University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

5. Regression Models with Dummy Explanatory Variables

Prof. Dr. Miroslav Verbič

miroslav.verbic@ef.uni-lj.si

www.miroslav-verbic.si

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION AMBA ACCREDITED

Ljubljana, October 2022

Basic definitions

DUMMY VARIABLES

↳ categorical variables with 2 possible values (for example gender)

The following names occur for such a variable:

- **dummy**
- **binary**
- **dihotomous**
- **indicator**
- **artificial**

Observed unit has the analysed property → **Value 1**

Observed unit does not have the analysed property → **Value 0**

Basic definitions

WARNINGS

1. **A categorical variable (attribute) with *m* possible values requires (*m* – 1) dummy variables.**

   "Dummy trap" – Perfect multicollinearity!

2. **Value of the categorical variable that is assigned value 0 for the dummy variable, is called the base (benchmark, control, reference) value.**

3. **Switching the assigned values (0 to 1 and vice versa) for dummy variables does not affect the (absolute) values of regression coefficients.**

Regression models with dummy explanatory variables

Dummy trap:

gender	D_1	D_2
M	1	0
F	0	1

Slide 2.

$D_{1i} + D_{2i} = 1, \forall i$

One can write: $D_{1i} = 1 - D_{2i}$ or $D_{2i} = 1 - D_{1i}$, resulting in perfect multicollinearity.

↳ dummy trap

to avoid this we always use $m - 1$ values of dummy variables (for example gender 1 - male, 0 - female)

A, B and C are ANOVA models

Analysis of variance models

ANALYSIS OF VARIANCE (ANOVA) MODELS

A Regression model with one attribute with two possible values

Types of regression models with dummy explanatory variables

A
- y – gross wage
- gender (male, female): D = { 0; female, 1; male

$$y_i = \beta_1 + \beta_2 D_i + u_i$$

$$E(y_i | D_i = 0) = \beta_1$$
$$E(y_i | D_i = 1) = \beta_1 + \beta_2$$

⇒ average gross wage females

⇒ average gross wage males

this is how the model looks like

$H_0 : \beta_2 = 0 \ ; \ H_1 : \beta_2 \neq 0$

$|t(b_2)| > t_c$

we can reject the H_0.

β_2 is the difference between the 2 genders.

Analysis of variance models

B Regression model with one attribute with several (three) possible values

$$y_i = \beta_1 + \beta_2 D_{1i} + \beta_3 D_{2i} + u_i$$

B
- y – gross wage
- education (elementary school, college, university):

$D_1 = \begin{cases} 0; & \text{other} \\ 1; & \text{college} \end{cases}$ $D_2 = \begin{cases} 0; & \text{other} \\ 1; & \text{university} \end{cases}$

elementary school is the base value

$$E(y_i | D_{1i} = 0, D_{2i} = 0) = \beta_1$$ → average gross wage of people with elementary school

$$E(y_i | D_{1i} = 1, D_{2i} = 0) = (\beta_1 + \beta_2)$$ → average gross wage of people with college education

$$E(y_i | D_{1i} = 0, D_{2i} = 1) = (\beta_1 + \beta_3)$$ → average gross wage of people with university education

? $H_0 : \beta_2 = 0 \; ; \; H_1 : \beta_2 \neq 0$ $\qquad$ $H_0 : \beta_3 = 0 \; ; \; H_1 : \beta_3 \neq 0$

i $|t(b_2)| > t_c$ $\qquad$ $|t(b_3)| > t_c$

$\beta_3 - \beta_1$

average difference in gross wages between people with uni. education and people with elementary education.

Analysis of variance models

C Regression model with two attributes with two values each

$$y_i = \beta_1 + \beta_2 D_{1i} + \beta_3 D_{2i} + u_i$$

$$E(y_i | D_{1i} = 0, D_{2i} = 0) = \beta_1$$ → average savings of unemployed females

$$E(y_i | D_{1i} = 1, D_{2i} = 0) = (\beta_1 + \beta_2)$$

$$E(y_i | D_{1i} = 0, D_{2i} = 1) = (\beta_1 + \beta_3)$$

$$E(y_i | D_{1i} = 1, D_{2i} = 1) = (\beta_1 + \beta_2 + \beta_3)$$ → average savings of an employed male

? $H_0 : \beta_2 = 0 \ ; \ H_1 : \beta_2 \neq 0$ $\quad$ $H_0 : \beta_3 = 0 \ ; \ H_1 : \beta_3 \neq 0$

$|t(b_2)| > t_c$ $\quad$ $|t(b_3)| > t_c$

average difference in savings between unemployed females and unemployed males

(C)

- y – savings
- gender (male, female):

$D_1 = \begin{cases} 0; \text{ female} \\ 1; \text{ male} \end{cases}$

- labour force participation (employed, unemployed):

$D_2 = \begin{cases} 0; \text{ unemployed} \\ 1; \text{ employed} \end{cases}$

Analysis of covariance models

ANALYSIS OF COVARIANCE (ANCOVA) MODELS

D **Regression model with one numerical explanatory variable and one attribute with two possible values**

- y – gross wage
- gender (male, female): $D = \begin{cases} 0; & \text{female} \\ 1; & \text{male} \end{cases}$
- x_3 – years of employment

$$y_i = \beta_1 + \beta_2 D_i + \beta_3 x_{3i} + u_i$$

$$E(y_i | D_i = 0, x_{3i}) = \beta_1 + \beta_3 x_{3i}$$

$$E(y_i | D_i = 1, x_{3i}) = (\beta_1 + \beta_2) + \beta_3 x_{3i}$$

expected gross wage of females conditional on their years of employment

?

