Sociology 7704: Regression Models for Categorical Data
Instructor: Natasha Sarkisian

OLS Regression Assumptions

A1. All independent variables are quantitative or dichotomous, and the dependent variable is quantitative, continuous, and unbounded. All variables are measured without error.
A2. All independent variables have some variation in value (non-zero variance).
A3. There is no exact linear relationship between two or more independent variables (no perfect multicollinearity).
A4. At each set of values of the independent variables, the mean of the error term is zero.
A5. Each independent variable is uncorrelated with the error term.
A6. At each set of values of the independent variables, the variance of the error term is the same (homoscedasticity).
A7. For any two observations, their error terms are not correlated (lack of autocorrelation).
A8. At each set of values of the independent variables, error term is normally distributed.
A9. The change in the expected value of the dependent variable associated with a unit increase in an independent variable is the same regardless of the specific values of other independent variables (additivity assumption).
A10. The change in the expected value of the dependent variable associated with a unit increase in an independent variable is the same regardless of the specific values of this independent variable (linearity assumption).

A1-A7: Gauss-Markov assumptions: If these assumptions hold, the resulting regression estimates are BLUE (Best Linear Unbiased Estimates).

Unbiased: if we were to calculate that estimate over many samples, the mean of these estimates would be equal to the mean of the population (i.e., on average we are on target).

Best (also known as efficient): the standard deviation of the estimate is the smallest possible (i.e., not only are we on target on average, but we don't deviate too far from it).

If A8-A10 also hold, the results can be used appropriately for statistical inference (i.e., significance tests, confidence intervals).

OLS Regression diagnostics and remedies

1. Multivariate Normality

OLS is not very sensitive to non-normally distributed errors but the efficiency of estimators decreases as the distribution substantially deviates from normal (especially if there are heavy tails). Further, heavily skewed distributions are problematic as they question the validity of the mean as a measure for central tendency and OLS relies on means. Therefore, we usually test for nonnormality of residuals' distribution and if it's found, attempt to use transformations to remedy the problem.

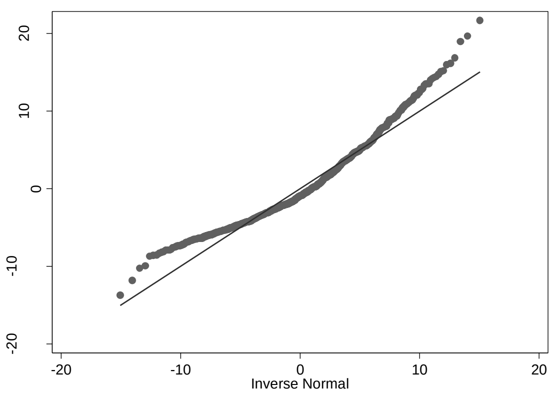
To test normality of error terms distribution, first, we generate a variable containing residuals:

```
. reg agekdbrn educ born sex mapres80 age
```

```
      Source |       SS       df       MS              Number of obs =    1089
-------------+------------------------------           F(  5,  1083) =   49.10
       Model |  5760.17098     5  1152.0342            Prob > F      =  0.0000
    Residual |   25412.492  1083  23.4649049           R-squared     =  0.1848
-------------+------------------------------           Adj R-squared =  0.1810
       Total |   31172.663  1088  28.6513447           Root MSE      =  4.8441

------------------------------------------------------------------------------
    agekdbrn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .6158833   .0561099    10.98   0.000     .5057869    .7259797
        born |   1.679078   .5757599     2.92   0.004     .5493468    2.808809
         sex |  -2.217823   .3043625    -7.29   0.000     -2.81503   -1.620616
    mapres80 |   .0331945   .0118728     2.80   0.005     .0098982    .0564909
         age |   .0582643   .0099202     5.87   0.000     .0387993    .0777293
       _cons |   13.27142   1.252294    10.60   0.000     10.81422    15.72861
------------------------------------------------------------------------------

. predict resid1, resid
(1676 missing values generated)
```

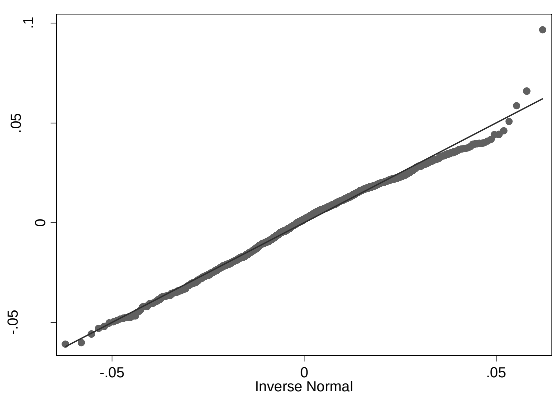
Next, we can use any of the tools we used above to evaluate the normality of distribution for this variable. For example, we can construct the qnorm plot:

```
. qnorm resid1
```



In this case, residuals deviate from normal quite substantially. We could check whether transforming the dependent variable using the transformation we identified above would help us:

```
. quietly reg agekdbrnrr educ born sex mapres80 age
. predict resid2, resid
(1676 missing values generated)
. qnorm resid2
```



Looks much better – the residuals are essentially normally distributed although it looks like there are a few outliers in the tails. We could further examine the outliers and influential observations; we'll discuss that later.

