Lecture 7: OLS with qualitative information

Dummy variables

- Dummy variable: an indicator that says whether a particular observation is in a category or not
  - Like a light switch: on or off
  - Most useful values: 1 & 0
- Example, predicting school attachment:
  - schattach = $\beta_1 + \beta_2$male+u
  - The variable 'male' is equal to 1 for all males, and 0 for all females.

Example, cont.

- For males: schattach-hat = $\beta_1+1^*\beta_2 = \beta_1+\beta_2 = 7.83+.17=8.00$
- For females: schattach-hat = $\beta_1+0^*\beta_2 = \beta_1 = 7.83$

```
. reg schattach male

      Source |       SS       df       MS              Number of obs =    6574
-------------+------------------------------           F(  1,  6572) =   11.12
       Model |  45.2251677     1  45.2251677           Prob > F      =  0.0009
    Residual |  26719.3529  6572  4.06563495           R-squared     =  0.0017
-------------+------------------------------           Adj R-squared =  0.0015
       Total |   26764.578  6573  4.07189686           Root MSE      =  2.0163

------------------------------------------------------------------------------
   schattach |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .1659059   .0497434     3.34   0.001     .0683925    .2634192
       _cons |   7.829004   .0354564   220.81   0.000     7.759498     7.89851
------------------------------------------------------------------------------
```

Example, cont.

- To test for significant differences between two groups, we look at the estimate and standard error for the coefficient on the dummy variable.
- If we fail to reject the null that the coefficient is zero, this means that we have no evidence that the two groups differ in their means (or adjusted means) for the dependent variable.
- In the simple regression case, the regression is simply reporting the average of the dependent variable for the two groups, and whether they're statistically different

Example, cont.

```
. ttest schattach, by(male)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |    3234    7.829004    .0363044    2.064566    7.757822    7.900186
       1 |    3340     7.99491    .0340618    1.968524    7.928126    8.061694
---------+--------------------------------------------------------------------
combined |    6574    7.913295    .0248876    2.017894    7.864507    7.962083
---------+--------------------------------------------------------------------
    diff |           -.1659059    .0497434               -.2634192   -.0683925
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -3.3352
Ho: diff = 0                                     degrees of freedom =     6572

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0004         Pr(|T| > |t|) = 0.0009          Pr(T > t) = 0.9996
```

Qualitative variables with 2+ categories

- A qualitative variable with more than two categories can also be analyzed using dummy variables. We have to create more than one dummy variable to do so.
- Let’s say we have three race categories: white, black and other, and one race variable:
  - race=1 if white
  - race=2 if black
  - race=3 if other

Qualitative variables with 2+ categories, cont.

- What happens if we enter this race variable into a regression? Gibberish! Never do this.
- A one unit increase in a qualitative variable is meaningless.
- In order to assess race differences in school attachment, we have to create a dummy variable for each race, and enter any *two* of these into the regression model.
- In general, if there are *j* discrete categories, we need to enter *j-1* dummy variables into the regression model

Qualitative variables with 2+ categories, cont.

- Why *j-1*?
- If we were to include *j* categories, these variables would always sum to 1, and the regression wouldn’t run because of perfect multicollinearity.
- So, how do we create these new variables?

Qualitative variables with 2+ categories, cont.

```
. tab race

       race |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      3,467       52.74       52.74
          2 |      1,897       28.86       81.59
          3 |      1,210       18.41      100.00
------------+-----------------------------------
      Total |      6,574      100.00
```

Technique 1:

```
. gen white=race==1 if race~=.
. gen black=race==2 if race~=.
. gen other=race==3 if race~=.


. summ white black other

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       white |      6574    .5273806    .4992877          0          1
       black |      6574     .288561    .4531278          0          1
       other |      6574    .1840584    .3875613          0          1
```

Qualitative variables with 2+ categories, cont.

Technique 2:

```
. tab race, gen(racecat)

       race |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      3,467       52.74       52.74
          2 |      1,897       28.86       81.59
          3 |      1,210       18.41      100.00
------------+-----------------------------------
      Total |      6,574      100.00

. summ racecat*

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    racecat1 |      6574    .5273806    .4992877          0          1
    racecat2 |      6574     .288561    .4531278          0          1
    racecat3 |      6574    .1840584    .3875613          0          1
```

Qualitative variables with 2+ categories, cont.

Technique 3:

```
. reg schattach i.race
i.race            _Irace_1-3          (naturally coded; _Irace_1 omitted)

      Source |       SS       df       MS              Number of obs =    6574
-------------+------------------------------           F(  2,  6571) =   52.70
       Model |  422.549964     2  211.274982           Prob > F      =  0.0000
    Residual |  26342.0281  6571  4.00883093           R-squared     =  0.0158
-------------+------------------------------           Adj R-squared =  0.0155
       Total |   26764.578  6573  4.07189686           Root MSE      =  2.0022

------------------------------------------------------------------------------
   schattach |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Irace_2 |  -.5825364   .0571798   -10.19   0.000    -.6946274   -.4704454
    _Irace_3 |  -.1250742   .0668533    -1.87   0.061    -.2561284      .00598
       _cons |   8.104413   .0340042   238.34   0.000     8.037754    8.171072
------------------------------------------------------------------------------
```

Qualitative variables with 2+ categories, cont.

- How are the regression results interpreted?
- Using the variables created using technique 1, because they have the most descriptive names, we have the following regression model:
- Schattach = β_1+$\beta_2$black+ $\beta_3$other+ *u*

Qualitative variables with 2+ categories, cont.

