回帰分析I

13．ダミー変数

1

イントロダクション

- これまで数多くの回帰分析のシミュレーションをしてきましたが、そこでの説明変数はすべて「量的」なものとしていました。
- 実際の「量的」な変数というと、例えば、
  - 売上高、営業利益、経常利益
  - 外国人持株比率、女性役員比率
  - 所得、賃金、消費、労働時間
  - 年齢、教育年数、GPA

などが挙げられます、他にいくらでもありますが。

- 「量的」な変数は「連続的」な変数です。
- これらは「どれぐらい大きいか（小さいか）」という情報を与えます。
- 回帰分析では、「量的」な要因だけでなく、「質的」な要因も考慮に入れることが可能です。

2

- そして「質的」な変数を分析に入れることは、ほとんどの分析においてとても重要になります。
- 「質的」な変数の例としては、
  - 企業が属する産業（製造業、建設業、運輸・通信業、卸売・小売業など）
  - 個人の性別（男性か女性か）、信仰する宗教、居住する県

などが挙げられます。ですから離散的な変数になります。

- 同じ（ような）情報でも、手持ちのデータによって、「量的」か「質的」かが決まることもあります。
  - 例えば、「教育年数」が変数としてデータセットに存在することもあるし、「中卒、高卒、大卒（以上）」のような形でデータセットに存在することもあります。
  - 前者であれば「量的」な変数だし、後者なら「質的」な変数ですね、どちらも教育水準を測る変数ですが。

3

- この講義ノートでは、「質的」な変数、特に「質的」な説明変数について取り扱います。
- 「質的」な要因をどのように分析の中に取り込めばいいか、を理解することが目的です。
- 具体的には以下のことを学びます。
  - ダミー変数の使い方
  - ダミー変数の使い方： 複数のカテゴリーへの対応
  - ダミー変数の使い方： 序数的な情報への対応

4

質的な要因の変数化:ダミー変数

- 質的な要因は、多くの場合、バイナリー(0か1)の形で変数化されます。
- 「男性・女性」を質的な要因の例として挙げましたが、、、、
- 「している・していない」、「ある・ない」なども質的な要因といえます、よってバイナリー(0か1)の形で変数化されます。
  - 企業がR&Dをしている、していない
  - 企業が業績連動型ストック・オプションを導入している
  - 個人がPCを所有している、していない
  - 個人が結婚している、していない
- 回帰分析では、このようなバイナリー(0か1)変数のことを「**ダミー変数**」と呼ぶのが一般的です。

5

変数化の仕方

- 「男性・女性」はバイナリー(0か1)の形で変数化できますが、二通りありますね。
- 一つ目は、男性なら1の値をとり、女性なら0の値をとる変数を作る。変数名は$MALE$とでもしておきましょう。
- もう一つは、女性なら1の値をとり、男性なら0の値をとる変数を作る。変数名は$FEMALE$とでもしておきましょう。
- 両方の変数を分析の中に入れることは意味がないですよね、完全な多重共線性が起きますから。
  - $FEMALE + MALE = 1$です。
- よって、どちらかだけ分析の中に入れればいいのですが、どっちがいいとかある?
- ないです、どちらをいれても構いません。

6

- $MALE$をモデルに入れたとき、$FEMALE$をモデルに入れたとき、推定結果は一見違うように見えます。
- でも実は同じ結果です。
- それはそうです、$MALE$と$FEMALE$は基本的に同じ情報を有するわけですから。
- これについては後ほど確認してみましょう。
- また、ここでは「男性・女性」を例として使いましたが、今の話はすべてのダミー変数に当てはまります。
- おまけ: なんで0−1なの? 0−1じゃなきゃだめなの?
  - そんな疑問を持たれる方いらっしゃると思います。
  - 実は0−1じゃなくとも構いません。(二つの異なる数ならなんでもOK)。
  - でも0−1を使いたい理由があります、それは係数の解釈が容易になるためです。
  - 係数の解釈の容易さ以外で言えば、結果に変わりはありません。

7

ダミー(説明)変数の使い方

- それではダミー変数がモデルにおいてどのような働きをするかを見てみましょう。
- 次のモデルを考えます。

$$WAGE = \beta_0 + \delta_0 FEMALE + \beta_1 EDUC + U, \quad E(U|FEMALE, EDUC) = 0$$

$WAGE$: 時間当たり賃金
$FEMALE$: = 1 (女性) = 0 (男性)
$EDUC$: 教育年数
U: 誤差項

ここでδ_0は$FEMALE$の係数です。

注: これ以降、係数を表すとき、$\beta_j (j = 1, \dots, k)$ 以外の記号を使うことがあります。今後は特に断りは入れません。

8

$$WAGE = \beta_0 + \delta_0 FEMALE + \beta_1 EDUC + U, \quad E(U|FEMALE, EDUC) = 0$$