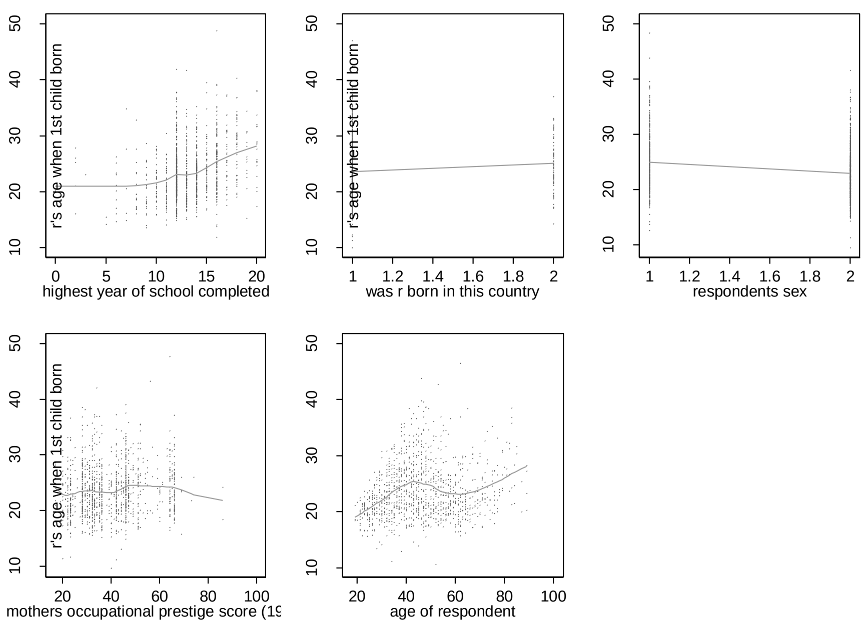
2. Linearity

We looked at bivariate plots to assess linearity during the screening phase, but bivariate plots do not tell the whole story - we are interested in partial relationships, controlling for all other regressors. We can try plots for such relationship using mrunning command. Let's download that first:

```
. search mrunning

Keyword search
        Keywords:  mrunning
          Search:  (1) Official help files, FAQs, Examples, SJs, and STBs
Search of official help files, FAQs, Examples, SJs, and STBs
SJ-5-3  gr0017  . . . . . . . . . . . . . . A multivariable scatterplot smoother
        (help mrunning, running if installed) . . . . P. Royston and N. J. Cox
        Q3/05   SJ 5(3):405--412
        presents an extension to running for use in a
        multivariable context
```

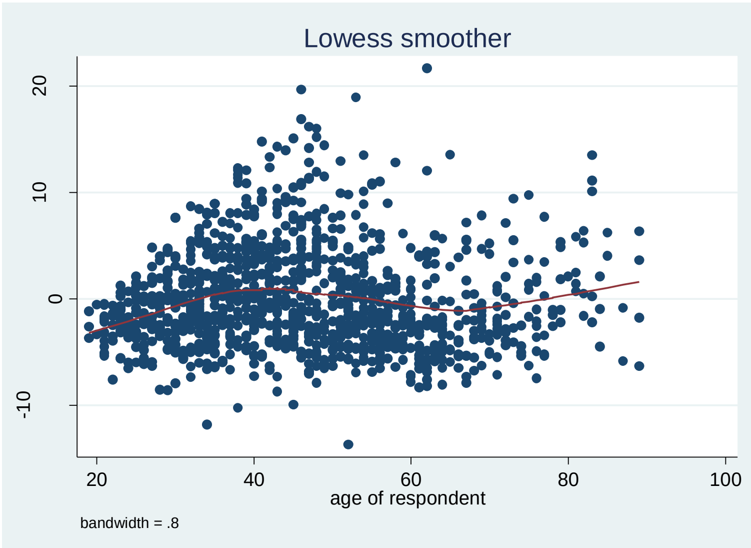
Click on gr0017 to install the program. Now we can use it:

```
. mrunning agekdbrn educ born sex mapres80 age
1089 observations, R-sq = 0.2768
```



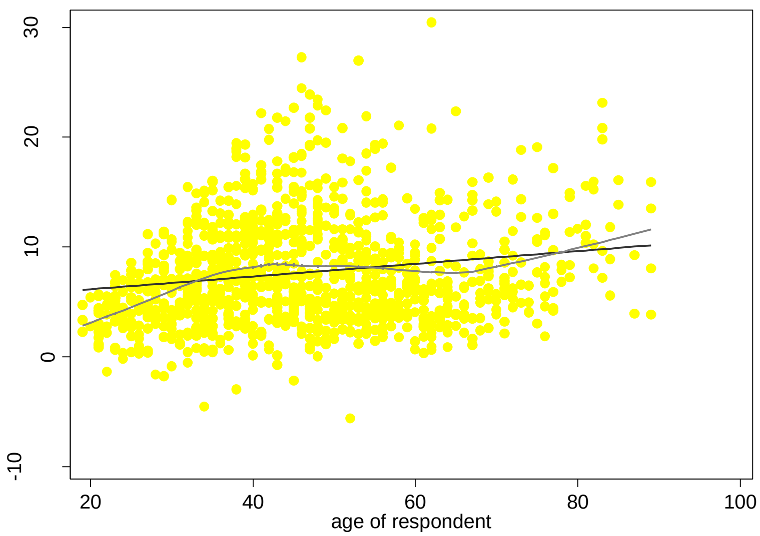
We can clearly see some substantial nonlinearity for educ and age; mapres80 doesn't look quite linear either. We can also run our regression model and examine the residuals. One way to do so would be to plot residuals against each continuous independent variable:

```
.lowess resid1 age
```



We can detect some nonlinearity in this graph. A more effective tool for detecting nonlinearity in such multivariate context is so-called augmented component plus residual plots, usually with lowess curve:

```
. acprplot age, lowess mcolor(yellow)
```



In addition to these graphical tools, there are also a few tests we can run. One way to diagnose nonlinearities is so-called omitted variables test. It searches for a pattern in residuals that could suggest that a power transformation of one of the variables in the model is omitted. To find such factors, it uses either the powers of the fitted values (which means, in essence, powers of the linear combination of all regressors) or the powers of individual regressors in predicting Y. If it finds a significant relationship, this suggests that we probably overlooked some nonlinear relationship.

```
. ovtest
Ramsey RESET test using powers of the fitted values of agekdbrn
       Ho:  model has no omitted variables
                F(3, 1080) =      2.74
                  Prob > F =      0.0423
```

```
. ovtest, rhs
(note:  born dropped due to collinearity)
(note:  sex dropped due to collinearity)
(note:  born^3 dropped due to collinearity)
(note:  born^4 dropped due to collinearity)
(note:  sex^3 dropped due to collinearity)
(note:  sex^4 dropped due to collinearity)

Ramsey RESET test using powers of the independent variables
       Ho:  model has no omitted variables
               F(11, 1074) =     14.84
                  Prob > F =      0.0000
```

Looks like we might be missing some nonlinear relationships. We will, however, also explicitly check for linearity for each independent variable. We can do so using Box-Tidwell test. First, we need to download it:

```
. net search boxtid
(contacting http://www.stata.com)

3 packages found (Stata Journal and STB listed first)
-----------------------------------------------------

sg112_1 from http://www.stata.com/stb/stb50
    STB-50 sg112_1.  Nonlin. reg. models with power or exp. func. of covar. /
    STB insert by / Patrick Royston, Imperial College School of Medicine, UK;
    / Gareth Ambler, Imperial College School of Medicine, UK. / Support:
    proyston@rpms.ac.uk and gambler@rpms.ac.uk / After installation, see
```

