y がoutput, x がinputとすると、このような定式化で限界生産物逓減の効果をとらえることができる

係数の意味の把握の仕方

- b_1, b_2 の値をもとに $\frac{\partial y}{\partial x} = b_1 + 2b_2x$ の値を計算させる
(いくつか方法があります)
 - Excelに回帰係数の値をコピーし、異なる x に対応する $\partial y / \partial x$ の値を計算させる
 - R等の統計ソフトの中で、回帰分析の係数を取り出して計算させる方法
- Excel またはR等で、 $yhat = b_1x + b_2x^2$ を計算し（他の説明変数は無視して）、 x と $yhat$ のグラフを描かせる
- Rの場合
 - 回帰分析の結果が`wage1.lm`のようなオブジェクトに保存されているとする
 - `coef(オブジェクト名)` で回帰係数を取り出せる
 - `coef(オブジェクト名)[1]`とすると1番目の係数（定数項）を取り出せる
 - `coef(オブジェクト名)[2]`で2番目の係数を取り出せ、これをもとに $yhat$ や dy/dx の値を計算すればよい

回帰分析の推計値の取り出し方 (R)

回帰分析の結果は`summary(object)`で取り出せたが、他の情報も取り出せる

`summary(object)` 回帰分析の結果の要約

`coef(object)` 係数の推定値

`resid(object)` 残差

`fitted(object)` 回帰モデルの推定値

`deviance(object)` 残差平方和

`plot(object)` 残差のチェックのためのグラフ

`confint(object)` 係数の信頼区間

コマンドラインで、`coefficients(wage1.lm)` または `coef(wage1.lm)` とタイプすると推計された係数が出力される

`coef(wage1.lm)[1]` `coef(wage1.lm)[2]` で係数ベクトルの1番めの要素と2番めの要素が出力される

Rでの変数の作成方法

- コマンドラインで

新変数名 <- 計算式

で作成できる

例) `lnwage <- log(wage)`
`exper2 <- exper * exper`
`exper2 <- exper^2`

- 新変数は、そのままではデータフレームの外に作られる→上の式で、新変数名をデータフレーム名\$新変数 として、新変数をデータフレームに追加し、そのデータフレームを保存しておく、後で利用できる

- 回帰式の中での指定→計算式で指定することもできる。

`log( )`はそのまま使えるようだが、2次式等は`I( )`関数を用いる

`lm(log(wage) ~ educ)`

`lm(wage ~ educ + I(educ^2))`

tenureの2乗 (tenursq)を説明変数に加えた回帰

```
lm(formula = lwage ~ educ + exper + tenure + tenursq)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.04272	-0.27978	-0.02262	0.27206	1.40011

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	0.2756792	0.1029287	2.678	0.007632	**
educ	0.0897291	0.0072651	12.351	< 2e-16	***
exper	0.0032729	0.0017169	1.906	0.057154	.
tenure	0.0465254	0.0071852	6.475	2.19e-10	***
tenursq	-0.0009321	0.0002478	-3.761	0.000188	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4354 on 521 degrees of freedom

Multiple R-squared: 0.3341, Adjusted R-squared: 0.329

F-statistic: 65.35 on 4 and 521 DF, p-value: < 2.2e-16

tenureが1単位増加した場合の効果

- 前頁の回帰分析がwage1.lmに保存されているとして， tenureの係数は4番目， tenursq(=tenure^2)の係数は5番目なので， Rstudioで次のようにタイプする(# 以下はコメントなので無視)