- $FEMALE$の係数δ_0は何を表すでしょうか、丁寧に見ていきましょう。
- まず$FEMALE = 1$のときの$WAGE$の条件付期待値を考えます。

$$E(WAGE|FEMALE = 1, EDUC) = \beta_0 + \delta_0 + \beta_1 EDUC + E(U|FEMALE = 1, EDUC)$$

となります。

- ここで$E(U|FEMALE, EDUC) = 0$ですから$E(U|FEMALE = 1, EDUC) = 0$
- 従って、

$$E(WAGE|FEMALE = 1, EDUC) = \beta_0 + \delta_0 + \beta_1 EDUC$$

9

$$WAGE = \beta_0 + \delta_0 FEMALE + \beta_1 EDUC + U, \quad E(U|FEMALE, EDUC) = 0$$

- 次に$FEMALE = 0$のとき（すなわち男性）の$WAGE$の条件付期待値を考えましょう。

$$E(WAGE|FEMALE = 0, EDUC) = \beta_0 + \beta_1 EDUC + E(U|FEMALE = 0, EDUC)$$

- ここで$E(U|FEMALE, EDUC) = 0$ですから$E(U|FEMALE = 0, EDUC) = 0$
- 従って、

$$E(WAGE|FEMALE = 0, EDUC) = \beta_0 + \beta_1 EDUC$$

10

- まとめます:

$$E(WAGE|FEMALE = 0, EDUC) = \beta_0 + \beta_1 EDUC$$
$$E(WAGE|FEMALE = 1, EDUC) = \beta_0 + \delta_0 + \beta_1 EDUC$$
$$\delta_0 = E(WAGE|FEMALE = 1, EDUC) - E(WAGE|FEMALE = 0, EDUC)$$

- 二つの条件付き期待値において$EDUC$は同じです。
- 従って、δ_0は

「他の要因を一定にしたときの、女子と男子の（平均的な）賃金の差」

を捉えたものです。

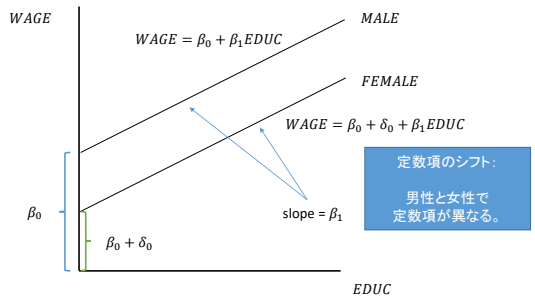
- 他の要因が一定なのに、性別の違いだけで賃金に差があるなら、「差別あり」ってことですかね。

11

- 理解を深めるために図を描いてみましょう。

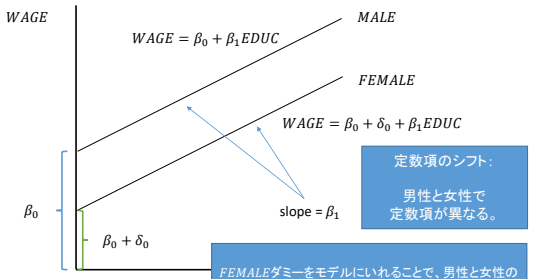
$$E(WAGE|FEMALE = 0, EDUC) = \beta_0 + \beta_1 EDUC$$
$$E(WAGE|FEMALE = 1, EDUC) = \beta_0 + \delta_0 + \beta_1 EDUC$$

- ここでは$\delta_0 < 0$（女性の賃金の方が低い）として図を描いてみます。



12

- 理解を深めるために図を描いてみましょう。 $E(WAGE|FEMALE=0,EDUC)=\beta_0+\beta_1 EDUC$ $E(WAGE|FEMALE=1,EDUC)=\beta_0+\delta_0+\beta_1 EDUC$
- ここでは$\delta_0<0$（女性の賃金の方が低い）として図を描い$EDUC$てみます。



定数項のシフト：
男性と女性で
定数項が異なる。

$FEMALE$ダミーをモデルにいれることで、男性と女性の定数項が異なる可能性を考慮に入れることができる。

13

- 男性の定数項はβ_0、女性の定数項は$\beta_0+\delta_0$。
- これは男性を「基準となるグループ」として選んだということです。
  - 「ベース・グループ」とか「ベンチマーク・グループ」などと呼ばれます。
  - 比較はこのグループに対して、ということ。
- まず基準となる男性の定数項β_0があり、、、
- 男子の定数項と比較して、女子の定数項はどれだけ大きいのか（小さいのか）をδ_0が捉える、、、