We select this first one, sg112_1, and install it. Now use it:

```
. boxtid reg agekdbrn educ born sex mapres80 age
Iteration 0:  Deviance =  6483.522
Iteration 1:  Deviance =  6470.107 (change = -13.41466)
Iteration 2:  Deviance =   6469.55 (change = -.5577601)
Iteration 3:  Deviance =  6468.783 (change = -.7663782)
Iteration 4:  Deviance =    6468.6 (change = -.1832873)
Iteration 5:  Deviance =  6468.496 (change =  -.103788)
Iteration 6:  Deviance =  6468.456 (change = -.0399491)
Iteration 7:  Deviance =  6468.438 (change = -.0177698)
Iteration 8:  Deviance =   6468.43 (change = -.0082658)
Iteration 9:  Deviance =  6468.427 (change = -.0035944)
Iteration 10:  Deviance =  6468.425 (change = -.0018104)
Iteration 11:  Deviance =  6468.424 (change = -.0008303)
-> gen double Ieduc__1 = X^2.6408-2.579607814 if e(sample)
-> gen double Ieduc__2 = X^2.6408*ln(X)-.9256893949 if e(sample)
   (where: X = (educ+1)/10)
-> gen double Imapr__1 = X^0.4799-1.931881531 if e(sample)
-> gen double Imapr__2 = X^0.4799*ln(X)-2.650956804 if e(sample)
   (where: X = mapres80/10)
-> gen double Iage__1 = X^-3.2902-.0065387933 if e(sample)
-> gen double Iage__2 = X^-3.2902*ln(X)-.009996425 if e(sample)
   (where: X = age/10)
-> gen double Iborn__1 = born-1 if e(sample)
-> gen double Isex__1 = sex-1 if e(sample)
[Total iterations: 33]
Box-Tidwell regression model
      Source |       SS       df       MS           Number of obs =    1089
-------------+------------------------------           F(  8,  1080) =   38.76
       Model |  6953.00253     8  869.125317           Prob > F      =  0.0000
    Residual |  24219.6605  1080  22.4256115           R-squared     =  0.2230
-------------+------------------------------           Adj R-squared =  0.2173
```

```
       Total |   31172.663  1088  28.6513447           Root MSE      =  4.7356
------------------------------------------------------------------------------
    agekdbrn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    Ieduc__1 |   1.215639   .7083273     1.72   0.086     -.174215    2.605492
    Ieduc_p1 |     .00374   .8606987     0.00   0.997    -1.685091    1.692571
    Imapr__1 |   1.153845    9.01628     0.13   0.898    -16.53757    18.84525
    Imapr_p1 |   .0927861   2.600166     0.04   0.972    -5.009163    5.194736
     Iage__1 |  -67.26803   42.28364    -1.59   0.112    -150.2354    15.69937
     Iage_p1 |  -.4932163   53.49507    -0.01   0.993    -105.4593    104.4728
    Iborn__1 |   1.380925   .5659349     2.44   0.015     .2704681    2.491381
     Isex__1 |  -2.017794    .298963    -6.75   0.000    -2.604408    -1.43118
       _cons |   25.14711   .2955639    85.08   0.000     24.56717    25.72706
------------------------------------------------------------------------------
educ      |   .5613397     .05549      10.116   Nonlin. dev. 11.972  (P = 0.001)
      p1 |    2.64077   .7027411      3.758
------------------------------------------------------------------------------
mapres80 |   .0337813   .0115436      2.926   Nonlin. dev. 0.126   (P = 0.724)
      p1 |   .4798773    1.28955      0.372
------------------------------------------------------------------------------
age      |   .0534185   .0098828      5.405   Nonlin. dev. 39.646  (P = 0.000)
      p1 |  -3.290191   .8046904     -4.089
------------------------------------------------------------------------------
Deviance: 6468.424.
```

Here, we interpret the last three portions of output, and more specifically the P values there. P=0.001 for educ and P=0.000 for age suggests that there is some nonlinearity with regard to these two variables. Mapres80 appears to be fine. With regard to remedies, the process here is the same as we discussed earlier when talking about bivariate linearity. Once remedies are applied, it is a good idea to retest using these multivariate screening tools.

3. Outliers, Leverage Points, and Influential Observations

A single observation that is substantially different from other observations can make a large difference in the results of regression analysis. For this reason, unusual observations (or small groups of unusual observations) should be identified and examined. There are three ways that an observation can be unusual:

Outliers: In univariate context, people often refer to observations with extreme values (unusually high or low) as outliers. But in regression models, an outlier is an observation that has unusual value of the dependent variable given its values of the independent variables – that is, the relationship between the dependent variable and the independent ones is different for an outlier than for the other data points. Graphically, an outlier is far from the pattern defined by other data points. Typically, in a regression model, an outlier has a large residual.

Leverage points: An observation with an extreme value (either very high or very low) on a single predictor variable or on a combination of predictors is called a point with high leverage. Leverage is a measure of how far a value of an independent variable deviates from the mean of that variable. In the multivariate context, leverage is a measure of each observation's distance from the multidimensional centroid in the space formed by all the predictors. These leverage points can have an effect on the estimates of regression coefficients.

Influential Observations: A combination of the previous two characteristics produces influential observations. An observation is considered influential if removing the observation substantially changes the estimates of coefficients. Observations that have just one of these two characteristics (either an outlier or a high leverage point but not both) do not tend to be influential.

Thus, we want to identify outliers and leverage points, and especially those observations that are both, to assess and possibly minimize their impact on our regression model. Furthermore, outliers, even when they are not influential in terms of coefficient estimates, can unduly inflate the error variance. Their presence may also signal that our model failed to capture some important factors (i.e., indicate potential model specification problem).

In the multivariate context, to identify outliers, we want to find observations with high residuals; and to identify observations with high leverage, we can use the so-called hat-values -- these measure each observation's distance from the multidimensional centroid in the space formed by all the regressors. We can also use various influence statistics that help us identify influential observations by combining information on outlierness and leverage.