```
b1 <- coef(wage1.lm)[4] # tenureの係数をb1に代入
b2 <- coef(wage1.lm)[5] # tenursqの係数をb2に代入
x <- seq(from=0, to= 45) # 0から45まで1刻みの変数を作成
dydx <- b1 + 2 * b2 * x # dy/dxの計算
yhat <- b1*x + b2 * x^2 # yhatの計算
plot(x, dydx) # 横軸がx, 縦軸がdydxのグラフを描かせる
plot(x, yhat) # 横軸がx, 縦軸がyhatのグラフを描かせる
```

線グラフにしたい場合はplot(x,yhat, type="l") とする

問題(2)

- データセット : wage1.xls or wage1.raw
- $\ln(\text{wage})$ (wageの対数値)を被説明変数にし, educ, exper, tenure, tenure^2 を説明変数にして回帰分析を行え。
- tenureの範囲を調べよ。
 - `range()`, `summary()`, `table()`関数などが使える
- tenureが1年増加したとき, wageは何%増加するか
 - tenure=0, 5, 10, 20, 30, 40のそれぞれの場合について
- 上の回帰分析の係数の値を用い, tenureとwageの関係をグラフで表せ (他の変数は無視する)。

ダミー変数

- 経済分析に用いるデータの中には、所得や消費のように連続量で表されるデータもありますが、女性かそうでないとか働いているかいないかなどのように、質を表す変数もあります。
- ある性質を満たしていれば1、満たしていなければ0のような質を表す変数をダミー変数(dummy variable)といいます。

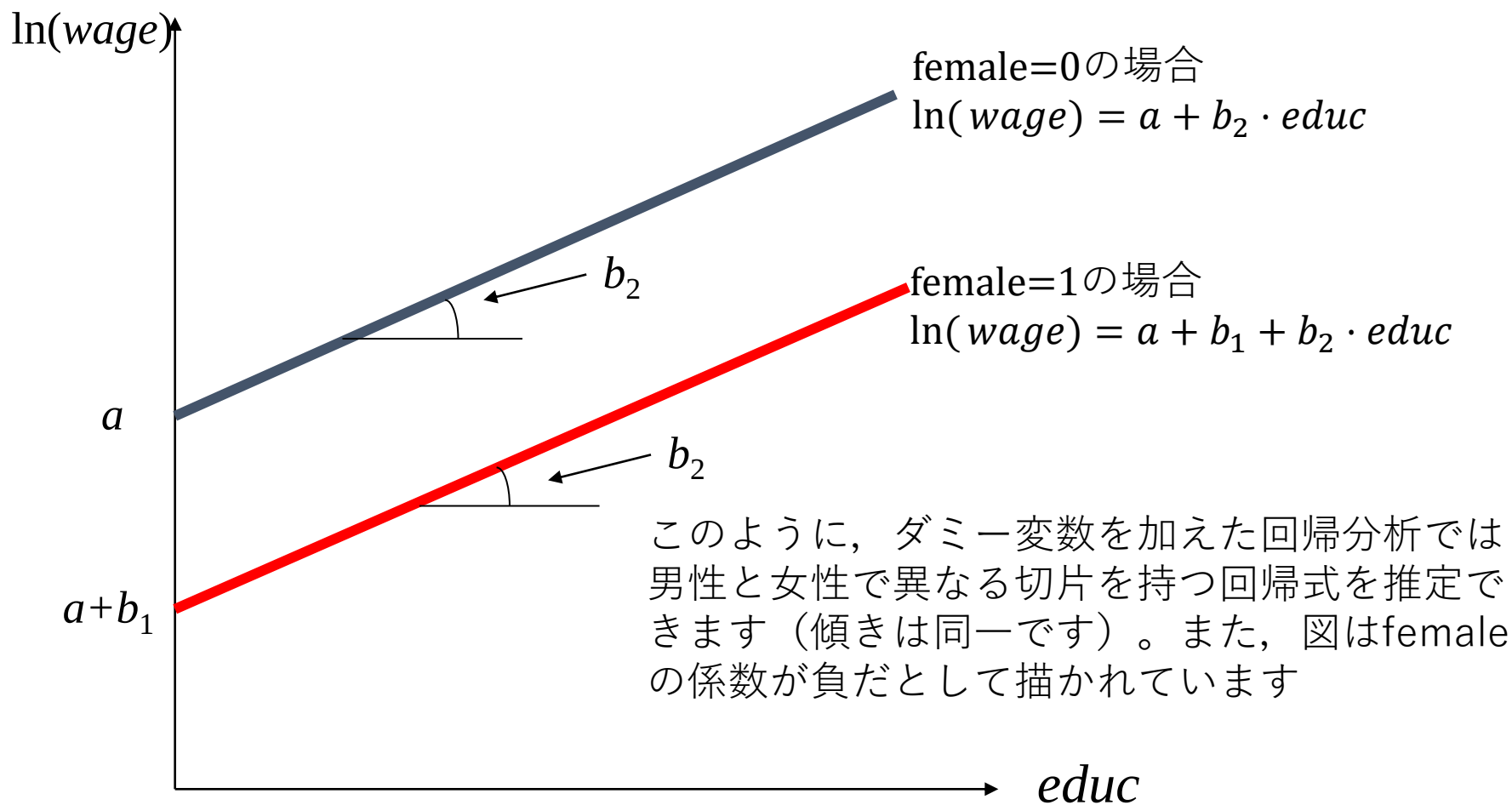