という構造です。

- それでは、女性を「基準となるグループ」に選ぶとどうなるでしょうか？
  - $FEMALE$ダミーの代わりに$MALE$ダミーをモデルに入れるということです。
  - モデルは次ページのようになります。

14

$$WAGE=\alpha_0+\gamma_0 MALE+\beta_1 EDUC+U$$

先ほどと同じ手順で以下の式が得られます。

$$E(WAGE|MALE=0,EDUC)=\alpha_0+\beta_1 EDUC$$
$$E(WAGE|MALE=1,EDUC)=\alpha_0+\gamma_0+\beta_1 EDUC$$

- これは、女性を基準となるグループとして選んだ時、女性の定数項がα_0、男性の定数項が$\alpha_0+\gamma_0$となることを示しています。
- 男性を基準となるグループとして選んだときは、男性の定数項はβ_0、女性の定数項は$\beta_0+\delta_0$でした。
- 従って、$\alpha_0=\beta_0+\delta_0$、$\alpha_0+\gamma_0=\beta_0$という関係が成り立ちます。
  - 同じことだけど表現の仕方が異なる、ということですね。
  - 従って、基準となるグループはどちらを選んでもかまいません。
  - ただし、どちらのグループを基準にしたか、は頭に入れておいてください。

15

実際に推定してみましょう

- データ: WAGE1.DAT. 出典：Wooldridge, J.M. (2006) *Introductory Econometrics*, Thomson, South-western
- 記述統計：

```
. summarize wage if female == 1
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        wage |       252    4.587659    2.529363        .53      21.63

. summarize wage if female == 0
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        wage |       274    7.099489    4.160858        1.5      24.98
```

- 時間当たり賃金は、平均すると、女性の方が男性に比べて2.51ドル低いです。
- この差は$WAGE$と$FEMALE$の単回帰モデルにおける$FEMALE$の係数の推定値と一致します。

```
. reg wage female
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  1,   524) =   68.54
       Model |  828.220467     1  828.220467            Prob > F      =  0.0000
    Residual |  6332.19382   524  12.0843394            R-squared     =  0.1157
-------------+------------------------------           Adj R-squared =  0.1140
       Total |  7160.41429   525  13.6388844            Root MSE      =  3.4763

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -2.51183   .3034092    -8.28   0.000    -3.107878   -1.915782
       _cons |   7.099489   .2100082    33.81   0.000     6.686928    7.51205
------------------------------------------------------------------------------
```

16

実際に推定してみましょう

- 時間当たり賃金は、平均すると、女性の方が男性に比べて2.51ドル低い。
- しかもその差は統計的に有意。
- しかしながら、これをもって、賃金において女性差別があると結論づけることはできません。
- なぜなら、男女間で教育水準に差（平均して0.47年）があるからです。

```
. summarize educ if female == 1
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        educ |        252    12.31746    2.472642          0         18

. summarize educ if female == 0
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        educ |        274    12.78832    3.002882          2         18
```

17

- 仮に、男女の賃金差が、「教育水準の差が生み出す生産性の差」を単に反映したものであるならば、賃金において女性差別があるということにはなりませんよね。
- ここで知りたいのは、教育水準が与える賃金への影響をコントロールした上で、男女に賃金差があるかどうかです。
- ここに回帰分析の意味があります。
- やってみましょう。

18

- モデル：$WAGE = \beta_0 + \delta_0 FEMALE + \beta_1 EDUC + U$

```
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  2,   523) =   91.32
       Model |  1853.25304     2  926.626518            Prob > F      =  0.0000
    Residual |  5307.16125   523  10.1475359            R-squared     =  0.2588
-------------+------------------------------           Adj R-squared =  0.2560
       Total |  7160.41429   525  13.6388844            Root MSE      =  3.1855

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -2.273362   .2790444    -8.15   0.000    -2.821547   -1.725176
        educ |   .5064521   .0503906    10.05   0.000     .4074592    .605445
       _cons |   .6228168   .6725334     0.93   0.355     -.698382    1.944016
------------------------------------------------------------------------------
```

- $FEMALE$の係数がゼロという帰無仮説は1%水準で棄却されました。
- 推定値は「他の要因を一定にした時、平均すると、女性の賃金は男性の賃金に比べると2.27ドル低い」ことを示しています。
  - 教育の影響をコントロールしないとき、賃金格差は2.51ドルでした。
  - 欠落変数バイアスが生じていた、ということですかね。