To obtain these various statistics in Stata, we use predict command. Here are some values we can obtain using predict, with the rule-of-thumb cutoff values for statistics used in outlier diagnostics:

Predict option	**Result**	**Cutoff value (n=sample size, k=parameters)**
xb	xb, fitted values (linear prediction); the default	
stdp	standard error of linear prediction	
residuals	residuals	
stdr	standard error of the residual	
rstandard	standardized residuals (residuals divided by standard error)	
rstudent	studentized (jackknifed) residuals, recommended for outlier diagnostics (for each observation, the residual is divided by the standard error obtained from a model that includes a dummy variable for that specific observation)	|rstudent|> 2
lev (hat)	hat values, measures of leverage (diagonal elements of hat matrix)	Hat >(2k+2)/n
*dfits	DFITS, influence statistic based on studentized residuals and hat values	|DFits|>2*sqrt(k/n)
*welsch	Welsch Distance, a variation on dfits	|WelschD|>3*sqrt(k)
cooksd	Cook's distance, an influence statistic based on dfits and indicating the distance between coefficient vectors when the jth observation is omitted	CooksD >4/n
*covratio	COVRATIO, a measure of the influence of the jth observation on the variance-covariance matrix of the estimates	|CovRatio-1|>3k/n
*dfbeta(varname)	DFBETA, a measure of the influence of the jth observation on each coefficient (the difference between the regression coefficient when the jth observation is included and when it is excluded, divided by the estimated standard error of the coefficient)	|DFBeta|> 2/sqrt(n)

*Note: Starred statistics are only available for the estimation sample; unstarred statistics are available both in and out of sample; type predict ... if e(sample) ... if you want them only for the estimation sample.

So we could obtain and individually examine various outlier and leverage statistics, e.g.,

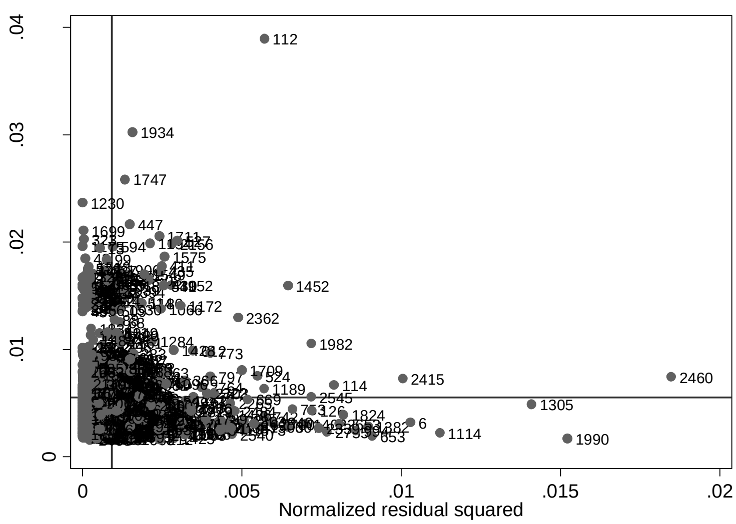
```
.predict hats, lev
.predict resid, resid
.predict rstudent, rstudent
```

For instance, we can then find the observations with the highest leverage values:

```
. sum hats if e(sample), det
                          Leverage
-------------------------------------------------------------
      Percentiles      Smallest
 1%       .00176       .0015777
 5%     .0021025       .0016196
10%     .0023401         .00162       Obs                1089
25%     .0030041       .0016511       Sum of Wgt.        1089

50%     .0041908                      Mean           .0055096
                        Largest       Std. Dev.       .004043
75%      .006332       .0236406
90%      .010143       .0258473       Variance       .0000163
95%     .0155289       .0302377       Skewness       2.466179
99%     .0198167        .038942       Kurtosis       11.40481

. list id hats if hats>.023 & hats~=. & e(sample)
      +-----------------+
      |   id       hats |
      |-----------------|
   3. | 1934   .0302377 |
  10. |  112    .038942 |
  17. | 1230   .0236406 |
2447. | 1747   .0258473 |
      +-----------------+
```

But the best way to graphically examine both leverage values and residuals at the same time is the leverage versus the residuals squared plot (L-R plot) (you can replicate it by creating a scatterplot of hat values and residuals squared):