- White mean = $\beta_1+\beta_2*0+\beta_3*0=\beta_1$
- Black mean = $\beta_1+\beta_2*1+\beta_3*0=\beta_1+\beta_2$
- ‘Other’ mean = $\beta_1+\beta_2*0+\beta_3*1=\beta_1+\beta_3$
- Each coefficient, β_2 and β_3 tests the difference between the associated category and the omitted one.
  - Here, β_2 is the difference between whites and blacks, β_3 is the difference between whites and ‘others’.
- To test other differences, either run a new regression with a different omitted variable, or:
  - `test black=other`

Qualitative variables with 2+ categories, cont.

```
. reg schattach black other

      Source |       SS       df       MS              Number of obs =    6574
-------------+------------------------------           F(  2,  6571) =   52.70
       Model |  422.549964     2  211.274982           Prob > F      =  0.0000
    Residual |  26342.0281  6571  4.00883093           R-squared     =  0.0158
-------------+------------------------------           Adj R-squared =  0.0155
       Total |   26764.578  6573  4.07189686           Root MSE      =  2.0022

------------------------------------------------------------------------------
   schattach |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |  -.5825364   .0571798   -10.19   0.000    -.6946274   -.4704454
       other |  -.1250742   .0668533    -1.87   0.061    -.2561284      .00598
       _cons |   8.104413   .0340042   238.34   0.000     8.037754    8.171072
------------------------------------------------------------------------------
```

Qualitative variables with 2+ categories, cont.

```
. reg schattach white other

      Source |       SS       df       MS              Number of obs =    6574
-------------+------------------------------           F(  2,  6571) =   52.70
       Model |  422.549964     2  211.274982           Prob > F      =  0.0000
    Residual |  26342.0281  6571  4.00883093           R-squared     =  0.0158
-------------+------------------------------           Adj R-squared =  0.0155
       Total |   26764.578  6573  4.07189686           Root MSE      =  2.0022

------------------------------------------------------------------------------
   schattach |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |   .5825364   .0571798    10.19   0.000     .4704454    .6946274
       other |   .4574622   .0736636     6.21   0.000     .3130575    .6018669
       _cons |   7.521877   .0459701   163.63   0.000      7.43176    7.611993
------------------------------------------------------------------------------
```

Qualitative variables with 2+ categories, cont.

- It is possible for none of the dummy variable coefficients in a set to be statistically significantly different from zero, but for the set to jointly be statistically significant.
  - If the middle category (on levels of DV) is omitted, it may not differ significantly from any other categories, but several included categories may differ from one another
  - To test joint significance in Stata, run the F-test for restricted/unrestricted models:

```
. test black other

 ( 1)  black = 0
 ( 2)  other = 0

       F(  2,  6571) =   52.70
            Prob > F =    0.0000
```

Qualitative variables in multiple regression

- When dummy variables are included in multiple regression, they are interpreted as the expected difference in the outcome variable between groups, *holding all other included variables constant.*
- As more variables are included, the magnitude of the dummy variable coefficients tends to decrease. The raw differences are explained by other differences between the groups.

Qualitative variables in multiple regression, example

```
. reg schattach black other msgpa antipeer

      Source |       SS       df       MS              Number of obs =    6574
-------------+------------------------------           F(  4,  6569) =  265.92
       Model |  3729.91174     4  932.477936           Prob > F      =  0.0000
    Residual |  23034.6663  6569  3.50657121           R-squared     =  0.1394
-------------+------------------------------           Adj R-squared =  0.1388
       Total |   26764.578  6573  4.07189686           Root MSE      =  1.8726

------------------------------------------------------------------------------
   schattach |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |  -.3320044   .0542555    -6.12   0.000    -.4383629    -.225646
       other |  -.0257976   .0626731    -0.41   0.681    -.1486572     .097062
       msgpa |   .3420146    .027845    12.28   0.000     .2874293    .3965999
    antipeer |  -.3588477   .0137435   -26.11   0.000    -.3857895   -.3319059
       _cons |   7.769536   .0933073    83.27   0.000     7.586624    7.952449
------------------------------------------------------------------------------
```

Qualitative variables in multiple regression, cont.

- In the simple regression, the black/white difference in school attachment was .58, but when middle school grades and anti-social peers are controlled, the difference drops to .33
  - You might claim that 43 percent of the black-white gap in school attachment is “explained” by association with antisocial peers and low middle school grades.
  - Of course, low m.s. grades probably results from earlier low school attachment.
- The constant is no longer interpreted as the mean school attachment for whites. It is now the expected school attachment for whites with a 0.00 middle school gpa (not in the data), and a 0 on the antisocial peer scale.

Qualitative variables in multiple regression, cont.

- Testing for joint significance of a set of dummy variables proceeds as before

```
. test black other

 ( 1)  black = 0
 ( 2)  other = 0

       F(  2,  6569) =   19.93
            Prob > F =    0.0000
```

- Notice the F statistic is now much smaller, but still statistically significant.

Multiple sets of dummy variables

- Say you want to look at gender differences and race differences (black, white, other). There are a few different ways to do this:
- First, consider all the possible categories:

	White	Black	Other
Male			
Female			

Multiple sets of dummy variables

- Example 1, assumes that race and gender don't interact (column & row effects, not cells): Schattach = $\beta_1+\beta_2$male+ $\beta_3$black+ $\beta_4$other+ u
- This assumption is twofold
  1. The difference between males and females is the same in each race category.
  2. The difference between races is the same for males and females.
- To calculate the expected school attachment for any group, plug in the appropriate zeros and ones.

Multiple sets of dummy variables, cont

- Example 2, interactive model, different effect for each cell:
- Schattach = $\beta_1+\beta_2$male+ $\beta_3$black+ $\beta_4$other+$\beta_5$male*black+ $\beta_6$male*other+ u
- The two assumptions in Example 1 are dropped.
- Expected school attachment:
  - Black males=$\beta_1+\beta_2+\beta_3+\beta_5$
  - Black females=$\beta_1+\beta_3$
  - White males=$\beta_1+\beta_2$
  - White females= β_1
  - Other males=$\beta_1+\beta_2+\beta_4+\beta_6$
  - Other females=$\beta_1+\beta_4$

Multiple sets of dummy variables, cont

- Example 3 (equivalent to #2 but simpler to interpret, cell effects only):
- Schattach = β_1+$\beta_2$male*black+ $\beta_3$female*black+ $\beta_4$white*female+$\beta_5$male*other+ $\beta_6$female*other+ u
- Expected school attachment:
  - Black males=β_1+ β_2
  - Black females= β_1+ β_3
  - White males=β_1
  - White females= β_1+ β_4
  - Other males= β_1+ β_5
  - Other females= β_1+ β_6

Multiple sets of dummy variables, cont

- The models in examples 2 and 3 will have identical model diagnostics, and either can be compared to the model in example 1 (the restricted model) to jointly test that the interaction terms are equal to zero.
- We'll contrast Example 1 & Example 2.
- We use the F-test for restricted vs. unrestricted models, where the fully interactional model is unrestricted.