19

- 次に男性ダミーを使ったモデルを推定してみましょう。
- モデル：$WAGE = \alpha_0 + \gamma_0 MALE + \beta_1 EDUC + U$

```
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  2,   523) =   91.32
       Model |  1853.25304     2  926.626518            Prob > F      =  0.0000
    Residual |  5307.16125   523  10.1475359            R-squared     =  0.2588
-------------+------------------------------           Adj R-squared =  0.2560
       Total |  7160.41429   525  13.6388844            Root MSE      =  3.1855

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   2.273362   .2790444     8.15   0.000     1.725176    2.821547
        educ |   .5064521   .0503906    10.05   0.000     .4074592    .605445
       _cons |  -1.650545    .652317    -2.53   0.012    -2.932028   -.3690617
------------------------------------------------------------------------------
```

- $MALE$の係数がゼロという帰無仮説は1%水準で棄却されました。
- 推定値は「他の要因を一定にした時、平均すると、男性の賃金は女性の賃金に比べると2.27ドル高い」ことを示しています。
- さっきと言い方は違いますが、同じことを言っていますね。
- どちらをベース・グループにしても結果は同じということです。

20

- モデル: $WAGE = \beta_0 + \delta_0 FEMALE + \beta_1 EDUC + U$

```
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  2,   523) =   91.32
       Model |  1853.25304     2  926.626518            Prob > F      =  0.0000
    Residual |  5307.16125   523  10.1475359            R-squared     =  0.2588
-------------+------------------------------           Adj R-squared =  0.2560
       Total |  7160.41429   525  13.6388844            Root MSE      =  3.1855

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -2.273362   .2790444    -8.15   0.000    -2.821547   -1.725176
        educ |   .5064521   .0503906    10.05   0.000     .4074592    .605445
       _cons |   .6228168   .6725334     0.93   0.355    -.698382    1.944016
------------------------------------------------------------------------------
```

- モデル: $WAGE = \alpha_0 + \gamma_0 MALE + \beta_1 EDUC + U$

```
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  2,   523) =   91.32
       Model |  1853.25304     2  926.626518            Prob > F      =  0.0000
    Residual |  5307.16125   523  10.1475359            R-squared     =  0.2588
-------------+------------------------------           Adj R-squared =  0.2560
       Total |  7160.41429   525  13.6388844            Root MSE      =  3.1855

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   2.273362   .2790444     8.15   0.000     1.725176    2.821547
        educ |   .5064521   .0503906    10.05   0.000     .4074592    .605445
       _cons |  -1.650545    .652317    -2.53   0.012    -2.932028   -.3690617
------------------------------------------------------------------------------
```


- モデル: $WAGE = \beta_0 + \delta_0 FEMALE + \beta_1 EDUC + U$

```
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  2,   523) =   91.32
       Model |  1853.25304     2  926.626518            Prob > F      =  0.0000
    Residual |  5307.16125   523  10.1475359            R-squared     =  0.2588
-------------+------------------------------           Adj R-squared =  0.2560
       Total |  7160.41429   525  13.6388844            Root MSE      =  3.1855

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -2.273362   .2790444    -8.15   0.000    -2.821547   -1.725176
        educ |   .5064521   .0503906    10.05   0.000     .4074592    .605445
       _cons |   .6228168   .6725334     0.93   0.355    -.698382    1.944016
------------------------------------------------------------------------------
```

男性の定数項:0.6228
女性の定数項:0.6228-2.7733 = -1.6505

- モデル: $WAGE = \alpha_0 + \gamma_0 MALE + \beta_1 EDUC + U$

```
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  2,   523) =   91.32
       Model |  1853.25304     2  926.626518            Prob > F      =  0.0000
    Residual |  5307.16125   523  10.1475359            R-squared     =  0.2588
-------------+------------------------------           Adj R-squared =  0.2560
       Total |  7160.41429   525  13.6388844            Root MSE      =  3.1855

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   2.273362   .2790444     8.15   0.000     1.725176    2.821547
        educ |   .5064521   .0503906    10.05   0.000     .4074592    .605445
       _cons |  -1.650545    .652317    -2.53   0.012    -2.932028   -.3690617
------------------------------------------------------------------------------
```

女性の定数項:-1.6505
男性の定数項:-1.6505+2.2733=0.6228

完全に一致

- モデル: $WAGE = \beta_0 + \delta_0 FEMALE + \beta_1 EDUC + U$

```
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  2,   523) =   91.32
       Model |  1853.25304     2  926.626518            Prob > F      =  0.0000
    Residual |  5307.16125   523  10.1475359            R-squared     =  0.2588
-------------+------------------------------           Adj R-squared =  0.2560
       Total |  7160.41429   525  13.6388844            Root MSE      =  3.1855

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -2.273362   .2790444    -8.15   0.000    -2.821547   -1.725176
        educ |   .5064521   .0503906    10.05   0.000     .4074592    .605445
       _cons |   .6228168   .6725334     0.93   0.355    -.698382    1.944016
------------------------------------------------------------------------------
```