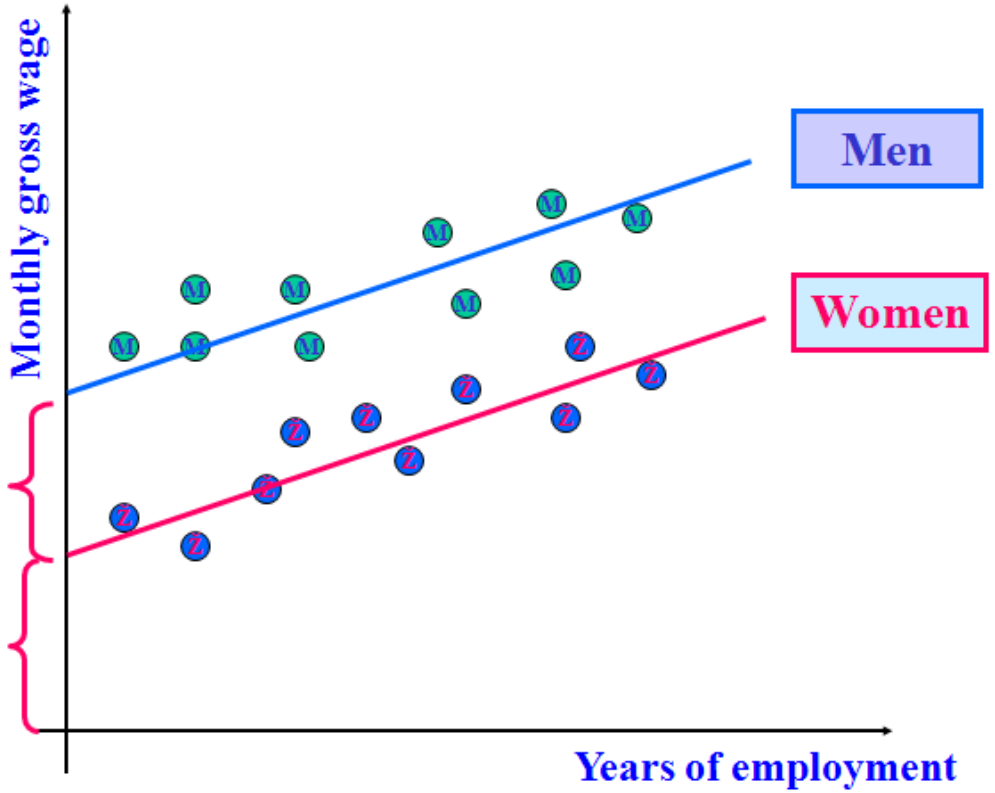
$$H_0 : \beta_2 = 0 \ ; \ H_1 : \beta_2 \neq 0$$

i

$$|t(b_2)| > t_c$$

Analysis of covariance models

average difference between the wages of males and females



Regression coefficient β_2 is also called the *differential intercept coefficient.*

Analysis of covariance models

E **Regression model with one numerical explanatory variable and one attribute with several (three) possible values**

$$y_i = \beta_1 + \beta_2 D_{1i} + \beta_3 D_{2i} + \beta_4 x_{4i} + u_i$$

expected gross wage of people with elementary education

$$E(y_i | D_{1i} = 0, D_{2i} = 0, x_{4i}) = \beta_1 + \beta_4 x_{4i}$$

conditional on their years of employment

$$E(y_i | D_{1i} = 1, D_{2i} = 0, x_{4i}) = (\beta_1 + \beta_2) + \beta_4 x_{4i}$$

$$E(y_i | D_{1i} = 0, D_{2i} = 1, x_{4i}) = (\beta_1 + \beta_3) + \beta_4 x_{4i}$$

? $H_0 : \beta_2 = 0 \ ; \ H_1 : \beta_2 \neq 0$ $\qquad$ $H_0 : \beta_3 = 0 \ ; \ H_1 : \beta_3 \neq 0$

i $|t(b_2)| > t_c$ $\qquad$ $|t(b_3)| > t_c$

(E)

- y – gross wage
- education (elementary school, college, university):

$D_1 = \begin{cases} 0; \text{ other} \\ 1; \text{ college} \end{cases}$ $D_2 = \begin{cases} 0; \text{ other} \\ 1; \text{ university} \end{cases}$