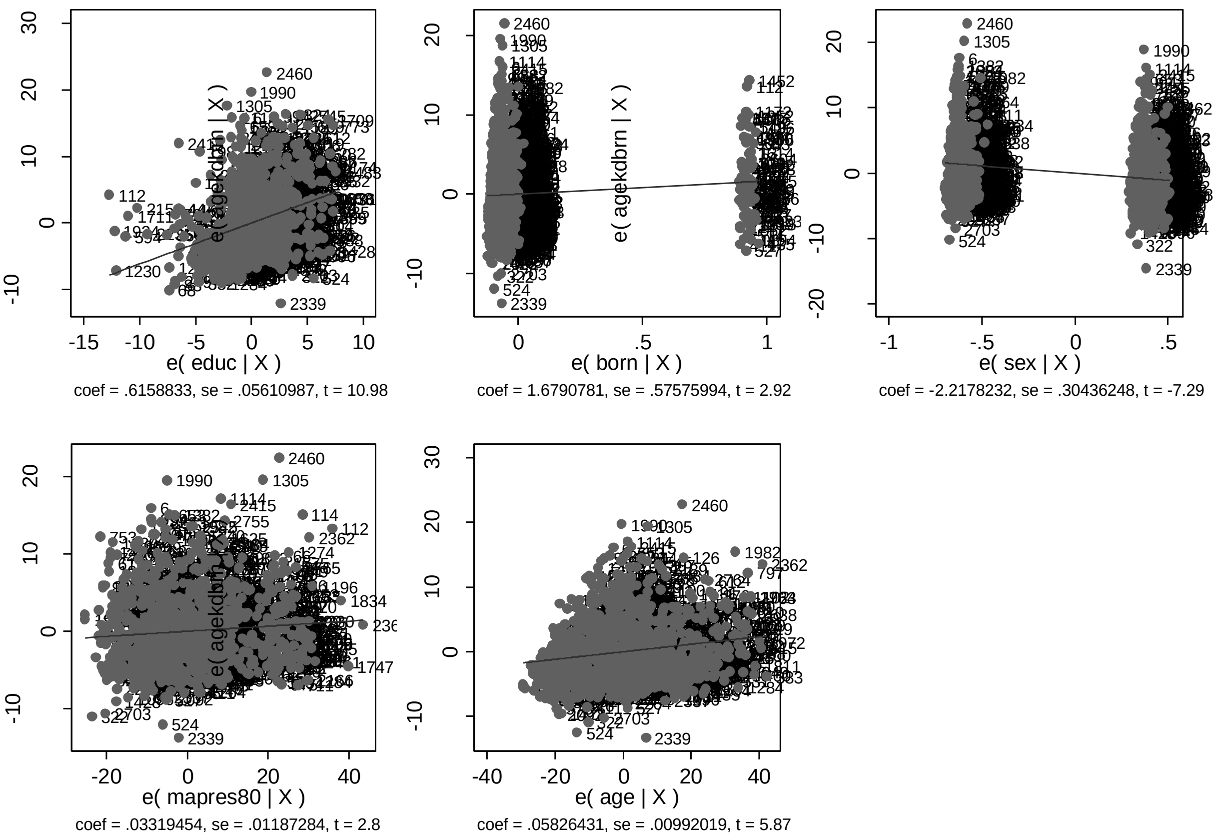
```
.lvr2plot, mlabel(id)
```



There are many observations with high leverage and residuals; we would be especially concerned about 112, 1934, 2460, 1452 etc.

Added variable plots (avplots) is another tool we can use to identify outliers and leverage points – in this case, we can see them in relationship to the slopes. Note that you can also obtain these plots one by one using avplot command, e.g. avplot educ, mlabel(id)

```
.avplots, mlabel(id)
```



Observation #2460 is the first one that looks especially suspicious - that's an outlier, a high residual observation; same thing with 1305. Looks like these are people who had their first child very late in life. As for high leverage observations, not too many stand out on this graph, although #112 might be one – looks like that might be a foreign born individual with very little education who had their first child relatively late in life.

To supplement these graphs, we can use a number of influence statistics that combine information on outlier status and leverage -- DFITS, Welsch's D, Cook's D, COVRATIO, and DFBETAs. It is usually a good idea to obtain a range of those to decide which cases are really problematic.

It makes sense to list the values of your dependent and independent variables for those observations that have values of these measures above the suggested cutoffs.
E.g., we get Cook's D (based on hat values and standardized residuals):

```
. predict cooksd if e(sample), cooksd
```

Don't forget to specify "if e(sample)" here – Cook's D is available out of sample as well!

NOTE: if you already generated a variable with this name (e.g. cooksd) but want to reuse the name, just use the drop command first: e.g., drop cooksd

Now we list those observations with high Cook's distance. The cutoff is 4/n so in this case, it's 4/1089=.00367309.

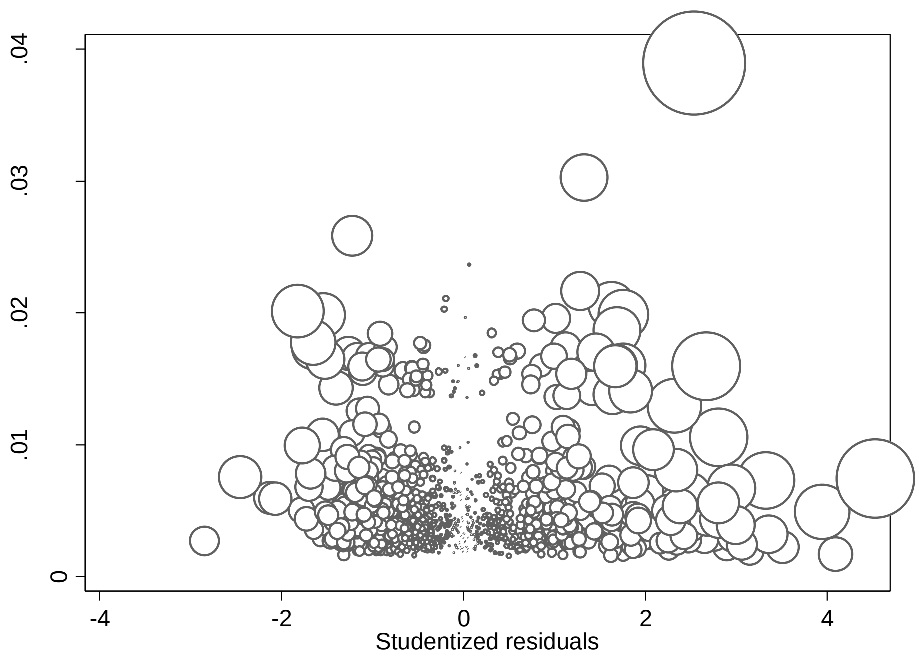
```
. sort cooksd
. list id agekdbrn educ born sex mapres80 age cooksd if cooksd>=4/1089 & cooksd~=.
      +--------------------------------------------------------------------+
      |   id   agekdbrn   educ   born      sex   mapres80   age     cooksd |
      |--------------------------------------------------------------------|
1031. | 1394         30     15     no   female         28    33   .0036766 |
1032. |   63         19     19    yes   female         34    64    .003683 |
1033. | 2484         37     17    yes   female         52    56   .0037003 |
1034. | 1906         29     10     no     male         23    39   .0037224 |
1035. |  994         38     15    yes   female         33    41    .003788 |
      |--------------------------------------------------------------------|
1036. |   22         19     12     no     male         44    23   .0038182 |
1037. | 1402         37     12    yes     male         33    42   .0038667 |
1038. |  742         36     13    yes     male         28    39   .0038726 |
1039. |  366         37     17    yes     male         66    44   .0041899 |
1040. | 2265         39     17    yes     male         52    55    .004212 |
      |--------------------------------------------------------------------|
1041. | 2703         16     16    yes     male         23    45    .004219 |
1042. | 1284         17     12    yes   female         64    76   .0043403 |
1043. | 2764         35     12    yes     male         23    75   .0044005 |
1044. | 1114         39     12    yes   female         46    46   .0044603 |
1045. | 2653         38     12    yes     male         32    43   .0044713 |
      |--------------------------------------------------------------------|
1046. |  322         13     16    yes   female         20    38   .0044766 |
1047. |  352         16      9     no   female         44    49   .0045471 |
1048. | 1382         39     12    yes     male         35    45   .0045595 |
1049. | 1990         42     13    yes   female         34    46   .0046982 |
1050. |  514         16     11     no   female         40    42   .0047655 |
      |--------------------------------------------------------------------|
1051. | 1186         30     12     no   female         30    44   .0049131 |
1052. |  669         37     18    yes   female         32    49    .005042 |
1053. | 1428         17     20    yes   female         32    28   .0052439 |
1054. |  753         35     13    yes   female         17    51   .0053052 |
1055. |  797         34     12    yes   female         35    83   .0054951 |
      |--------------------------------------------------------------------|
1056. |  126         38     15    yes   female         28    65   .0056446 |
1057. | 1824         41     16    yes     male         34    49   .0058367 |
1058. |    6         40     12    yes     male         29    47   .0059349 |
1059. |  447         26      6     no   female         23    55   .0060603 |
1060. | 1549         32     14     no   female         66    34   .0061423 |
      |--------------------------------------------------------------------|
1061. | 1066         32     13     no   female         47    40   .0062896 |
1062. |  612         36     18    yes   female         23    73   .0063017 |
1063. |  508         18     14     no   female         64    40   .0064009 |
1064. | 1747         24     17     no     male         86    36   .0065845 |
1065. | 1189         39     16    yes     male         23    62   .0066001 |
      |--------------------------------------------------------------------|
1066. |  773         37     20    yes   female         28    54   .0070942 |
1067. | 2545         42     18    yes     male         46    54   .0072636 |
1068. | 1709         38     20    yes   female         35    47   .0073801 |
1069. |  541         35     18     no   female         46    37   .0075467 |
1070. |  524         16     19    yes     male         42    34   .0075767 |
      |--------------------------------------------------------------------|
1071. |  430         35     18     no   female         44    38   .0075794 |
1072. | 1194         21     17     no   female         66    60   .0079331 |
```

```
1073. |  435          19     12      no     male         36     67   .0079604 |
1074. | 1172          33     14      no   female         32     39   .0080491 |
1075. |  411          21     18      no     male         51     30   .0082472 |
      |-------------------------------------------------------------------------|
1076. | 1952          31     12      no   female         20     40   .0083125 |
1077. | 1575          34     12      no     male         64     34   .0090088 |
1078. | 1934          25      0     yes     male         23     89    .009117 |
1079. | 1711          27      2     yes     male         36     69   .0093139 |
1080. |  114          37     12     yes   female         66     47   .0096068 |
      |-------------------------------------------------------------------------|
1081. | 2156          25      2     yes     male         20     33   .0104581 |
1082. |  527          22     20      no     male         44     43   .0112643 |
1083. | 2362          36     12     yes   female         64     83   .0117106 |
1084. | 1305          44     12     yes     male         56     53   .0125958 |
1085. | 2415          35      7     yes   female         42     48   .0133718 |
      |-------------------------------------------------------------------------|
1086. | 1982          37      8     yes     male         30     83   .0139673 |
1087. | 1452          41     16      no     male         36     47   .0191272 |
1088. | 2460          50     16     yes     male         64     62   .0251248 |
1089. |  112          32      2      no     male         63     38   .0434919 |
      +-------------------------------------------------------------------------+
```