- モデル: $WAGE = \alpha_0 + \gamma_0 MALE + \beta_1$

```
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  2,   523) =   91.32
       Model |  1853.25304     2  926.6265              Prob > F      =  0.0000
    Residual |  5307.16125   523  10.1475359            R-squared     =  0.2588
-------------+------------------------------           Adj R-squared =  0.2560
       Total |  7160.41429   525  13.6388844            Root MSE      =  3.1855

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   2.273362   .2790444     8.15   0.000     1.725176    2.821547
        educ |   .5064521   .0503906    10.05   0.000     .4074592    .605445
       _cons |  -1.650545    .652317    -2.53   0.012    -2.932028   -.3690617
------------------------------------------------------------------------------
```

係数の推定値は一致するけれど、
統計的有意性は？

- これ、推定値だけでなくて、実は推定量の標準誤差も同じです。

```
------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   2.273362   .2790444     8.15   0.000     1.725176    2.821547
        educ |   .5064521   .0503906    10.05   0.000     .4074592    .605445
       _cons |  -1.650545    .652317    -2.53   0.012    -2.932028   -.3690617
------------------------------------------------------------------------------
```

- 男性の定数項は$\alpha_0 + \gamma_0$、これの推定量$\hat{\alpha}_0 + \hat{\gamma}_0$の標準誤差は$lincom$というコマンドで計算できます。

```
. lincom _b[_cons] + _b[male]
 ( 1)  male + _cons = 0

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .6228168   .6725334     0.93   0.355    -.698382    1.944016
------------------------------------------------------------------------------
```

- 男性をベース・グループにした結果と完全に一致しました。

```
------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -2.273362   .2790444    -8.15   0.000    -2.821547   -1.725176
        educ |   .5064521   .0503906    10.05   0.000     .4074592    .605445
       _cons |   .6228168   .6725334     0.93   0.355    -.698382    1.944016
------------------------------------------------------------------------------
```

Stataのlincomコマンド

- lincomコマンドは結構便利なコマンドです。
- このコマンドは、モデルの推定後に使うことができて、、、、
- 複数の係数の線形関数について、その推定値、標準誤差、t値、p値、信頼区間を与えます。
- 例えば、教育年数が10年の女性の賃金の予測値は

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-2.273362	.2790444	-8.15	0.000	-2.821547	-1.725176
educ	.5064521	.0503906	10.05	0.000	.4074592	.605445
_cons	.6228168	.6725334	0.93	0.355	-.698382	1.944016

$\widehat{WAGE} = 0.623 - 2.273 - 0.506 \cdot 10$と計算できますが、統計的に有意にゼロと異なるのか？また信頼区間は？　などに興味があるとします。

- lincomコマンド発動です。


```
. lincom _b[_cons] + _b[female] + 10*_b[educ]
 ( 1)  female + 10*educ + _cons = 0
```

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	3.413976	.2321749	14.70	0.000	2.957866	3.870086

- 教育年数が10年の女性の賃金の予測値は3.414ドル。
- 95%信頼区間は[2.958, 3.870]
- 「時間当たり賃金はゼロ」という帰無仮説に対して、検定統計量のp値はほぼゼロ、、、

みたいな感じで使えます。

- ついでに、教育年数が20年の男性の賃金の予測値は、、、

```
. lincom _b[_cons]  + 20*_b[educ]
 ( 1)  20*educ + _cons = 0
```

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	10.75186	.4112118	26.15	0.000	9.944029	11.55969

- ダミー変数に関する基本的な実習はこれで終わりですが、もうちょっとだけ賃金の実証分析を続けましょう。
- 賃金水準に影響を与える要因は、教育水準だけではありません。
- 例えば、これまでに何年働いたか、また現在勤めている会社で何年働いたか、なども賃金水準に影響を与えるでしょう。
- そしてこれらの要因は$FEMALE$と相関している可能性があります（なぜ？）
- もしそうであるなら、現在の結果には欠落変数バイアスの問題がありますね。
- これらの要因もコントロールして結果がどう変わるか見てみましょう。


```
. reg  wage  female educ exper tenure
```

Source	SS	df	MS		
Model	2603.10658	4	650.776644	Number of obs	= 526
Residual	4557.30771	521	8.7472317	F(4, 521)	= 74.40
				Prob > F	= 0.0000
Total	7160.41429	525	13.6388844	R-squared	= 0.3635
				Adj R-squared	= 0.3587
				Root MSE	= 2.9576