- x_4 – years of employment

Analysis of covariance models

F **Regression model with one numerical explanatory variable and two attributes with two values each**

$$y_i = \beta_1 + \beta_2 D_{1i} + \beta_3 D_{2i} + \beta_4 x_{4i} + u_i$$

$$E(y_i | D_{1i} = 0, D_{2i} = 0, x_{4i}) = \beta_1 + \beta_4 x_{4i}$$

$$E(y_i | D_{1i} = 1, D_{2i} = 0, x_{4i}) = (\beta_1 + \beta_2) + \beta_4 x_{4i}$$

$$E(y_i | D_{1i} = 0, D_{2i} = 1, x_{4i}) = (\beta_1 + \beta_3) + \beta_4 x_{4i}$$

$$E(y_i | D_{1i} = 1, D_{2i} = 1, x_{4i}) = (\beta_1 + \beta_2 + \beta_3) + \beta_4 x_{4i}$$

average expected savings of single women conditioned on years of employment

conditioned on their years of employment

average expected savings of married men

? $H_0 : \beta_2 = 0 \ ; \ H_1 : \beta_2 \neq 0$ $\quad$ $H_0 : \beta_3 = 0 \ ; \ H_1 : \beta_3 \neq 0$

$|t(b_2)| > t_c$ $\quad$ $|t(b_3)| > t_c$

(F)

- y – savings
- gender (male, female):

$$D_1 = \begin{cases} 0; \text{ female} \\ 1; \text{ male} \end{cases}$$

- marital status (single, married):

$$D_2 = \begin{cases} 0; \text{ single} \\ 1; \text{ married} \end{cases}$$

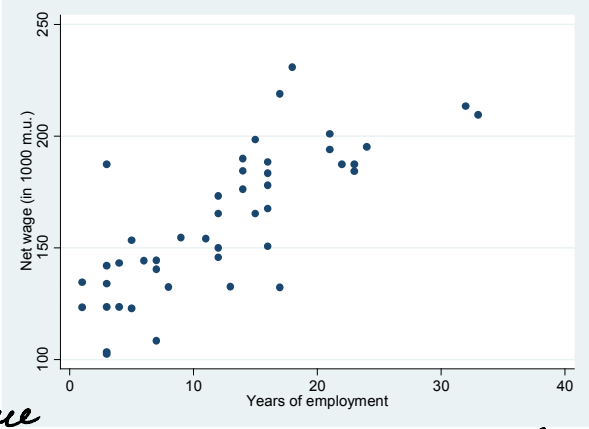
- x_4 – years of employment

Example 1: For 45 employees we gathered data in the file `wage1.dta` on their net wage (in 1,000 m.u.), period of employment (in years) and gender (dummy variable, which equals 1 for males and 2 for females). For gender we introduce a new dummy variable *D*, which has value of 1 for males and value of 0 for females.

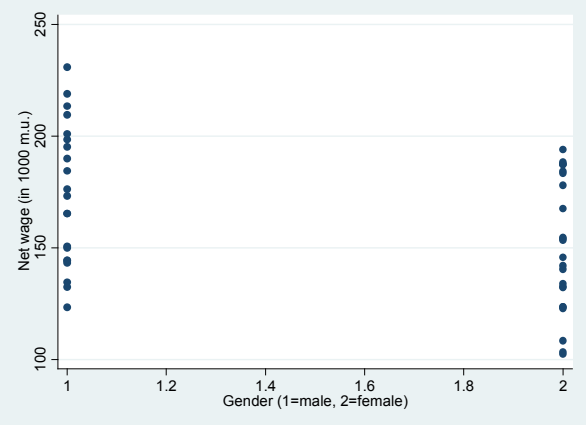
Estimate the regression models in which the dummy variable will appear in various ways and explain the obtained regression coefficients.

Computer printout of the results in Stata:

```
. scatter wage employment
```




```
. scatter wage gender
```



to introduce a dummy variable

```
. gen d=1
. replace d=0 if gender==2
(24 real changes made)

. label variable d "Gender of the employee (1=male, 0=female)"

. gen dalt=1
. replace dalt=0 if gender==1
(21 real changes made)

. label variable dalt "Gender of the employee (1=female, 0=male)"

. regress wage d

      Source |       SS       df       MS              Number of obs =      45
-------------+------------------------------           F(  1,    43) =    5.90
       Model |  5549.44883     1  5549.44883            Prob > F      =  0.0194
    Residual |  40416.5491    43  939.919746            R-squared     =  0.1207
-------------+------------------------------           Adj R-squared =  0.1003
       Total |  45965.9979    44  1044.68177            Root MSE      =  30.658

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           d |   22.25953   9.160863     2.43   0.019     3.784886    40.73417
       _cons |   151.3167   6.258061    24.18   0.000     138.6961    163.9372
------------------------------------------------------------------------------
```

a) $wage = f(gender)$:

$wage_i = \beta_1 + \beta_2 D_i + u_i$ (PRM)

$\widehat{wage}_i = b_1 + b_2 D_i$ (SRM)