Reminder: F-test for restricted/unrestricted models

$$F\left(k_{UR}-k_R, n-k_{UR}\right)=\frac{\left(SSR_R-SSR_{UR}\right)/\left(k_{UR}-k_R\right)}{SSR_{UR}/\left(n-k_{UR}\right)}$$

- Where SSR refers to the residual sum of squares, and k refers to the number of regressors (including the intercept).

Multiple sets of dummy variables, cont

```
. reg schattach male black other

      Source |       SS       df       MS              Number of obs =    6574
-------------+------------------------------           F(  3,  6570) =   38.33
       Model |  460.424139     3  153.474713           Prob > F      =  0.0000
    Residual |  26304.1539  6570  4.00367639           R-squared     =  0.0172
-------------+------------------------------           Adj R-squared =  0.0168
       Total |   26764.578  6573  4.07189686           Root MSE      =  2.0009

------------------------------------------------------------------------------
   schattach |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |    .151884   .0493821     3.08   0.002      .055079    .2486891
       black |  -.5776691   .0571649   -10.11   0.000    -.6897309   -.4656072
       other |  -.1240059   .0668112    -1.86   0.063    -.2549776    .0069658
       _cons |   8.025645   .0425518   188.61   0.000      7.94223    8.109061
------------------------------------------------------------------------------
```

Multiple sets of dummy variables, cont

```
. Xi: reg schattach i.male*i.race
i.male            _Imale_0-1          (naturally coded; _Imale_0 omitted)
i.race            _Irace_1-3          (naturally coded; _Irace_1 omitted)
i.male*i.race     _ImalXrac_#_#       (coded as above)

      Source |       SS       df       MS              Number of obs =    6574
-------------+------------------------------           F(  5,  6568) =   24.68
       Model |  493.631924     5  98.7263847           Prob > F      =  0.0000
    Residual |  26270.9461  6568  3.99983954           R-squared     =  0.0184
-------------+------------------------------           Adj R-squared =  0.0177
       Total |   26764.578  6573  4.07189686           Root MSE      =       2

------------------------------------------------------------------------------
   schattach |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Imale_1 |   .0187919   .0679791     0.28   0.782    -.1144691     .152053
    _Irace_2 |  -.7312178   .0806422    -9.07   0.000    -.8893027   -.5731329
    _Irace_3 |  -.2486438   .0957312    -2.60   0.009    -.4363081   -.0609794
_ImalXrac_~2 |   .3068158    .114286     2.68   0.007     .0827781    .5308535
_ImalXrac_~3 |    .241808   .1336071     1.81   0.070    -.0201053    .5037213
       _cons |   8.094667   .0489546   165.35   0.000     7.998701    8.190634
------------------------------------------------------------------------------
```

Multiple sets of dummy variables, cont

```
. di ((26304.15309-26270.9461)/2)/3.99983954
4.1510403

. di Ftail(2,6568,4.1510403)
.01578936
```

OR:

```
. test  _ImalXrac_1_2 _ImalXrac_1_3

 ( 1)  _ImalXrac_1_2 = 0
 ( 2)  _ImalXrac_1_3 = 0

       F(  2,  6568) =    4.15
            Prob > F =    0.0158
```

Multiple sets of dummy variables, review

```
. reg schattach male black other maleblack maleother antipeer
[cut]
------------------------------------------------------------------------------
   schattach |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   -.128352   .0645282    -1.99   0.047    -.2548482   -.0018557
       black |  -.6122396   .0764103    -8.01   0.000    -.7620287   -.4624506
       other |  -.2271114   .0905682    -2.51   0.012    -.4046545   -.0495682
   maleblack |   .3895791   .1081594     3.60   0.000     .1775516    .6016067
   maleother |   .2983765   .1264131     2.36   0.018     .0505657    .5461874
    antipeer |  -.3834192    .013802   -27.78   0.000    -.4104757   -.3563627
       _cons |   8.870465    .054081   164.02   0.000     8.764449    8.976482
------------------------------------------------------------------------------
```

- What does the constant represent?
- What does the coefficient on male represent?
- What is the difference in school attachment between black males and black females, holding antisocial peers constant?
- What is the difference in school attachment for black males and white males?
- What is the difference in school attachment for black females and white females?

Other points

- It is ok to include an entire set of dummy variables only if they are not mutually exclusive
  - If a ‘1’ is allowed for more than one category, like multiple reasons for dropout, or multiple ethnic identities
  - If a ‘0’ is allowed on all the categories, like types of arrest.
- This changes the interpretation of the coefficient to the difference between that single category and everyone else.

Other points, cont.

- By construction, any set of mutually exclusive dummy variables are highly negatively correlated. This is to be expected, and is not a multicollinearity issue.
- If you have 1 or more tiny groups, consider pooling them. You’ll have little power with such small groups anyway. “Tiny” is relative.

Ordinal variables

- If X is quantitative but discrete, we force some assumptions on its measurement in a regression model
  - The meaning of the distance between any two adjacent values must be constant.
- For example, in some of the previous regression models, we included a supposedly continuous variable called *antipeer*. In fact, *antipeer* takes on only 6 values, 0 through 5, indicating how many antisocial behaviors 50% or more of one's friends are involved in.
- Does moving from a 0 to 1 mean the same thing as moving from a 1 to a 2?

Ordinal variables

- Often, the line between discrete and continuous is fuzzy.
  - Likert scale: 5 different values
  - School expectations in NLSY: 101 different values
- We can test the assumption that an ordinal variable can be modeled in a linear fashion by creating dummy variables for each category.
- When there are too many discrete values, we might create a set of dummy variables, each representing a range of values.