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-1.810852	.2648252	-6.84	0.000	-2.331109	-1.290596
educ	.5715048	.0493373	11.58	0.000	.4745802	.6684293
exper	.0253959	.0115694	2.20	0.029	.0026674	.0481243
tenure	.1410051	.0211617	6.66	0.000	.0994323	.1825778
_cons	-1.567939	.7245511	-2.16	0.031	-2.991339	-.144538

- 差はかなり小さくなってきましたが、まだありそうですね。
- ただし、まだこれでも「差別」があると結論づけることはできません。
  - 他にも重要な変数がモデルから抜け落ちている、と考えられるからです。
  - 考えてみてください。
  - それに加えてモデルの関数形も適切なものではありません。
  - 従属変数は、賃金そのものよりも、賃金の対数値を使うのが一般的です。
  - データを使っていろいろ遊んでみるといいと思います。

ダミー変数の使い方：複数カテゴリーへの対応

- 質的な情報が、複数のカテゴリーで表される場合があります。
- 例えば、個人レベルの分析であれば、居住する都道府県はこれに当たります。
- 企業レベルの分析であれば、属する産業（製造業、建設業、運輸・通信業、卸売・小売業など）はこの例です。
- ここでは教育水準を例として使って考えます。
- データセットに、高卒未満、高卒、大卒以上のいずれであるかの情報があるとしましょう。
- カテゴリーは3つです。

29

- データセットに、高卒未満、高卒、大卒以上のいずれであるかの情報があるとしましょう。
- こういう場合、それぞれのカテゴリーにダミー変数を設定します。

$$JH=\begin{cases}1 & \text{高卒未満}\\ 0 & \text{それ以外}\end{cases}\quad H=\begin{cases}1 & \text{高卒}\\ 0 & \text{それ以外}\end{cases}\quad UN=\begin{cases}1 & \text{大卒以上}\\ 0 & \text{それ以外}\end{cases}$$

- これらのダミーですが、3つのダミー変数のうち必ず一つは1の値を取ります。
- また、あるダミー変数が1の値をとれば、それ以外ダミー変数は0の値をとります。
- 従って、$JH+H+UN=1$が成り立ちます。

30

$$WAGE=\alpha_0+\gamma_0 MALE+\beta_1 EDUC+U$$

- 教育水準に関して、$EDUC$の代わりに、高卒未満、高卒、大卒以上という情報があったとしましょう。
- この場合、JH、H、UNのダミーを使います。
- ただし、モデルに三つのダミーすべてを入れることはできません、$JH+H+UN=1$より完全な多重共線性が生じますから。
- なので二つだけ使います。どの組み合わせでもOKです、同じ結果が得られます。
  - これは男性ダミー・女性ダミーと場合と同じロジックによります。
- 使わなかったダミー変数（＝1）のカテゴリーが、「ベース・カテゴリー」になります。

31

- ここではH、Uのダミーを使ったモデルを考えることにします。

$$WAGE=\alpha_0+\gamma_0 MALE+\beta_1 H+\beta_2 UN+U$$

- 高卒未満の人: $E(WAGE|MALE,H=0,UN=0)=\alpha_0+\gamma_0 MALE$
- 高卒の人: $E(WAGE|MALE,H=1,UN=0)=\alpha_0+\beta_1+\gamma_0 MALE$
- 大卒以上の人: $E(WAGE|MALE,H=0,UN=1)=\alpha_0+\beta_2+\gamma_0 MALE$
- これらの条件付き期待値において$MALE$は同じです。
- 従って、β_1は「ベース・カテゴリーである高卒未満の人に比べて、高卒の人の時間当たり賃金は平均するといくら高い（低い）のか」を捉えます。
- β_2は「ベース・カテゴリーである高卒未満の人に比べて、大卒以上の人の時間当たり賃金は平均するといくら高い（低い）のか」を捉えます。

32

実際に推定してみましょう

- $EDUC$からJH、H、UNを作って、$EDUC$は無かった体でモデルを推定します。

Source	SS	df	MS		
Model	1958.4824	3	652.827466	Number of obs =	526
Residual	5201.93189	522	9.96538677	F(3, 522) =	65.51
				Prob > F =	0.0000
Total	7160.41429	525	13.6388844	R-squared =	0.2735
				Adj R-squared =	0.2693
				Root MSE =	3.1568

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
male	2.205731	.2799592	7.88	0.000	1.655746	2.755716
h	1.763788	.3444719	5.12	0.000	1.087066	2.440509
un	4.593163	.4335902	10.59	0.000	3.741367	5.444959
_cons	2.839769	.3313102	8.57	0.000	2.188904	3.490634