$\widehat{wage}_i = 151.32 + 22.26 D_i$ – female

$p(b_2) = 0.019 < \alpha = 0.05.$

1

There is statistical significance since the p value is lower than 0,05 => there is a difference between male and female average wages.

```
. tab d, sum(wage)

  Gender of |
        the |
   employee |
    (1=male, | Summary of Net wage (in 1000 m.u.)
  0=female) |        Mean   Std. Dev.       Freq.
------------+------------------------------------
          0 |   151.31667   29.742015          24
          1 |   173.57619   31.678887          21
------------+------------------------------------
      Total |   161.70444   32.321537          45
```

```
. regress wage employment d
```

b)

```
      Source |       SS       df       MS              Number of obs =      45
-------------+------------------------------           F(  2,    42) =   36.07
       Model |     29052.7     2    14526.35           Prob > F      =  0.0000
    Residual |  16913.2979    42  402.697568           R-squared     =  0.6320
-------------+------------------------------           Adj R-squared =  0.6145
       Total |  45965.9979    44  1044.68177           Root MSE      =  20.067

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  employment |   2.975623   .3894964     7.64   0.000     2.189588    3.761659
           d |   12.67731   6.126038     2.07   0.045      .314466    25.04015
       _cons |   118.9568    5.89244    20.19   0.000     107.0653    130.8482
------------------------------------------------------------------------------
```

```
. gen demployment=d*employment

. regress wage employment d demployment
```

c)

```
      Source |       SS       df       MS              Number of obs =      45
-------------+------------------------------           F(  3,    41) =   23.48
       Model |  29052.7495     3  9684.24985           Prob > F      =  0.0000
    Residual |  16913.2484    41  412.518253           R-squared     =  0.6320
-------------+------------------------------           Adj R-squared =  0.6051
       Total |  45965.9979    44  1044.68177           Root MSE      =  20.311

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  employment |   2.970835   .5886187     5.05   0.000     1.782095    4.159574
           d |   12.57033   11.56706     1.09   0.284    -10.78982    35.93047
 demployment |    .008684   .7926433     0.01   0.991    -1.592092     1.60946
       _cons |   119.0088   7.626532    15.60   0.000     103.6067    134.4109
------------------------------------------------------------------------------
```

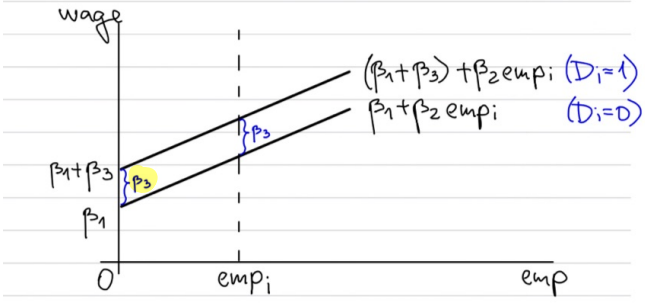
dummy variable is not statistically significant

```
. list wage employment d demployment

     +---------------------------------+
     |  wage   employ~t   d   demplo~t |
     |---------------------------------|
  1. | 150.6         16   1         16 |
  2. | 213.4         32   1         32 |
  3. | 108.4          7   0          0 |
  4. | 123.6          4   0          0 |
  5. |   194         21   0          0 |
     |---------------------------------|
  6. | 154.1         11   0          0 |
  7. | 184.5         14   1         14 |
  8. | 173.2         12   1         12 |
  9. | 167.5         16   0          0 |
     +---------------------------------+
```

b) $wage = f(emp, gender)$:

$wage_i = \beta_1 + \beta_2 emp_i + \beta_3 D_i + u_i$ (PRM)



↳ look at next page

$\widehat{wage}_i = b_1 + b_2\, emp_i + b_3 D_i$ (SRM)

$\widehat{wage}_i = 118.96 + 2.976\, emp_i + 12.677\, D_i$

$p(b_3) = 0.045 < \alpha = 0.05.$

Interpretation: PDF, p. 2.

b_2: If the period of employment increases by 1 year, then the net wage will increase on average by 2,976 m.u., for both genders. ceteris paribus is not necessary because there are no other variables

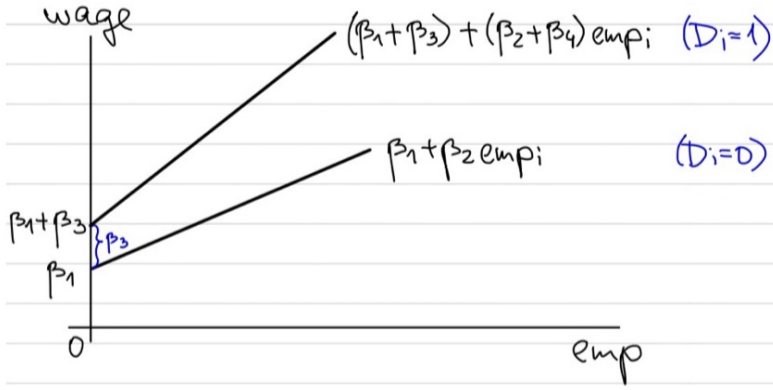
b_3: The expected net wage estimate for males is 12,677 m.u. higher than the expected net wage estimate for females, conditional on their period of employment. (which? any!)

c) The effect of period of employment on net wage depends on gender:

$wage_i = \beta_1 + \beta_2\, emp_i + \beta_3 D_i + \beta_4 D_i \cdot emp_i + u_i$ (PRM)

multiplicative or interaction dummy variable

↳ by doing this you allow for different slopes (nakon)



$\widehat{wage}_i = b_1 + b_2\,emp_i + b_3 D_i + b_4 D_i \cdot emp_i$ (SRM)

$\widehat{wage}_i = 119.01 + 2.971\,emp_i + 12.570\,D_i + 0.009\,D_i \cdot emp_i$

↳ reg. coef. of dummy variable

How come?