That's quite a few; the largest Cook's D belong to observations 112, 2460, and 1452. All of those stood out in graphs as well, so we want to investigate those, but first we might want to examine other indices (e.g. DFITS, COVRATIO, etc.) as well. In the end, we want to identify and further investigate those observations that are consistently problematic across a range of diagnostic tools.

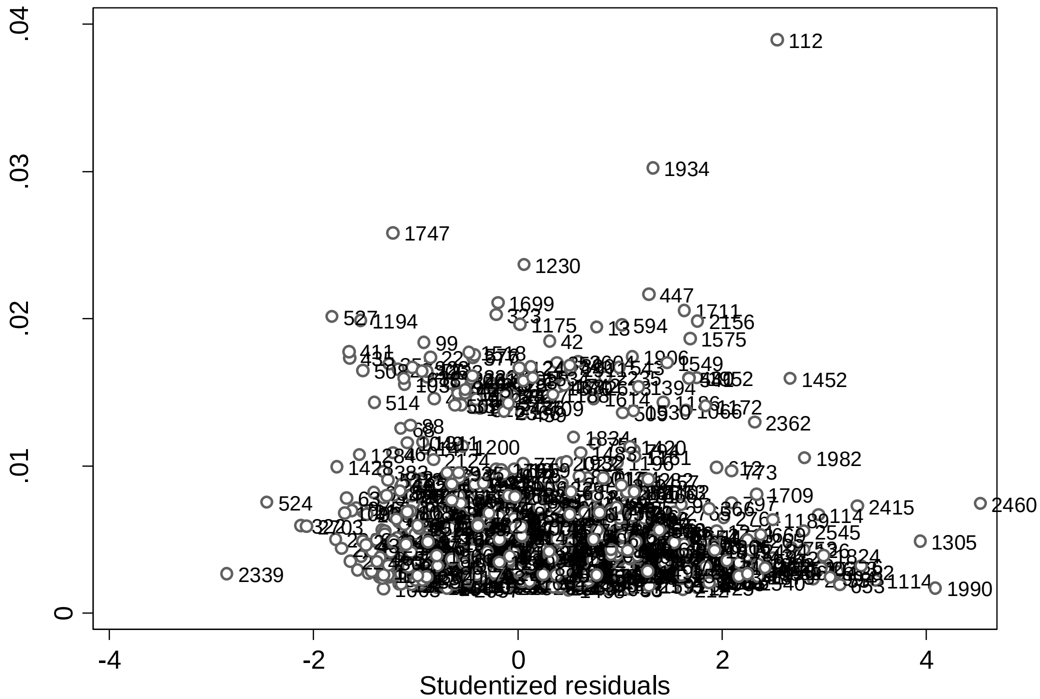
E.g., we can combine the information on high leverage, high studentized residual, and Cook's D:

```
.scatter hats rstudent [w=cooksd] , mfc(white)
```