(例)

- 女性ならば1, そうでなければ0
- 結婚していれば1, そうでなければ0
- 大学卒ならば1, そうでなければ0
- educ, wage, exper → 連続変数
- 一般に、0または1をとるような変数をダミー変数と呼ぶ

ダミー変数(2)

- 定数項ダミー
- 傾きに関するダミー
- 3つ以上のカテゴリーを持つ変数の場合
 - 学歴
 - 中卒または高校中退
 - 高卒, 大卒未満
 - 大卒以上
 - 職業
 - 事務職
 - 研究職
 - 営業
 - 現場

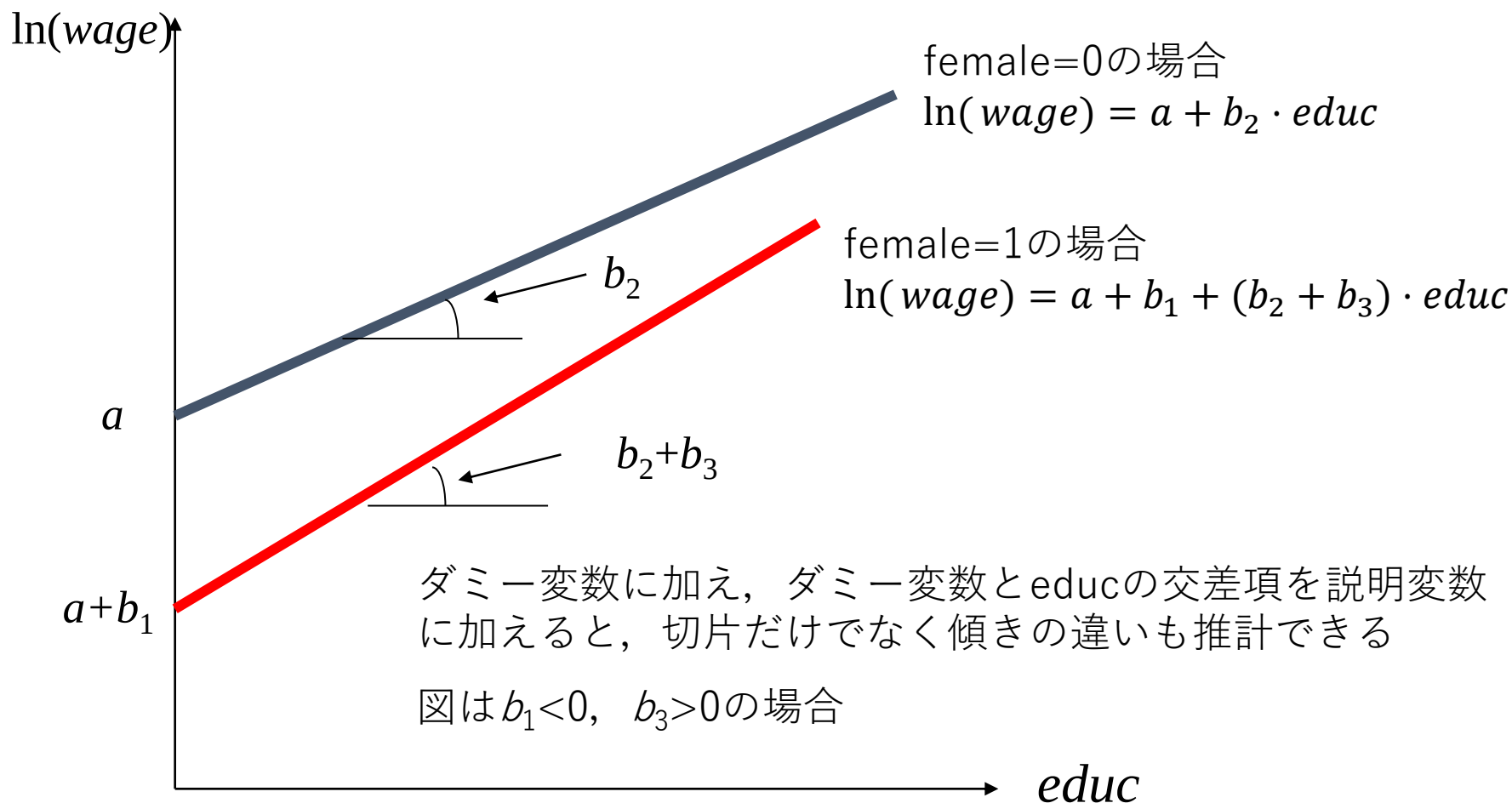
定数項ダミー — female (女性なら1, そうでなければ0) というダミー変数を用いた回帰を考えます

$$\ln(wage) = a + b_1 \cdot female + b_2 \cdot educ$$


傾きのダミー

educとfemaleの交差項を回帰式の説明変数に加えると傾きの違いも推計できる

$$\ln(wage) = a + b_1 \cdot female + b_2 \cdot educ + b_3 \cdot (female \times educ)$$



問題 (3)

- データセット : wage1.xls or wage1.raw
- femaleダミー変数を説明変数に加えた回帰を行え。
また、他の変数が一定の値をとった場合、女性の賃金は男性の賃金にくらべてどのくらい異なるか。
 - 被説明変数 lwage
 - 説明変数 educ, exper, tenure, female
- educ と femaleの交差項を回帰式の説明変数に加えて回帰分析を行え。educ(教育年数)の1年の増加が賃金に与える効果を男女別に求めよ。

Rでは次のコマンドでeducとfemaleの交差項が作成できます
(変数名はed_fmとした場合)