$p(b_3) = 0.284 > \alpha = 0.05.$ PDF, p. 2.

$p(b_4) = 0.991 > \alpha = 0.05.$

Interpretation (for teaching purposes):

b_2: If the period of employment increases by 1 year, then the net wage of females will increase on average by 2,971 m.u.

b_2+b_4: If the period of employment increases by 1 year, then the net wage of males will increase on average by 2,980 m.u.

↳ if it were statistically significant this is how you would interpret it.

b_4: If the period of employment increases by 1 year, then the net wage of males increases on average by 9 m.u. more than that of females.

b_3: For persons with zero years of employment ($empj=0$), the expected net wage of males is 12,570 m.u. higher than that of females.

10.	144.3	6	1	6
11.	165.3	15	1	15
12.	103.4	3	0	0
13.	154.7	9	0	0
14.	219	17	1	17
15.	188.4	16	0	0
16.	201	21	1	21
17.	153.4	5	0	0
18.	132.7	13	0	0
19.	183.4	16	0	0
20.	165.3	12	1	12
21.	187.4	22	0	0
22.	132.4	17	0	0
23.	123.6	3	0	0
24.	176.3	14	1	14
25.	187.4	23	0	0
26.	134.5	1	1	1
27.	102.3	3	0	0
28.	198.4	15	1	15
29.	150	12	1	12
30.	140.4	7	0	0
31.	184.3	23	0	0
32.	143.1	4	1	4
33.	187.5	3	0	0
34.	132.5	8	1	8
35.	190	14	1	14
36.	145.7	12	0	0
37.	123.4	1	1	1
38.	142	3	0	0
39.	195.3	24	1	24
40.	123	5	0	0
41.	144.4	7	1	7
42.	134	3	0	0
43.	231	18	1	18
44.	178	16	0	0
45.	209.6	33	1	33

d) Dummy trap:

$D_{alt} = \begin{cases} 0; \text{ males} \\ 1; \text{ females} \end{cases}$

$D + D_{alt} = 1 \Rightarrow$ perfect multicollinearity.

PDF, pp. 3-4.

```
. regress wage d dalt
note: dalt omitted because of collinearity

      Source |       SS       df       MS              Number of obs =      45
-------------+------------------------------           F(  1,    43) =    5.90
       Model |  5549.44883     1  5549.44883            Prob > F      =  0.0194
    Residual |  40416.5491    43  939.919746            R-squared     =  0.1207
-------------+------------------------------           Adj R-squared =  0.1003
       Total |  45965.9979    44  1044.68177            Root MSE      =  30.658

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           d |   22.25953   9.160863     2.43   0.019     3.784886    40.73417
        dalt |  (omitted)
       _cons |   151.3167   6.258061    24.18   0.000     138.6961    163.9372
------------------------------------------------------------------------------
```

Stata figured out that we had multicollinearity and it dropped the last variable

```
. regress wage employment d dalt
note: dalt omitted because of collinearity

      Source |       SS       df       MS              Number of obs =      45
-------------+------------------------------           F(  2,    42) =   36.07
       Model |     29052.7     2    14526.35           Prob > F      =  0.0000
    Residual |  16913.2979    42  402.697568           R-squared     =  0.6320
-------------+------------------------------           Adj R-squared =  0.6145
       Total |  45965.9979    44  1044.68177           Root MSE      =  20.067

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  employment |   2.975623   .3894964     7.64   0.000     2.189588    3.761659
           d |   12.67731   6.126038     2.07   0.045      .314466    25.04015
        dalt |  (omitted)
       _cons |   118.9568    5.89244    20.19   0.000     107.0653    130.8482
------------------------------------------------------------------------------
```

■

Example 2: We gathered a sample of data for 32 European countries for the year 2003. We have the following variables available (the data are provided in Stata Data file `health.dta`, while the programming code is given in Stata Do file `health-commands-l12.do`):

- life expectancy at birth (*LIFE*; in years);
- health expenditure per capita (*EXP*; in U.S. dollars);
- percentage of smokers among adults (*TOBACCO*);
- consumption of alcohol per capita (*ALCO*; in liters of distilled spirits).

We divided the countries into two groups based on whether they are EU15 member states (in this case the dummy variable *DEU* equals 1) or not (in this case the dummy variable *DEU* equals 0).

We estimated separately for each group the following regression model:

$$LIFE_i = \beta_1 + \beta_2 EXP_i + \beta_3 ALCO_i + \beta_4 TOBACCO_i + u_i,$$

and found based on the Chow test that the analyzed regression function differs between the aforementioned groups of countries. Fill in the findings with the use of dummy variables.