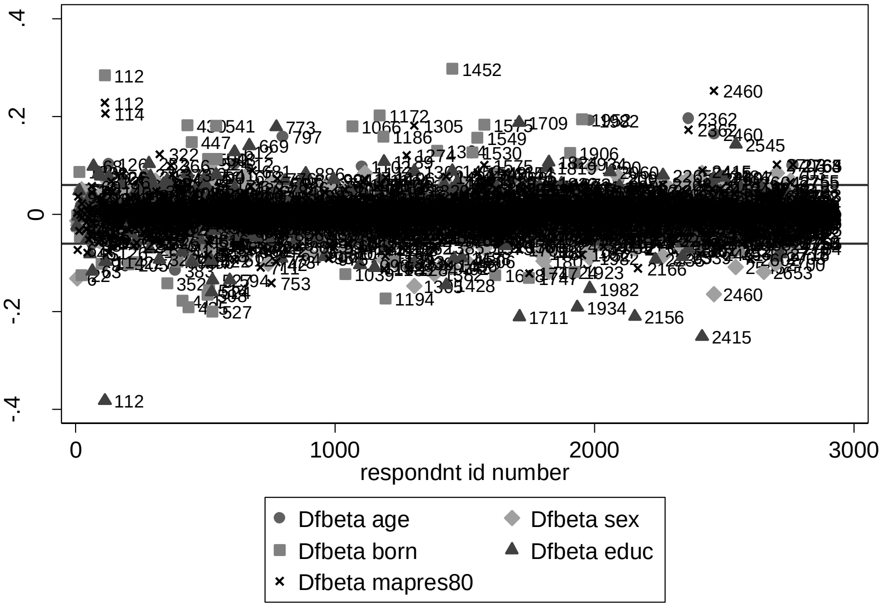
To identify problematic observations, let's replace circles with ID numbers:

```
. scatter hats student [w=cooksd] , mlabel(id)
```



Another set of index measures of influence, DFBETAs, focuses on one regression coefficient at a time. It is a normalized measure of the effect of each specific observation on a regression coefficient, estimated by omitting each observation and comparing the resulting coefficient to the coefficient with that observation included in the data. Positive DFBETA value indicates that an observation increases the value of the coefficient; negative value indicates a decrease in the coefficient due to that observation.

```
. dfbeta
(1676 missing values generated)
                           DFeduc:  DFbeta(educ)
(1676 missing values generated)
                           DFborn:  DFbeta(born)
(1676 missing values generated)
                            DFsex:  DFbeta(sex)
(1676 missing values generated)
                       DFmapres80:  DFbeta(mapres80)
(1676 missing values generated)
                            DFage:  DFbeta(age)

. di 2/sqrt(1089)
.06060606

. scatter  DFage DFsex DFborn DFeduc  DFmapres80 id, yline(.06 -.06) mlabel(id id id id
id)
```



Observations 112 and 2460 seem to have influence on a number of coefficients; others seem to have effects on specific coefficients, so we need to look into those that have particularly large effects.

<u>Remedies:</u>

Once you detected influential data points, you need to decide what to do with them. Typically, non-influential outliers and leverage points do not concern us much, although outliers do increase error variance. We also want to watch out for clusters of outliers, which may suggest an omitted variable. But influential points can have dramatic effects, and we definitely want to investigate those. Once we find them, there is no one clear-cut solution. They should not be ignored, but neither should they be automatically deleted. Typically, the presence of an influential point can mean one of the following:

A. Our model is correct, the influential point can be attributed to some kind of measurement error

B. The value of the influential point is observed correctly, but our model is not correct in that it cannot model the influential point well. Possible reasons for that: (a) The relationship between the dependent and the independent variable is not linear in the interval of values that includes the influential point; (b) There is another explanatory variable that can help account for that influential point; (c) The model has heteroskedasticity problems. Unfortunately, often it is not possible to determine which one is the case. But here's what you can do:

1. You have to investigate what makes these data points unusual —- make sure that you examine their values on all of the variables you use. This will help identify potential data entry errors or might provide other clues as to why these data points are unusual. E.g., we could check #112:

```
. list agekdbrn educ born sex mapres80 age if id==112
     +-------------------------------------------------+
     | agekdbrn   educ   born    sex   mapres80   age |
     |-------------------------------------------------|
 10. |       32      2     no   male         63    38 |
     +-------------------------------------------------+
```

Let's also get averages for all variables to compare:

```
. sum agekdbrn educ born sex mapres80 age if e(sample)
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    agekdbrn |      1089    23.66483    5.352695         11         50
        educ |      1089     13.3168    2.719027          0         20
        born |      1089    1.070707    .2564527          1          2
         sex |      1089    1.624426    .4844932          1          2
    mapres80 |      1089    39.44077    12.95284         17         86
         age |      1089     46.1258    15.06822         19         89
```

2. If you are considering omitting unusual data, you should investigate whether omitting these data points changes the results of your regression model. Try omitting them one by one and compare the coefficients with and without them: are there large changes? Let's check what happens if we omit #112:

```
. reg agekdbrn educ born sex mapres80 age, beta
      Source |       SS       df       MS              Number of obs =    1089
-------------+------------------------------           F(  5,  1083) =   49.10
       Model |  5760.17098     5  1152.0342            Prob > F      =  0.0000
    Residual |   25412.492  1083  23.4649049           R-squared     =  0.1848
-------------+------------------------------           Adj R-squared =  0.1810
       Total |   31172.663  1088  28.6513447           Root MSE      =  4.8441
------------------------------------------------------------------------------
    agekdbrn |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
        educ |   .6158833   .0561099    10.98   0.000                 .3128524
        born |   1.679078   .5757599     2.92   0.004                 .0804462
         sex |  -2.217823   .3043625    -7.29   0.000                -.2007438
    mapres80 |   .0331945   .0118728     2.80   0.005                 .0803266
         age |   .0582643   .0099202     5.87   0.000                 .1640182
       _cons |   13.27142   1.252294    10.60   0.000                        .
------------------------------------------------------------------------------
. reg agekdbrn educ born sex mapres80 age if id~=112, beta
      Source |       SS       df       MS              Number of obs =    1088
-------------+------------------------------           F(  5,  1082) =   50.04
       Model |  5841.74787     5  1168.34957           Prob > F      =  0.0000
    Residual |  25261.3762  1082  23.3469281           R-squared     =  0.1878
-------------+------------------------------           Adj R-squared =  0.1841
       Total |  31103.1241  1087  28.6137296           Root MSE      =  4.8319
------------------------------------------------------------------------------
    agekdbrn |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
        educ |     .63726   .0565958    11.26   0.000                 .3214802
        born |   1.515919   .5778803     2.62   0.009                 .0722698
         sex |  -2.187693   .3038273    -7.20   0.000                -.1980863
    mapres80 |    .030491   .0118905     2.56   0.010                 .0737543
         age |   .0583569   .0098953     5.90   0.000                 .1644404
       _cons |   13.20334   1.249428    10.57   0.000                        .
```

The actual effect of that observation on the coefficients of educ, mapres80, and born are rather pretty small; for each, beta changes by about 0.01.