Ordinal variables, example

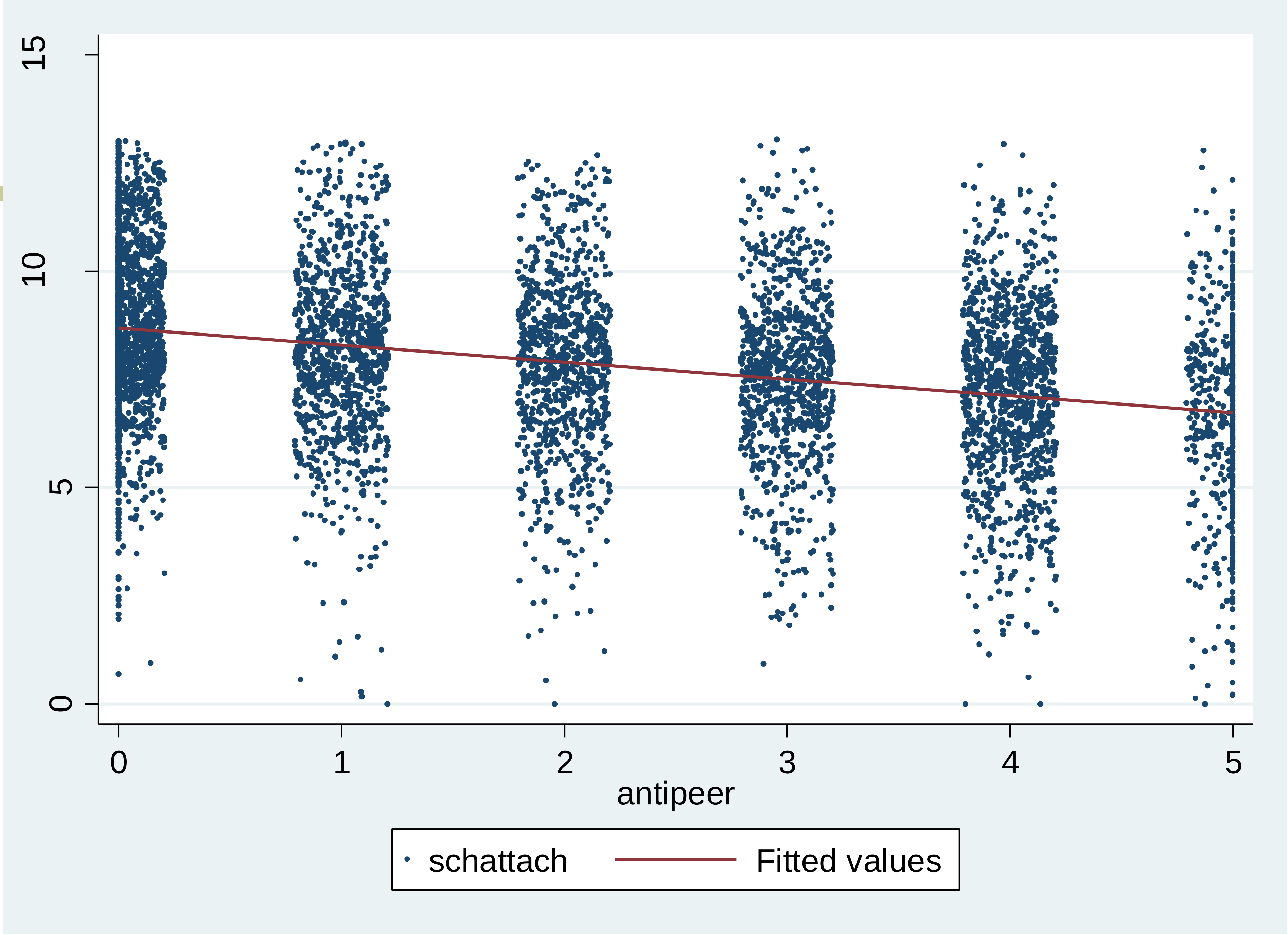
```
. reg schattach antipeer

      Source |       SS       df       MS              Number of obs =    6574
-------------+------------------------------           F(  1,  6572) =  824.32
       Model |  2982.9084     1  2982.9084           Prob > F      =  0.0000
    Residual | 23781.6696  6572  3.61863506           R-squared     =  0.1114
-------------+------------------------------           Adj R-squared =  0.1113
       Total |  26764.578  6573  4.07189686           Root MSE      =  1.9023

------------------------------------------------------------------------------
   schattach |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    antipeer |  -.3944241   .0137378   -28.71   0.000    -.4213545   -.3674936
       _cons |   8.691283   .0358428   242.48   0.000      8.62102    8.761547
------------------------------------------------------------------------------

. predict phat1
(option xb assumed; fitted values)

. twoway (scatter schattach antipeer, jitter(10) msize(tiny)) (line phat1
      antipeer, sort)
```