- 高卒は高卒未満に比べて、平均すると、1.76ドル時間当たり賃金が高い。
- 大卒以上は高卒未満に比べて、平均すると、4.59ドル時間当たり賃金が高い。
- 高卒と大卒以上の差は2.83ドル。

33

- JH、H、UN三つともモデルに入れると、、、

```
. reg wage male jh h un
note: un omitted because of collinearity
```

Source	SS	df	MS		
Model	1958.4824	3	652.827466	Number of obs =	526
Residual	5201.93189	522	9.96538677	F(3, 522) =	65.51
				Prob > F =	0.0000
Total	7160.41429	525	13.6388844	R-squared =	0.2735
				Adj R-squared =	0.2693
				Root MSE =	3.1568

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
male	2.205731	.2799592	7.88	0.000	1.655746	2.755716
jh	-4.593163	.4335902	-10.59	0.000	-5.444959	-3.741367
h	-2.829375	.3699455	-7.65	0.000	-3.55614	-2.10261
un	(omitted)					
_cons	7.432932	.3709959	20.04	0.000	6.704104	8.161761

- Stataは大卒以上をベース・カテゴリーに選びました。
- そのためUNをモデルから落としました。
- 結果ですが、、、
  - 高卒未満は大卒以上と比べると、平均して、4.59ドル時間当たり賃金が低い。
  - 高卒は大卒以上と比べると、平均して、2.83ドル賃金が低い。
- 先ほどの結果と同じですね。

34

- lincomを使っていくつか予測値を計算してみましょう。
- 女性で高卒未満:

```
. lincom _b[_cons] + _b[jh]
 ( 1)  jh + _cons = 0
```

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	2.839769	.3313102	8.57	0.000	2.188904	3.490634

- 男性で大卒以上

```
. lincom _b[_cons] + _b[male]
 ( 1)  male + _cons = 0
```

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	9.638663	.3291587	29.28	0.000	8.992025	10.2853

- 男性で高卒

```
. lincom _b[_cons] + _b[male] + _b[h]
 ( 1)  male + h + _cons = 0
```

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	6.809288	.2349197	28.99	0.000	6.347784	7.270792

35

- Stataでは同じことを別の方法でも推定可能です。
- 例えば、データセットによってはカテゴリーに番号が振ってあることがあります。
- 今回の例だと高卒未満なら1、高卒なら2、大卒以上なら3といった具合です。
- この変数の変数名をEDU_CATとでもしましょう。
- reg wage male i.edu_cat
- この i. というのは、Stataに「この変数はカテゴリカル変数です。このカテゴリカル変数からダミー変数を作って、それを使って推定して下さい」とお願いするためのものです。

36

```
. reg wage male i.educ_cat

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  3,   522) =   65.51
       Model |  1958.4824      3  652.827466            Prob > F      =  0.0000
    Residual |  5201.93189   522  9.96538677            R-squared     =  0.2735
-------------+------------------------------           Adj R-squared =  0.2693
       Total |  7160.41429   525  13.6388844            Root MSE      =  3.1568

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   2.205731   .2799592     7.88   0.000     1.655746    2.755716
             |
    educ_cat |
          2  |   1.763788   .3444719     5.12   0.000     1.087066    2.440509
          3  |   4.593163   .4335902    10.59   0.000     3.741367    5.444959
             |
       _cons |   2.839769   .3313102     8.57   0.000     2.188904    3.490634
------------------------------------------------------------------------------
```

- edu_cat = 1 (高卒未満)をベースカテゴリーにしていますね。
- この結果が、33ページのものと全く同じになっていることを確認してください。
- この i. という機能はカテゴリカル変数（ダミー変数も含みます）に使うことができます。
- カテゴリカル変数からダミー変数を自動的に作り、それを使って推定してくれるという優れものです。
- 後に扱うトピックである「交差項」のところで活躍しますので、覚えておいてください。

37

ダミー変数の使い方：序数的な情報への対応

- 情報の中には「序数的」なものがあります。
- **序数とは？（反対語は「基数」：物事の数量を表す）**
- 単に順番（序数）を表すもの。
- 2は1より「大きい」、1は3より「小さい」、っていうだけの情報。
- 1と2の差は1、1と3の差は2、とかっていうことじゃありません。
- 例えば、さっきのEDU_CATは序数的な変数です。
- EDU_CAT= 1は高卒未満（小さい）、EDU_CAT= 2 は高卒（大きい）、EDU_CAT= 3 は大卒以上（もっと大きい）。
- でも、この値そのものや、値の差に数量的な意味はありません。

