QS101: Introduction to Quantitative Methods in Social Science

Week 19: Multivariate Regression and Regression with Categorical Variables

Dr. Florian Reiche
Teaching Fellow in Quantitative Methods
Course Director BA Politics and Sociology
Deputy Director of Student Experience and Progression

March 5th, 2015

Your Regression Models I

Regression with Categorical Variables

- Dichotomous Categorical Variables
- Regression with a 1/2/3 Variable

Your Regression Models II

Your Regression Models I

Task

- Select two of the continuous variables from last week's handout

Task

- Select two of the continuous variables from last week's handout
- Run a multiple regression by typing
  - `regress c_fimngrs_dv` *indepvar1 indepvar2*

Task

- Select two of the continuous variables from last week's handout
- Run a multiple regression by typing
  - **`regress c_fimngrs_dv`** *`indepvar1 indepvar2`*
- Interpret the results:
  - What does the constant mean?
  - What does each slope coefficient indicate?
  - Are your results significant at the 95% level, and what does this mean?

Regression with Categorical Variables

Regression with Categorical Variables

Dichotomous Categorical Variables

The Setup

- You can enter a dichotomous categorical variable just like you would with a continuous one

Source of this section: `http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter3/statareg3.htm`

The Setup

- You can enter a dichotomous categorical variable just like you would with a continuous one
- Ensure, that the coding is 0/1, as the interpretation is easier

Source of this section: `http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter3/statareg3.htm`

The Setup

- You can enter a dichotomous categorical variable just like you would with a continuous one
- Ensure, that the coding is 0/1, as the interpretation is easier
- If necessary, recode, for example: **`recode c_sex 1=0 2=1`**

Source of this section: `http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter3/statareg3.htm`

The Setup

- You can enter a dichotomous categorical variable just like you would with a continuous one
- Ensure, that the coding is 0/1, as the interpretation is easier
- If necessary, recode, for example: **`recode c_sex 1=0 2=1`**
- Run the regression: **`regress c_fimngrs_dv c_sex`**

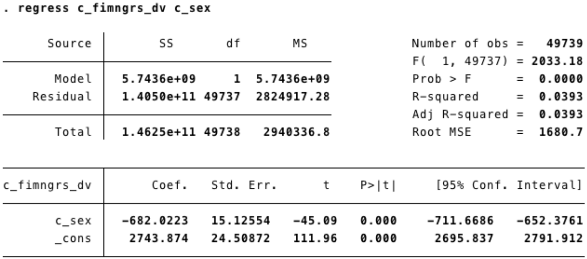
Source of this section: `http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter3/statareg3.htm`

The Setup

- You can enter a dichotomous categorical variable just like you would with a continuous one
- Ensure, that the coding is 0/1, as the interpretation is easier
- If necessary, recode, for example: **`recode c_sex 1=0 2=1`**
- Run the regression: **`regress c_fimngrs_dv c_sex`**
- The interpretation of the output is straightforward

Source of this section: `http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter3/statareg3.htm`

The Output



```
. regress c_fimngrs_dv c_sex

      Source |       SS       df       MS              Number of obs =   49739
-------------+------------------------------           F(  1, 49737) = 2033.18
       Model |  5.7436e+09     1  5.7436e+09           Prob > F      =  0.0000
    Residual |  1.4050e+11 49737  2824917.28           R-squared     =  0.0393
-------------+------------------------------           Adj R-squared =  0.0393
       Total |  1.4625e+11 49738   2940336.8           Root MSE      =  1680.7

------------------------------------------------------------------------------
c_fimngrs_dv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       c_sex |  -682.0223   15.12554   -45.09   0.000    -711.6686   -652.3761
       _cons |   2743.874   24.50872   111.96   0.000     2695.837    2791.912
------------------------------------------------------------------------------
```

Figure: Regression on the Influence of Gender on Monthly Income

Interpretation

- Remember the coding: Male=0, Female=1

Interpretation

- Remember the coding: Male=0, Female=1
- Now build the estimated regression equation

Interpretation

- Remember the coding: Male=0, Female=1
- Now build the estimated regression equation
- Male: $2743.87 - 682.02 \times 0 = 2743.87$

Interpretation

- Remember the coding: Male=0, Female=1
- Now build the estimated regression equation
- Male: $2743.87 - 682.02 \times 0 = 2743.87$
- Female: $2743.87 - 682.02 \times 1 = 2061.85$

Interpretation

- Remember the coding: Male=0, Female=1
- Now build the estimated regression equation
- Male: $2743.87 - 682.02 \times 0 = 2743.87$
- Female: $2743.87 - 682.02 \times 1 = 2061.85$
- The coefficient therefore tells you how much more or less the category coded as “1” would earn.