0
5
10
15
0
1
2
3
4
5
antipeer
schattach
Fitted values

Ordinal variables, example

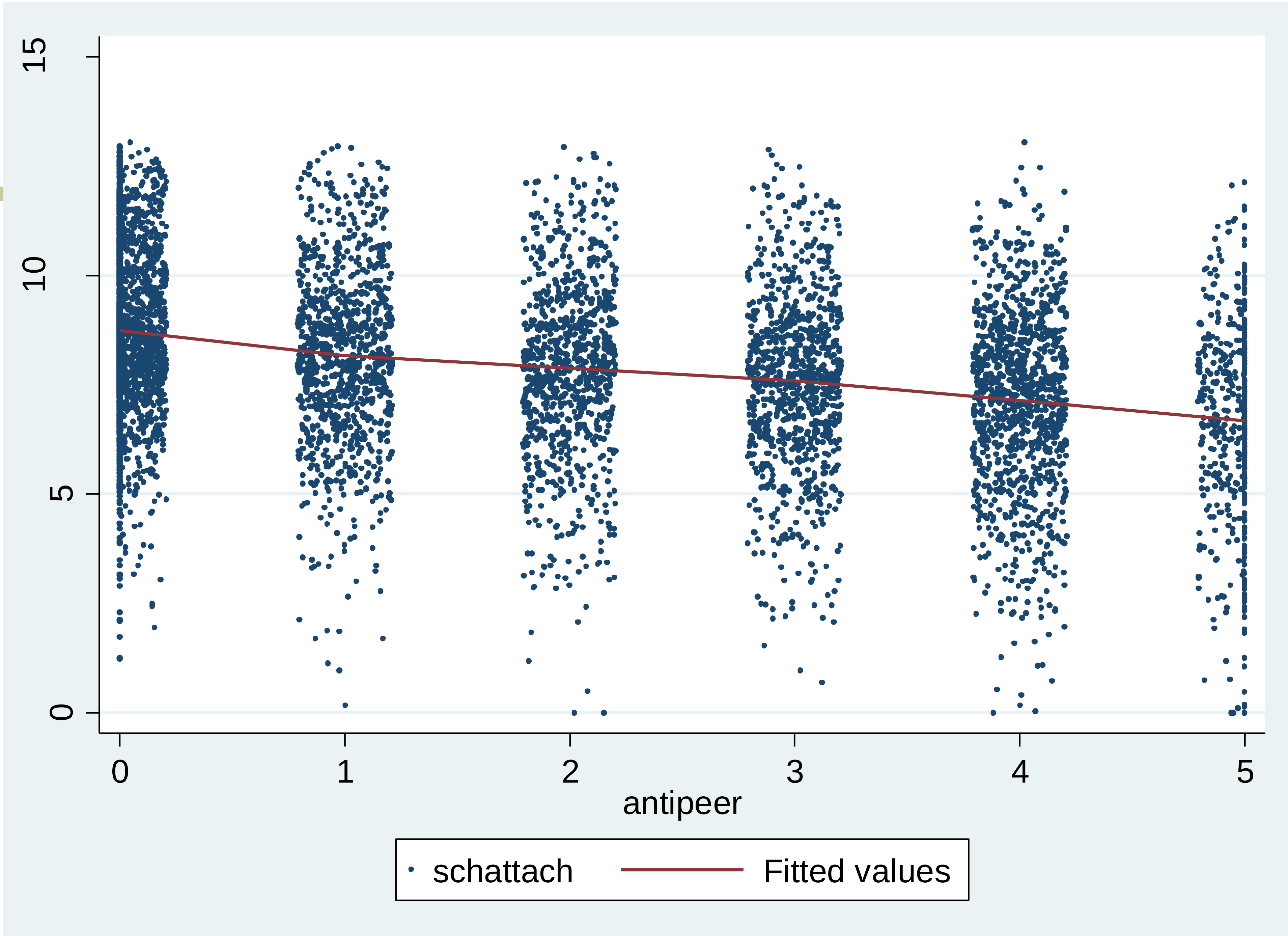
```
. reg schattach i.antipeer
i.antipeer        _Iantipeer_0-5      (naturally coded; _Iantipeer_0 omitted)

      Source |       SS       df       MS              Number of obs =    6574
-------------+------------------------------           F(  5,  6568) =  166.65
       Model |  3013.15977     5  602.631955           Prob > F      =  0.0000
    Residual |  23751.4183  6568  3.61623299           R-squared     =  0.1126
-------------+------------------------------           Adj R-squared =  0.1119
       Total |   26764.578  6573  4.07189686           Root MSE      =  1.9016

------------------------------------------------------------------------------
   schattach |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Iantipeer_1 |  -.5739035   .0727911    -7.88   0.000    -.7165977   -.4312093
_Iantipeer_2 |  -.8590046   .0751619   -11.43   0.000    -1.006346   -.7116628
_Iantipeer_3 |  -1.154187   .0752155   -15.35   0.000    -1.301634    -1.00674
_Iantipeer_4 |  -1.603894    .070838   -22.64   0.000    -1.742759   -1.465028
_Iantipeer_5 |  -2.064729   .0931166   -22.17   0.000    -2.247268    -1.88219
       _cons |   8.737697   .0428336   203.99   0.000     8.653729    8.821664
------------------------------------------------------------------------------

. predict phat2
(option xb assumed; fitted values)

. twoway (scatter schattach antipeer, jitter(10) msize(tiny)) (line phat2 antipeer, sort)
```



15
10
5
0
0
1
2
3
4
5
antipeer
schattach
Fitted values

Ordinal variables, example

- Assuming a constant linear effect, we estimated a change of -.39 in school attachment for each 1 point increase in the antisocial peer scale.
- Relaxing this assumption, we found effects of different magnitudes:
  - Moving from a 0 to a 1 associated with a .57 drop in attachment
  - Moving from a 1 to a 2 associated with a .285 drop in attachment
- Since the first model is nested within the second model, we can test whether allowing unequal changes between categories is more appropriate.

Ordinal variables, example

```
. di (23781.6696-23751.4183)/4
7.562825

. di 7.56285/3.61623299
2.0913614

. di Ftail(4,6568,2.0913614)
.07920167
```

- In this case, we detected some nonlinearity in the scale with respect to school attachment, but we can’t reject the assumption that the effect is linear at a .05 level, although we can at a .10 level.

Dummy variable interactions with continuous variables

- Dummy variables can also be interacted with continuous variables if we believe that the effect of the continuous variable is different for different groups.
- For example, if we feel that the relationship between test scores and school attachment differs by gender, we have to enter an interaction term into the regression model:
  - schattach = $\beta_1+\beta_2$male$+\beta_3$math$+\beta_4$male*math $+u$
- Both the intercept *and* the slope may differ for males and females in this regression.
- The relationship between test scores and school attachment now becomes:
  - For females: $\beta_1+\beta_2*0+\beta_3$math$+\beta_4*0*$math $= \beta_1+\beta_3$math
  - For males: $\beta_1+\beta_2*1+\beta_3$math$+\beta_4*1*$math $=\beta_1+\beta_2+(\beta_3+ \beta_4)$math

Dummy variable interactions with continuous variables

```
. reg schattach male math mathmale

      Source |       SS       df       MS              Number of obs =    6574
-------------+------------------------------           F(  3,  6570) =   66.59
       Model |  789.744404     3  263.248135           Prob > F      =  0.0000
    Residual |  25974.8336  6570  3.95355154           R-squared     =  0.0295
-------------+------------------------------           Adj R-squared =  0.0291
       Total |   26764.578  6573  4.07189686           Root MSE      =  1.9884

------------------------------------------------------------------------------
   schattach |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .1890154    .049529     3.82   0.000     .0919224    .2861084
        math |    .401583   .0389452    10.31   0.000     .3252378    .4779282
    mathmale |  -.0630184   .0539887    -1.17   0.243    -.1688538    .0428171
       _cons |   7.861137   .0351028   223.95   0.000     7.792324     7.92995
------------------------------------------------------------------------------
```

Dummy variable interactions with continuous variables, cont

- The coefficient on the interaction term tests the hypothesis that slope for males and females is the same.
- The male coefficient is the difference between males and females in school attachment when the math score is zero (the mean, in this case).
- The coefficient on math tests the hypothesis that the slope for females on math tests is equal to zero.
  - To do this same test for males, you have to test whether the sum of β_3 and β_4 is equal to zero, or rerun the regression with a female dummy variable.