Computer printout of the results in Stata:

```
. regress life exp alco tobacco

      Source |       SS       df       MS              Number of obs =      32
-------------+------------------------------           F(  3,    28) =   26.30
       Model |  413.850212     3  137.950071           Prob > F      =  0.0000
    Residual |  146.874565    28  5.24552017           R-squared     =  0.7381
-------------+------------------------------           Adj R-squared =  0.7100
       Total |  560.724777    31   18.087896           Root MSE      =  2.2903
```

smoking and alcohol consumption reduce the average life expectancy

```
------------------------------------------------------------------------------
        life |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   .0018569   .0004023     4.62   0.000     .0010329    .0026809
        alco |  -.6493606   .2805689    -2.31   0.028     -1.22408   -.0746412
     tobacco |  -.2238391   .0837702    -2.67   0.012    -.3954346   -.0522436
       _cons |   81.42053   2.720683    29.93   0.000     75.84746    86.99359
------------------------------------------------------------------------------
```

```
. gen dexp=deu*exp
. gen dalco=deu*alco
. gen dtobacco=deu*tobacco
```

```
. regress life exp alco tobacco deu
      Source |       SS       df       MS              Number of obs =      32
-------------+------------------------------           F(  4,    27) =   24.56
       Model |  439.854433     4  109.963608            Prob > F      =  0.0000
    Residual |  120.870344    27  4.47667939            R-squared     =  0.7844
-------------+------------------------------           Adj R-squared =  0.7525
       Total |  560.724777    31   18.087896            Root MSE      =  2.1158

------------------------------------------------------------------------------
        life |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   .0015803   .0003889     4.06   0.000     .0007823    .0023783
        alco |  -.4965685   .2668332    -1.86   0.074    -1.044065     .050928
     tobacco |  -.2039382   .0778272    -2.62   0.014    -.3636265   -.0442499
         deu |   2.144349   .8897162     2.41   0.023     .3188019    3.969896
       _cons |   79.81336   2.600354    30.69   0.000     74.47788    85.14885
------------------------------------------------------------------------------
```

average difference between EU countries and countries which are not in EU is 2,14 years

```
. regress life exp alco tobacco deu dexp dalco dtobacco
      Source |       SS       df       MS              Number of obs =      32
-------------+------------------------------           F(  7,    24) =   23.59
       Model |  489.576762     7  69.9395375            Prob > F      =  0.0000
    Residual |  71.1480148    24  2.96450062            R-squared     =  0.8731
-------------+------------------------------           Adj R-squared =  0.8361
       Total |  560.724777    31   18.087896            Root MSE      =  1.7218

------------------------------------------------------------------------------
        life |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   .0015388   .0003798     4.05   0.000     .0007548    .0023227
        alco |  -.5335537   .2317374    -2.30   0.030    -1.011836   -.0552712
     tobacco |  -.4019988   .0917503    -4.38   0.000    -.5913622   -.2126354
         deu |  -7.934472   4.162727    -1.91   0.069    -16.52592    .6569745
       dexp |  -.0014602   .0007901    -1.85   0.077    -.0030908    .0001705
      dalco |   .6710069   .8493576     0.79   0.437    -1.081981    2.423995
   dtobacco |   .4129869   .1349227     3.06   0.005     .1345201    .6914537
       _cons |   85.90512   3.096202    27.75   0.000     79.51487    92.29537
------------------------------------------------------------------------------
```

next page *

```
. test deu=dexp=dalco=0
 ( 1)  deu - dexp = 0
 ( 2)  deu - dalco = 0
 ( 3)  deu = 0

       F(  3,    24) =    4.20
            Prob > F =    0.0159
```

use this to test whether you can drop several variables from the model

E2 Application of dummy explanatory variables:

$LIFE_i = \beta_1 + \beta_2 EXP_i + \beta_3 ALCO_i + \beta_4 TOBACCO_i + \beta_5 DEU_i + u_i$

PDF, pp. 4-5.

$b_5 = 2.14$

$p(b_5) = 0.023 < \alpha = 0.05$

Are the differences between the two groups of countries really only in the average life expectancy at birth (or) perhaps the values of some of the regression coefficients also depend on placing the country in one of the two groups?

$LIFE_i = \beta_1 + \beta_2 EXP_i + \beta_3 ALCO_i + \beta_4 TOBACCO_i + \beta_5 DEU_i + \beta_6 DEU_i \cdot EXP_i + \beta_7 DEU_i \cdot ALCO_i + \beta_8 DEU_i \cdot TOBACCO_i + u_i$

PDF, p. 5

. **test dexp=dalco=0** — should we exclude only β_6 and β_7

```
 ( 1)  dexp - dalco = 0
 ( 2)  dexp = 0

       F(  2,    24) =    2.20
            Prob > F =    0.1332
```

. **regress life exp alco tobacco deu dtobacco**

```
      Source |       SS       df       MS              Number of obs =      32
-------------+------------------------------           F(  5,    26) =   29.44
       Model |  476.562498     5  95.3124995           Prob > F      =  0.0000
    Residual |  84.1622793    26  3.23701074           R-squared     =  0.8499
-------------+------------------------------           Adj R-squared =  0.8210
       Total |  560.724777    31   18.087896           Root MSE      =  1.7992

------------------------------------------------------------------------------
        life |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   .0012305   .0003467     3.55   0.001      .000518    .0019431
        alco |  -.5723521    .228013    -2.51   0.019     -1.04104   -.1036647
     tobacco |  -.4311841   .0945176    -4.56   0.000    -.6254678   -.2369004
         deu |  -10.55756   3.847028    -2.74   0.011    -18.46524    -2.64988
   dtobacco |   .4550991   .1351442     3.37   0.002     .1773063    .7328919
       _cons |   87.19745   3.114081    28.00   0.000     80.79636    93.59853
------------------------------------------------------------------------------
```

effect of smoking in EU member states

12 PIT - testiranje na roke ■

* 11

$$LIFE_i = \beta_1 + \beta_2 EXP_i + \beta_3 ALCO_i + \beta_4 TOBACCO_i + \beta_5 DEU_i + \beta_6 DEU_i \cdot EXP_i + \beta_7 DEU_i \cdot ALCO_i + \beta_8 DEU_i \cdot TOBACCO_i + u_i$$