Regression with Categorical Variables

Regression with a 1/2/3 Variable

The Setup

- If we have a predictor with three (or more) categories, we need to transform these

The Setup

- If we have a predictor with three (or more) categories, we need to transform these
- For example: new variable **`c_rel`**

The Setup

- If we have a predictor with three (or more) categories, we need to transform these
- For example: new variable **c_rel**
- Captures 3 religious categories: Christian (1), Muslim (2), and Other (3) (source: variable **c_oprlg1**)

The Setup

- If we have a predictor with three (or more) categories, we need to transform these
- For example: new variable **`c_rel`**
- Captures 3 religious categories: Christian (1), Muslim (2), and Other (3) (source: variable **`c_oprlg1`**)
- We need to create dummy variables from **`c_rel`**:

Creating Dummy Variables

Old Variable	New Variables		
c_rel	c_rel1	c_rel2	c_rel3
1	1	0	0
1	1	0	0
1	1	0	0
2	0	1	0
2	0	1	0
2	0	1	0
3	0	0	1
3	0	0	1
3	0	0	1

Generate the Dummies

- The command is: **`tabulate`** *`oldvar`*`,` **`gen(`***`oldvar`***`)`**

Generate the Dummies

- The command is: **`tabulate`** *`oldvar`*`,` **`gen(`***`oldvar`***`)`**
- For example: **`tabulate c_rel, gen(c_rel)`**

Generate the Dummies

- The command is: **tabulate** *oldvar*, **gen(***oldvar***)**
- For example: **tabulate c_rel, gen(c_rel)**
- You can check the coding for the first ten cases by typing: **list c_rel c_rel1 c_rel2 c_rel3 in 1/10, nolabel**

Running the Regression

- You include all but one of these dummies in your regression analysis

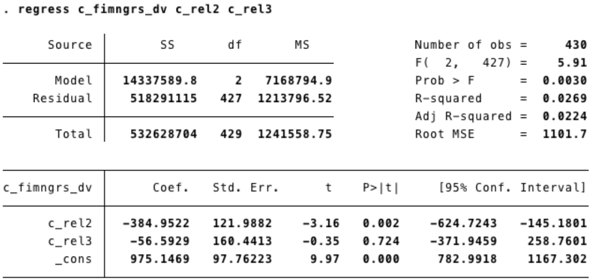
Running the Regression

- You include all but one of these dummies in your regression analysis
- This is your *reference category*

Running the Regression

- You include all but one of these dummies in your regression analysis
- This is your *reference category*
- For example: **`regress c_fimngrs_dv c_rel2 c_rel3`**

The Output



```
. regress c_fimngrs_dv c_rel2 c_rel3

      Source |       SS           df       MS      Number of obs   =       430
-------------+----------------------------------   F(  2,     427) =      5.91
       Model |  14337589.8         2   7168794.9   Prob > F        =    0.0030
    Residual |   518291115       427  1213796.52   R-squared       =    0.0269
-------------+----------------------------------   Adj R-squared   =    0.0224
       Total |   532628704       429  1241558.75   Root MSE        =    1101.7

------------------------------------------------------------------------------
c_fimngrs_dv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      c_rel2 |  -384.9522   121.9882    -3.16   0.002    -624.7243   -145.1801
      c_rel3 |   -56.5929   160.4413    -0.35   0.724    -371.9459    258.7601
       _cons |   975.1469   97.76223     9.97   0.000     782.9918    1167.302
------------------------------------------------------------------------------
```

Figure: Regression on the Influence of Religion on Monthly Income

Interpretation

- Here, **`c_rel1`** is omitted, so **`_cons`** shows the mean for a Christian person

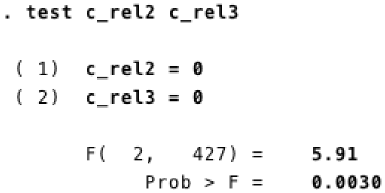
Interpretation

- Here, **`c_rel1`** is omitted, so **`_cons`** shows the mean for a Christian person
- The other coefficients tell you how much more, or less a Muslim or a person with the religion "other" earns, *relative* to a Christian person