Dummy variable interactions with continuous variables

```
. reg schattach male math mathmale

      Source |       SS       df       MS              Number of obs =    6574
-------------+------------------------------           F(  3,  6570) =   66.59
       Model |  789.744404     3  263.248135           Prob > F      =  0.0000
    Residual |  25974.8336  6570  3.95355154           R-squared     =  0.0295
-------------+------------------------------           Adj R-squared =  0.0291
       Total |   26764.578  6573  4.07189686           Root MSE      =  1.9884

------------------------------------------------------------------------------
   schattach |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .1890154    .049529     3.82   0.000     .0919224    .2861084
        math |    .401583   .0389452    10.31   0.000     .3252378    .4779282
    mathmale |  -.0630184   .0539887    -1.17   0.243    -.1688538    .0428171
       _cons |   7.861137   .0351028   223.95   0.000     7.792324     7.92995
------------------------------------------------------------------------------
```

Ignoring statistical significance:

- Does the male/female gap in school attachment increase or decrease as math scores increase?
- Is the effect of math score on school attachment greater for males or females?

Dummy variable interactions with continuous variables: Four general cases

1. No dummies, no interactions: one slope and intercept for all
2. Dummies: same slope for all, different levels (intercepts)
3. Dummies and interactions: different slopes and intercepts for each group (most general)
4. Interactions only: different slopes, same intercept (not normally used)

Another example + graphing interactions, a simplified conservatism model

```
. reg cons childs educ age tvhours inc06 male i.black##c.rel househ
```

Source	SS	df	MS
Model	186.320369	10	18.6320369
Residual	886.679628	1063	.834129471
Total	1073	1073	.999999997

Number of obs =	1074
F(10, 1063) =	22.34
Prob > F =	0.0000
R-squared =	0.1736
Adj R-squared =	0.1659
Root MSE =	.91331

conserv	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
childs	.0210998	.0207242	1.02	0.309	-.0195652	.0617649
educ	-.0156987	.0108031	-1.45	0.146	-.0368965	.0054991
age	-.0016935	.0022119	-0.77	0.444	-.0060337	.0026466
tvhours	.0122988	.0129304	0.95	0.342	-.0130733	.0376709
inc06	.0285528	.0064039	4.46	0.000	.0159872	.0411185
male	.1484773	.0583948	2.54	0.011	.0338951	.2630596
1.black	-.6813751	.0918733	-7.42	0.000	-.8616487	-.5011014
rel	.3138422	.0309431	10.14	0.000	.2531257	.3745587
black#c.rel						
1	-.3172576	.0927085	-3.42	0.001	-.4991701	-.1353451
househ	-.0226968	.0242022	-0.94	0.349	-.0701863	.0247927
_cons	.1229882	.2040162	0.60	0.547	-.277332	.5233085

What is the effect of religiosity?

Another example + graphing interactions, a simplified conservatism model

- Is there a statistically significant relationship between religiosity and conservatism for blacks?
- To test this, we ask Stata to test whether the sum of the religion effect and interaction term is equal to zero. This is the religion effect for blacks.
- But how do we refer to that weird interaction term in the previous regression? Using the “coeflegend” option will tell you.

```
. test _b[rel]+_b[1.black#c.rel]=0

 ( 1)  rel + 1.black#c.rel = 0

       F(  1,  1063) =    0.00
            Prob > F =    0.9692
```

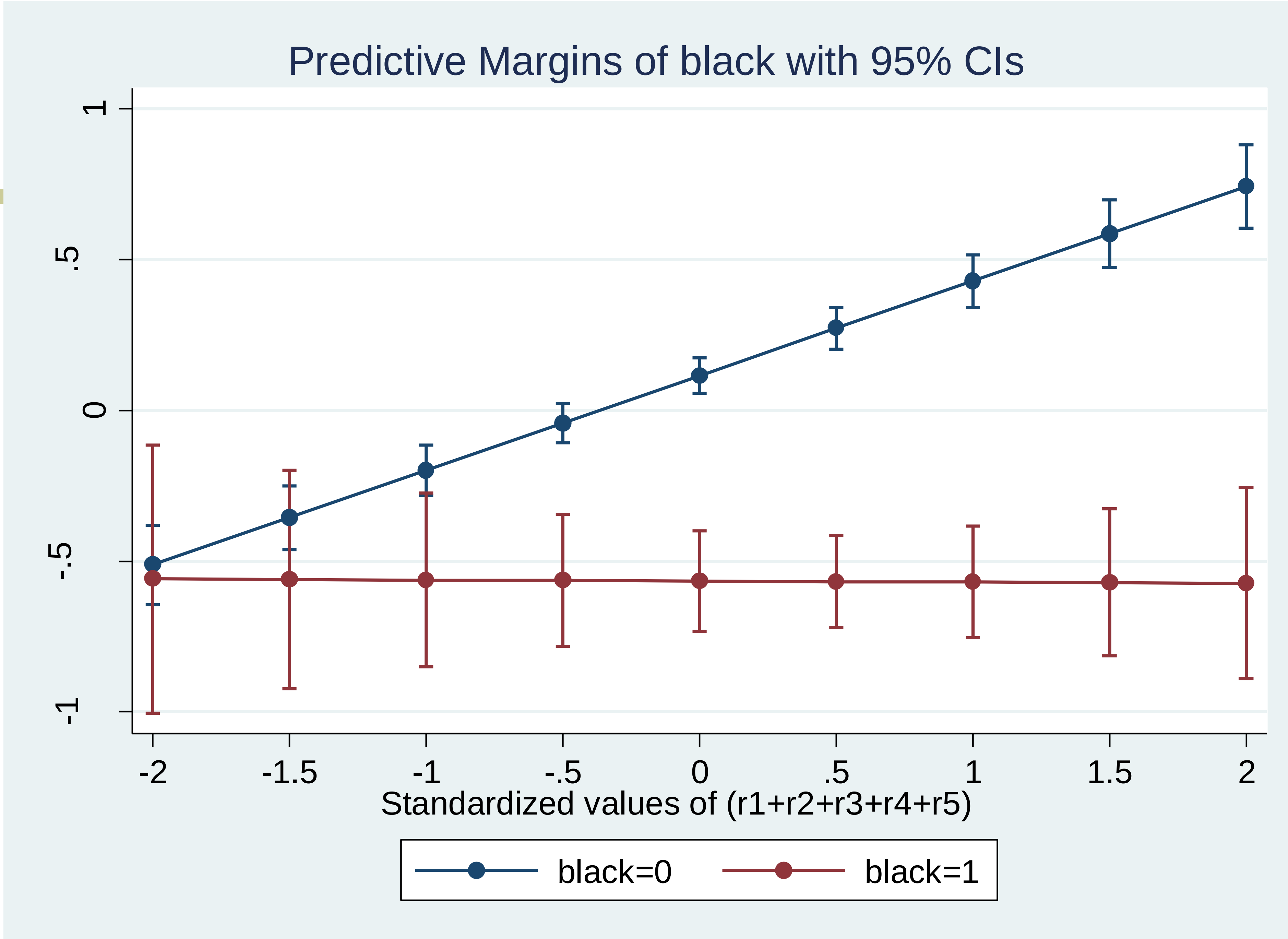
- This shows us that we cannot reject the null hypothesis that there is no relationship between religiosity and conservatism among blacks.
- Let’s look at this relationship visually. First, we need to use the correct margins command.