38

- 別の例として、、、、

この人の「身体的魅力」を五段階評価して下さい。1はもっとも悪い、5は最も良い、とします。（1, 2, 3, 4, 5）

という質問があったとしましょう。

- この質問から作られた変数も序数的と言えます。
- 魅力が1とか、、、この1に数量的な意味合いはありません。
- 2と1なら、2の方が魅力度が高い、っていうだけですよね。
- だから、数字を変えて、2、3、6、7、9にしたって同じ情報が伝わります。

39

- 身体的魅力が労働市場で有利に働いているかどうかを分析してみましょう。
- データ: BEAUTY.DAT. 出典：Wooldridge, J.M. (2006) *Introductory Econometrics*, Thomson, South-western
- このデータはHamermesh and Biddle (1994) "Beauty and the Labor Market," American Economic Review, 84, 1174-1194で使われたもの
- サンプルの中にいるそれぞれの人が調査員により5段階評価されています、身体的な魅力について。

  [homely（1）, quite plain（2）, average（3）, good looking（4）, strikingly beautiful or handsome（5）]

- この身体的魅力の序数的な変数（$LOOKS$）に加えて、賃金（$WAGE$）、教育水準（$EDUC$）、労働市場での経験年数（$EXPER$）、、、、などの変数がデータに含まれています。
- モデル、、、

40

$$\ln(WAGE) = \beta_0 + \beta_1 LOOKS + \cdots$$

- このモデルは、避けた方が良いです。
- この特定化は、homely（1）からquite plain（2）に変化したときと、good looking（4）からstrikingly beautiful （5）に変化したとき、賃金の変化は同じ、と仮定していることになります。
- そうとは言い切れないですよね。
- 従って、このような序数的な情報は、複数のダミー変数を使って捉えるのが一般的です（さっきの$EDU_CAT \rightarrow JH$、H、UN）。
- ここでは、例えば、

$$L1 = \begin{cases} 1 & homely \\ 0 & \text{それ以外} \end{cases} \quad L2 = \begin{cases} 1 & quite\ plain \\ 0 & \text{それ以外} \end{cases} \quad L3 = \begin{cases} 1 & average \\ 0 & \text{それ以外} \end{cases}$$

$$L4 = \begin{cases} 1 & good\ looking \\ 0 & \text{それ以外} \end{cases} \quad L5 = \begin{cases} 1 & str.beautiful \\ 0 & \text{それ以外} \end{cases}$$

とでもして、、、

41

$$\ln(WAGE) = \beta_0 + \beta_1 L2 + \beta_2 L3 + \beta_3 L4 + \beta_4 L5 \cdots$$

のようなモデルの方がベターです。

それではデータを見てみましょう。

tabulateコマンド、このように度数分布を与えます。

```
. tabulate looks
```

from 1 to 5	Freq.	Percent	Cum.
1	13	1.03	1.03
2	142	11.27	12.30
3	722	57.30	69.60
4	364	28.89	98.49
5	19	1.51	100.00
Total	1,260	100.00	

tabulateコマンド、変数二個並べると、クロス集計してくれます。

```
. tabulate looks female
```

from 1 to 5	=1 if female 0	1	Total
1	8	5	13
2	88	54	142
3	489	233	722
4	228	136	364
5	11	8	19
Total	824	436	1,260

tabulateを使って、データを理解すること、重要です。

42

```
. tabulate looks female
```

from 1 to 5	=1 if female 0	1	Total
1	8	5	13
2	88	54	142
3	489	233	722
4	228	136	364
5	11	8	19
Total	824	436	1,260

- 残念な感じの人、かなり少ないですね。これを一つのカテゴリーとして分析しても、、、
- また、いい意味で超絶の人も、かなり少ないですね。これを一つのカテゴリーとして分析しても、、、
- こういう時は、二つのカテゴリーをくっつけて一つにする、ことがよくあります。
- homelyとquite plainをくっつける（below averageカテゴリー）、good lookingとstrikingly beautiful （above average カテゴリー）をくっつける、なんてのは一つの手です。

43

$$\ln(WAGE) = \beta_0 + \beta_1 BLOWAVE + \beta_2 ABOVEAVE + \cdots$$

ここでは$AVERAGE$をベースカテゴリーにします。

- ですから、β_1の解釈は、「$AVERAGE$の人に比べて、$BLOWAVE$の人の賃金は、、、」、、、

β_2の解釈は、「$AVERAGE$の人に比べて、$ABOVEWAVE$の人の賃金は、、、」

となります。

- あとは、各自モデルを設定して自由に分析してみましょう。

44