Interpretation

- Here, **`c_rel1`** is omitted, so **`_cons`** shows the mean for a Christian person
- The other coefficients tell you how much more, or less a Muslim or a person with the religion "other" earns, *relative* to a Christian person
- Last step, test that the differences between the three groups are significant, by typing: **`test c_rel2 c_rel3`**

Regression with a 1/2/3 Variable



```
. test c_rel2 c_rel3

 ( 1)  c_rel2 = 0
 ( 2)  c_rel3 = 0

       F(  2,   427) =    5.91
            Prob > F =    0.0030
```

Figure: Test for Significant Differences of Income between Religious Groups

The Shortcut: **xi**

- We can save ourselves the faffing with generating the dummies by using the **xi** command

The Shortcut: **`xi`**

- We can save ourselves the faffing with generating the dummies by using the **`xi`** command
- For example: **`xi: regress c_fimngrs_dv i.c_rel`**

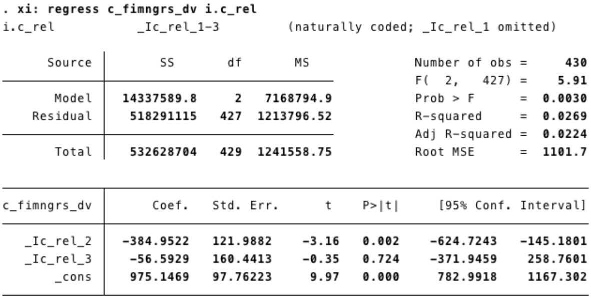
The Shortcut: `xi`

- We can save ourselves the faffing with generating the dummies by using the **xi** command
- For example: `xi: regress c_fimngrs_dv i.c_rel`
- STATA automatically leaves the first category (here: Christian) out

The Shortcut: `xi`

- We can save ourselves the faffing with generating the dummies by using the **`xi`** command
- For example: **`xi: regress c_fimngrs_dv i.c_rel`**
- STATA automatically leaves the first category (here: Christian) out
- If you want to omit a different category as your reference, you can tell STATA before running the regression: **`char c_rel[omit] 3`**

The Output

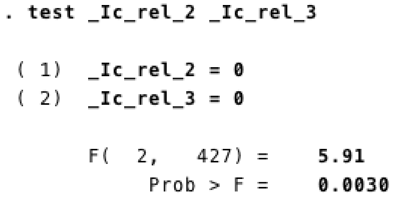


```
. xi: regress c_fimngrs_dv i.c_rel
i.c_rel           _Ic_rel_1-3         (naturally coded; _Ic_rel_1 omitted)

      Source |       SS       df       MS              Number of obs =     430
-------------+------------------------------           F(  2,   427) =    5.91
       Model |  14337589.8     2  7168794.9            Prob > F      =  0.0030
    Residual |   518291115   427  1213796.52           R-squared     =  0.0269
-------------+------------------------------           Adj R-squared =  0.0224
       Total |   532628704   429  1241558.75           Root MSE      =  1101.7

------------------------------------------------------------------------------
c_fimngrs_dv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Ic_rel_2 |  -384.9522   121.9882    -3.16   0.002    -624.7243   -145.1801
    _Ic_rel_3 |   -56.5929   160.4413    -0.35   0.724    -371.9459    258.7601
        _cons |   975.1469   97.76223     9.97   0.000     782.9918    1167.302
------------------------------------------------------------------------------
```

Figure: Regression on the Influence of Religion on Monthly Income

Regression with a 1/2/3 Variable

- The test command here is: **`test _Ic_rel2 _Ic_rel3`**



```
. test _Ic_rel_2 _Ic_rel_3

 ( 1)  _Ic_rel_2 = 0
 ( 2)  _Ic_rel_3 = 0

       F(  2,   427) =    5.91
            Prob > F =    0.0030
```

Your Regression Models II

Task

- Run a regression with a categorical variable

Task

- Run a regression with a categorical variable
- Recode the categorical variable as necessary before carrying out the **`regress`** command

Task

- Run a regression with a categorical variable
- Recode the categorical variable as necessary before carrying out the **`regress`** command
- Interpret the results:
  - What does the constant mean?
  - What does the slope coefficient indicate?
  - Are your results significant at the 95% level, and what does this mean?