PDF, p.5.

dtobacco minus tobacco

$b_8 = 0.41$

$p(b_8) = 0.005 < \alpha = 0.05$

Should we exclude all the insignificant relationship related to dummy explanatory variables from the model?

$H_0: \beta_j = 0, \forall j = 5,6,7$ $\quad F = 4.20$

$H_1: \beta_j \neq 0, \exists j = 5,6,7$ $\quad p(F) = 0.02$ No.

- if we want to drop variables out of the model we need to test them.

If we want to drop 3 coefficients out of the model we need to test them jointly.

This is fairly obvious. Why?

So, we should probably exclude only β_6 and β_7:

$H_0: \beta_j = 0, \forall j = 6,7$ $\quad F = 2.20$

$H_1: \beta_j \neq 0, \exists j = 6,7$ $\quad p(F) = 0.13$ Yes.

The final model should then be:

$$LIFE_i = \beta_1 + \beta_2 EXP_i + \beta_3 ALCO_i + \beta_4 TOBACCO_i + \beta_5 DEU_i + \beta_6 DEU_i \cdot TOBACCO_i + u_i$$

$p(b_5) = 0.011 < \alpha = 0.05$ PDF, p.6.

$p(b_6) = 0.002 < \alpha = 0.05$

Interpretation:

b_4: If the percentage of smokers increases by 1 p.p., then on average, ceteris paribus, the life expectancy at birth in non-EU member states decreases by 0.43 years.

$b_4 + b_6$: If the percentage of smokers increases by 1 p.p., then on average, ceteris paribus, the life expectancy at birth in EU member states does not change ($-0.43 + 0.45 = 0.02$, not statistically significant).

Comparison of two regression models

$$y_i = \beta_1 + \beta_2 x_i + u_i$$

a) Chow test

$$F = \frac{(RSS - RSS_1 - RSS_2)/k}{(RSS_1 + RSS_2)/(n_1 + n_2 - 2k)}$$

b) Application of dummy variables

Multiplicative or interaction dummy variable

$$y_i = \beta_1 + \beta_2 D_i + \beta_3 x_i + \beta_4 (D_i x_i) + u_i$$

$$E(y_i | D_i = 0, x_i) = \beta_1 + \beta_3 x_i$$

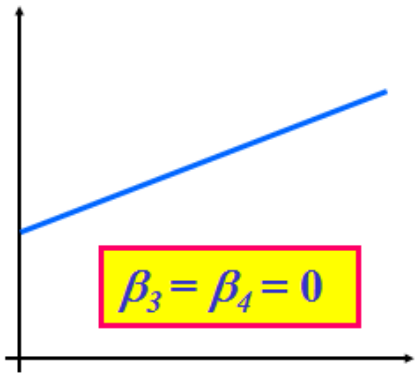
$$E(y_i | D_i = 1, x_i) = (\beta_1 + \beta_2) + (\beta_3 + \beta_4) x_i$$

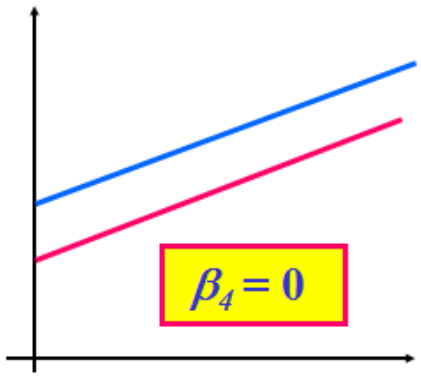
$H_0 : \beta_2 = 0 \; ; \; H_1 : \beta_2 \neq 0$ $\qquad$ $H_0 : \beta_4 = 0 \; ; \; H_1 : \beta_4 \neq 0$

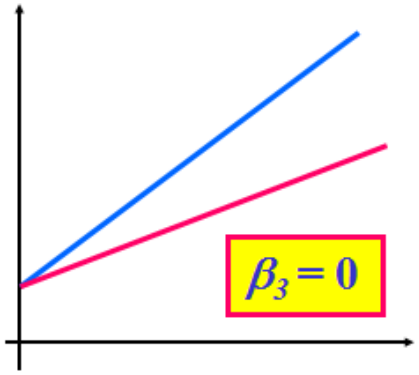
Comparison of two regression models

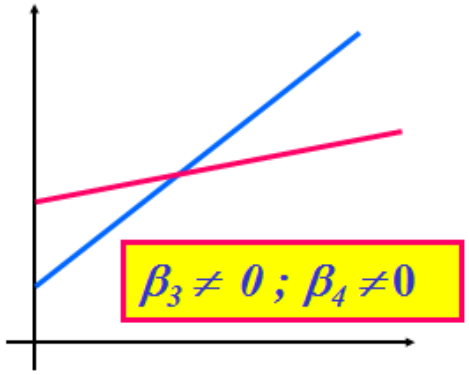
Application of one dummy variable offers four possibilities (models)