```
ed_fm <- educ * female
```

問題 (4)

- 次の回帰を行う
 - 被説明変数 lwage
 - 説明変数 educ, tenure, exper, female, female*educ, female*tenure, female*exper
- 男女別に回帰分析を行う
 - 説明変数を educ, tenure, exper として回帰
 - ダミー変数を用いた回帰と結果を比較せよ。

男女別に回帰分析を行うには

- R : `lm()` でsubset関数を使う
 - `lm(y~x1+x2, subset=(female==0))` #female =0 を満たすサンプルのみで回帰
 - `lm(y~x1+x2, subset=(female==1))` #female =1 を満たすサンプルのみで回帰

3つ以上のカテゴリー

	中卒	高卒	大卒
ed1	0	1	0
ed2	0	0	1

- 学歴ダミーのケース
 - 中卒, 高卒, 大卒 の3つのカテゴリー
 - 3つのカテゴリーがある場合, 2つのダミー変数をつくる
 - あるカテゴリーをベースにして比較を行うため
 - ed1 (高卒なら 1, そうでなければ 0)
 - ed2 (大卒なら 1, そうでなければ 0)
 - ed1, ed2は中卒に比べた高卒, 大卒の比較 (切片の違い)
- 高卒と大卒の比較は?
- 3つダミー変数を作るとどうなるか?
- N 種類のカテゴリー $\rightarrow N-1$ 個のダミー変数

問題 (5)

- 教育年数の影響は、連続変数で捉えるのではなく、学歴別に調べた方がよいかもしれない
- 教育年数の分布を調べよ
- 教育年数から次のような学歴ダミー変数を作れ
 - 高卒ダミー（高卒以上大卒未満）
 $\text{educ} \geq 12$ かつ $\text{educ} < 16$
 - 大卒ダミー（大卒以上）
 $\text{educ} \geq 16$
- 次の回帰分析を行え
 - 被説明変数：lwage, 説明変数：学歴ダミー, その他の変数（exper, tenure, female）

変数の作成方法(R)

コマンドラインで

```
ed1 <- (educ < 16) & (educ >= 12)
ed2 <- (educ >= 16)
```

- 論理式を用いて1(TRUE)または0 (FALSE) の変数を作成
- ed1 (高卒ダミー) とed2 (大卒ダミー) はTRUEとFALSEの2値をとる。このままで回帰分析に使える
- Rでの論理演算子
 - == 等しい
 - & and
 - | or
 - xor どちらか1つだけが真
 - ! 否定