Another example + graphing interactions, a simplified conservatism model

```
. margins black, at(rel=(-2(.5)2))
```

	Margin	Delta-method Std. Err.	t	P>|t|	[95% Conf.	Interval]
_at#black						
1 0	-.5125179	.0672325	-7.62	0.000	-.6444415	-.3805944
1 1	-.5593778	.2262175	-2.47	0.014	-1.003261	-.1154942
2 0	-.3555968	.0538881	-6.60	0.000	-.4613359	-.2498578
2 1	-.5610855	.1852863	-3.03	0.003	-.9246538	-.1975171
3 0	-.1986757	.0420282	-4.73	0.000	-.2811434	-.116208
3 1	-.5627932	.1462946	-3.85	0.000	-.8498521	-.2757343
4 0	-.0417546	.03328	-1.25	0.210	-.1070565	.0235473
4 1	-.5645009	.1112999	-5.07	0.000	-.7828933	-.3461085
5 0	.1151665	.0304545	3.78	0.000	.0554086	.1749243
5 1	-.5662086	.0853679	-6.63	0.000	-.7337174	-.3986998
6 0	.2720876	.0350163	7.77	0.000	.2033787	.3407965
6 1	-.5679163	.0781164	-7.27	0.000	-.7211961	-.4146365
7 0	.4290087	.0447609	9.58	0.000	.3411789	.5168384
7 1	-.569624	.093974	-6.06	0.000	-.7540197	-.3852283
8 0	.5859298	.0570936	10.26	0.000	.4739009	.6979587
8 1	-.5713317	.1243967	-4.59	0.000	-.8154226	-.3272408
9 0	.7428509	.0706721	10.51	0.000	.6041781	.8815236
9 1	-.5730394	.1613456	-3.55	0.000	-.8896315	-.2564474

Now, just type “marginsplot” to see the magic.


Predictive Margins of black with 95% CIs
1
.5
0
-.5
-1
-2
-1.5
-1
-.5
0
.5
1
1.5
2
Standardized values of (r1+r2+r3+r4+r5)
black=0
black=1

Chow test revisited

- If we want to test whether our full model is the same across different groups, we run a Chow test.
- Let’s run a Chow test with three subgroups: white, black & other

Chow test revisited

Unrestricted model (three groups):

$$Y = \beta_{0w} + \beta_{1w} X_1 + \beta_{2w} X_2 + \ldots + \beta_{kw} X_k + u$$

$$Y = \beta_{0b} + \beta_{1b} X_1 + \beta_{2b} X_2 + \ldots + \beta_{kb} X_k + u$$

$$Y = \beta_{0o} + \beta_{1o} X_1 + \beta_{2o} X_2 + \ldots + \beta_{ko} X_k + u$$

Restricted model (pooled):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + u$$

$$\beta_{0w} = \beta_{0b} = \beta_{0o} = \beta_0, \beta_{1w} = \beta_{1b} = \beta_{1o} = \beta_1, etc$$ ←restrictions

Chow test, restricted model

```
. reg schattach male antipeer math

      Source |       SS       df       MS              Number of obs =    6574
-------------+------------------------------           F(  3,  6570) =  311.14
       Model |  3329.53159     3  1109.84386           Prob > F      =  0.0000
    Residual |  23435.0464  6570  3.56697815           R-squared     =  0.1244
-------------+------------------------------           Adj R-squared =  0.1240
       Total |   26764.578  6573  4.07189686           Root MSE      =  1.8886

------------------------------------------------------------------------------
   schattach |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .0742307   .0468661     1.58   0.113    -.0176422    .1661035
    antipeer |  -.3708353   .0138827   -26.71   0.000    -.3980498   -.3436208
        math |   .2547431   .0259727     9.81   0.000      .203828    .3056581
       _cons |   8.638187   .0442618   195.16   0.000     8.551419    8.724954
------------------------------------------------------------------------------
```

Chow test, unrestricted model (part 1)

```
. reg schattach male antipeer math if white==1

      Source |       SS       df       MS              Number of obs =    3467
-------------+------------------------------           F(  3,  3463) =  180.23
       Model |  1795.34448     3  598.448159           Prob > F      =  0.0000
    Residual |   11498.858  3463  3.32049033           R-squared     =  0.1350
-------------+------------------------------           Adj R-squared =  0.1343
       Total |  13294.2025  3466  3.83560372           Root MSE      =  1.8222

------------------------------------------------------------------------------
   schattach |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |  -.1111335   .0625294    -1.78   0.076    -.2337317    .0114647
    antipeer |  -.3988284   .0188781   -21.13   0.000    -.4358416   -.3618151
        math |   .2296658    .036011     6.38   0.000     .1590608    .3002707
       _cons |   8.867848   .0594835   149.08   0.000     8.751222    8.984474
------------------------------------------------------------------------------
```