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 D_t + \beta_4 (x_{2t} D_t) + u_t$$





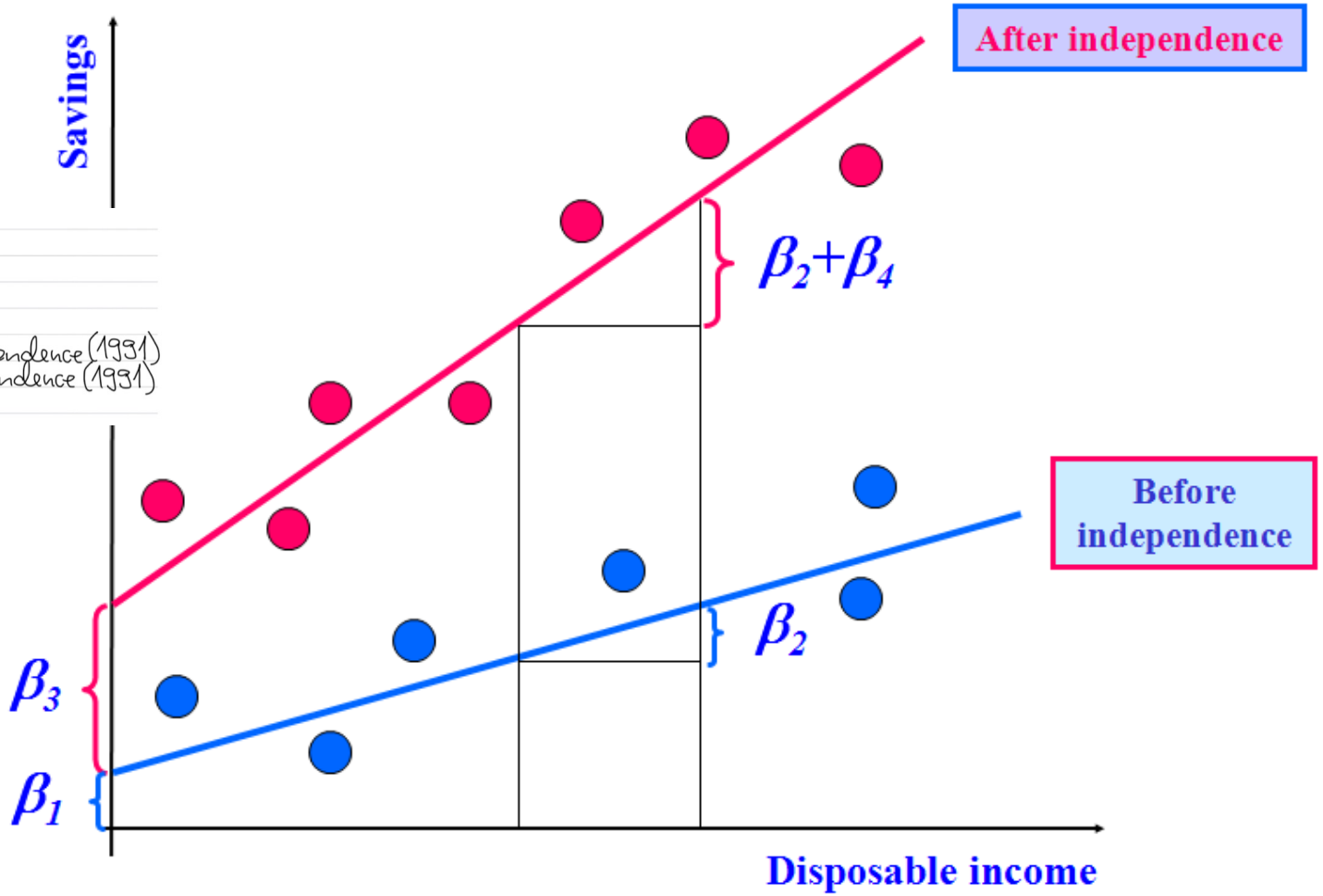




Comparison of two regression models

Slide 12:

y – savings
x_2 – disposable income
$D = \begin{cases} 0; & \text{years before the independence (1991)} \\ 1; & \text{years after the independence (1991)} \end{cases}$



Comparison of two regression models

Advantages of dummy variables compared to statistical tests (Chow test)

1. **Only one regression model is estimated**

2. **Differences in one or more regression coefficients allowed**

   β_3 – differential intercept coefficient – with the dummy
   β_4 – differential slope coefficient

Slide 13:

Chow test: $F \sim F_{k, n_1 + n_2 - 2k}$

3. **No degrees of freedom are lost when analysing statistical significance**

Look at example 2 => slide 267

University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

5. Regression Models with Dummy Explanatory Variables

Prof. Dr. Miroslav Verbič

miroslav.verbic@ef.uni-lj.si

www.miroslav-verbic.si

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION AMBA ACCREDITED

Ljubljana, October 2022