Chow test, unrestricted model (part 2)

```
. reg schattach male antipeer math if black==1

      Source |       SS       df       MS              Number of obs =    1897
-------------+------------------------------           F(  3,  1893) =   62.65
       Model |  768.503536     3  256.167845           Prob > F      =  0.0000
    Residual |  7740.83858  1893  4.08919101           R-squared     =  0.0903
-------------+------------------------------           Adj R-squared =  0.0889
       Total |  8509.34212  1896  4.48804964           Root MSE      =  2.0222

------------------------------------------------------------------------------
   schattach |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .2880907   .0934448     3.08   0.002     .1048251    .4713563
    antipeer |  -.3490126   .0274412   -12.72   0.000    -.4028307   -.2951944
        math |   .1306803   .0533725     2.45   0.014     .0260051    .2353554
       _cons |   8.229186   .0920434    89.41   0.000     8.048669    8.409703
------------------------------------------------------------------------------
```

Chow test, unrestricted model (part 3)

```
. reg schattach male antipeer math if other==1

      Source |       SS       df       MS              Number of obs =    1210
-------------+------------------------------           F(  3,  1206) =   55.61
       Model |  551.504621     3  183.834874           Prob > F      =  0.0000
    Residual |  3986.97885  1206  3.30595261           R-squared     =  0.1215
-------------+------------------------------           Adj R-squared =  0.1193
       Total |  4538.48347  1209   3.7539152           Root MSE      =  1.8182

------------------------------------------------------------------------------
   schattach |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .1819946   .1048028     1.74   0.083    -.0236215    .3876107
    antipeer |  -.3088765   .0301016   -10.26   0.000    -.3679339   -.2498191
        math |   .3282606   .0598712     5.48   0.000     .2107974    .4457239
       _cons |   8.556299   .0968935    88.31   0.000       8.3662    8.746397
------------------------------------------------------------------------------
```

F-test for restricted/unrestricted models, Chow test example

- Chow test proceeds as follows:

$$F\left(k_{UR}-k_R, n-k_{UR}\right)=\frac{\left(SSR_R-SSR_{UR}\right)/\left(k_{UR}-k_R\right)}{SSR_{UR}/\left(n-k_{UR}\right)}$$

$$F\left(12-4, 6574-12\right)=\frac{\left(23435-\left(11499+7741+3987\right)\right)/\left(12-4\right)}{\left(11499+7741+3987\right)/\left(6574-12\right)}$$

$$F(8, 6562)=\frac{208/8}{23227/6562}=7.35, \; (p<.001)$$

- Reject the null. The model differs by race.

Linear Probability Model

- Although not very common in criminology, it is possible to run multiple regression with a dummy variable as the dependent variable.
- The key to understanding what this type of regression means:
  - the expected value of Y conditional on X is the same as the probability that Y=1 conditional on X.
- So a 1 unit increase in an independent variable is associated with a β increase in the probability that Y=1.

Linear Probability Model example (Loeffler, 2013)

Table 2. Estimated Effects of Imprisonment on Recidivism within 5 Years

Variables	(1) Unconditional		(2) OLS		(3) 2SLS		(4) 2SLS	
	b	(SE)	*b*	(SE)	*b*	(SE)	*b*	(SE)
Prison	.088***	(.007)	.031***	(.008)	.075	(.104)	−.035	(.113)
Female			−.007	(.009)	−.002	(.015)	−.027	(.034)
White			−.079***	(.011)	−.079***	(.011)	−.128***	(.037)
Hispanic			−.112***	(.010)	−.112***	(.011)	−.091**	(.034)
Age			−.008***	(.002)	−.009**	(.003)	−.003	(.007)
Age squared			.000*	(.000)	.000*	(.000)	.000	(.000)
Prior_conv			.037***	(.003)	.031*	(.015)	.036	(.019)
Age at first			−.004***	(.001)	−.004***	(.001)	−.004*	(.002)
Prior arrests			.006***	(.000)	.006***	(.000)	.008***	(.001)
Chg. class 2			−.058***	(.007)	−.061***	(.010)	−.056*	(.024)
Chg. class 3			−.135***	(.009)	−.137***	(.010)	−.080**	(.030)
Instruments	None		None		All judges		High/low judges	
N	20,297		20,297		20,297		1,879	

- Dependent variable is felony re-arrest (0/1)
- Model 1 shows that those who were previously imprisoned were subsequently re-arrested at a higher rate (8.8 percentage points higher)
- Controlling for other characteristics reduces this to 3.1 percentage points – fancy stuff follows in models 3 and 4

Linear Probability Model example (Brezina et al, 2009 in *Criminology*)

Table 2. OLS Estimates of the Effects of AED on Offending Behaviors—Wave 1 and Wave 2

	Burglary	Graffiti	Assault	Property damage	Theft	Robbery	Pulled knife or gun	Shot or stabbed
Probability of being killed by 21 < 50%	−.023***	−.031***	−.052***	−.042***	−.028***	−.030***	−.035***	−.025***
	(.004)	(.005)	(.006)	(.006)	(.004)	(.005)	(.004)	(.003)
Probability of being killed by 21 > 50%	.044***	.045***	.041**	.033*	.056***	.034**	.073***	.051***
	(.014)	(.016)	(.018)	(.018)	(.014)	(.015)	(.015)	(.013)
Probability of living up to 35 < 50%	.053***	.034***	.057***	.055***	.041***	.053***	.057***	.047***
	(.009)	(.010)	(.012)	(.012)	(.009)	(.010)	(.009)	(.008)
Probability of living up to 35 > 50%	−.022***	−.032***	−.030***	−.048***	−.023***	−.024***	−.019***	−.016***
	(.003)	(.003)	(.004)	(.004)	(.002)	(.003)	(.002)	(.002)
County fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	34,479	34,460	34,462	34,461	34,482	34,483	34,504	34,497

NOTES: Heteroskedasticity corrected robust standard errors are in parentheses.
*$p < .10$; **$p < .05$; ***$p < .01$.

Next time:

Homework 8 Problems 7.1, C7.4, C7.6, C7.8

Read: Wooldridge Chapter 8