Also, try omitting the most persistent influential points as a group and examine the effects. If there are large changes in coefficients, you might use that to justify omitting a few (but only very few) observations from the model – but you will also have to explain what is so special about these cases.

3. To reduce the incidence of high leverage points, consider transforming skewed variables and/or topcoding/bottomcoding variables to bring univariate outliers closer to the rest of the distribution (e.g. coding incomes of >$100,000 to $100,000 so that these high values do not stand out), like we did when we discussed data screening (and if that was done at that stage, it reduces the chances that problems emerge in multivariate context).

4. If unusual data come in clusters, you may have to introduce another variable to control for their unusualness, or you might want to deal with them in a separate regression model.

5. Robust regression is another option when one observes substantial problems with influential data. The Stata rreg command performs a robust regression using iteratively reweighted least squares, i.e., assigning a weight to each observation with higher weights given to better behaved observations, while extremely unusual data can have their weights set to zero so that they are not included in the analysis at all.

```
. rreg agekdbrn educ born sex mapres80 age, gen(wt)
Robust regression                                      Number of obs =    1089
                                                       F(  5,  1083) =   52.34
                                                       Prob > F      =  0.0000
------------------------------------------------------------------------------
    agekdbrn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .6518023   .0539119    12.09   0.000     .5460186    .7575859
        born |   1.792079   .5532063     3.24   0.001     .7066014    2.877556
         sex |  -2.012778     .29244    -6.88   0.000    -2.586591   -1.438965
    mapres80 |   .0275798   .0114078     2.42   0.016      .005196    .0499637
         age |   .0522715   .0095316     5.48   0.000      .033569     .070974
       _cons |   12.34444   1.203239    10.26   0.000     9.983493    14.70538
------------------------------------------------------------------------------

. sum wt, det
                  Robust Regression Weight
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .2138941              0
 5%     .5965052       .0007363
10%     .7419349       .0035576       Obs                1089
25%     .8782627       .0726816       Sum of Wgt.        1089

50%     .9564363                      Mean           .9001565
                        Largest       Std. Dev.      .1513337
75%      .988214       .9999998
90%     .9983087       .9999999       Variance       .0229019
95%     .9996306              1       Skewness      -2.926814
99%     .9999847              1       Kurtosis       12.98754
```

Comparing the robust regression results with the OLS results on the previous page, we see that even though there are a few small differences, the coefficients, standard errors, and p-values are quite similar. Despite the minor problems with influential data that we observed while doing our diagnostics, the robust regression analysis yielded quite similar results, suggesting that these problems are indeed minor. If the results of OLS and robust regression were substantially different, we would need to further investigate what problems in our OLS model caused the difference. If it is impossible to resolve such problems, then the robust regression results should be viewed as more trustworthy.

4. Additivity.

First and foremost, we should always use our theory insights to consider the need for interactions. We can have interactions between dummies (or sets of dummies), a dummy (or a set of dummies) and a continuous variable, or two continuous variables. To avoid multicollinearity problems, you should code your dummies 0/1 and mean-center those continuous variables that are involved in interaction terms.

```
. gen sexd=sex-1
. gen bornd=born-1
(6 missing values generated)

. for var age educ mapres80: sum X \ gen Xmean=X-r(mean)
->  sum age
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |      2751    46.28281    17.37049         18         89
->  gen agemean=age-r(mean)
(14 missing values generated)

->  sum educ
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        educ |      2753    13.36397    2.973924          0         20
->  gen educmean=educ-r(mean)
(12 missing values generated)

->  sum mapres80
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    mapres80 |      1619    40.96912    13.63189         17         86
->  gen mapres80mean=mapres80-r(mean)
(1146 missing values generated)
```

A user-written program "fitint" helps find statistically significant two-way interactions, so it can be used as a diagnostic tool.

```
. net search fitint
```

Click on: `fitint from http://fmwww.bc.edu/RePEc/bocode/f`

```
. fitint reg agekdbrn bornd sexd agemean educmean mapres80mean, twoway(bornd sexd
agemean educmean mapres80mean) factor(bornd sexd)
      Source |       SS       df       MS              Number of obs =    1089
-------------+------------------------------           F( 15,  1073) =   17.65
       Model |  6169.67284    15  411.311523           Prob > F      =  0.0000
    Residual |  25002.9902  1073   23.301948           R-squared     =  0.1979
-------------+------------------------------           Adj R-squared =  0.1867
       Total |   31172.663  1088  28.6513447           Root MSE      =  4.8272
------------------------------------------------------------------------------
    agekdbrn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Ibornd_1 |   1.710533   .9779923     1.75   0.081    -.2084614    3.629527
    _Isexd_1 |   -2.21852   .3179507    -6.98   0.000    -2.842395   -1.594644
     agemean |   .0587138   .0171439     3.42   0.001     .0250744    .0923532
    educmean |   .4551926   .0888308     5.12   0.000     .2808908    .6294943
mapres80mean |    .033156   .0203674     1.63   0.104    -.0068085    .0731205
   _Ibornd_1 |  (dropped)
    _Isexd_1 |  (dropped)
_IborXsex_~1 |   .1211157   1.271076     0.10   0.924    -2.372961    2.615193
```

```
   _Ibornd_1 |  (dropped)
     agemean |  (dropped)
_IborXagem~1 |   .0048469   .0568729     0.09   0.932    -.1067477    .1164415
   _Ibornd_1 |  (dropped)
    educmean |  (dropped)
_IborXeduc~1 |  -.2922046    .210566    -1.39   0.166    -.7053724    .1209631
   _Ibornd_1 |  (dropped)
mapres80mean |  (dropped)
_IborXmapr~1 |   .0046759   .0414082     0.11   0.910    -.0765743    .0859261
    _Isexd_1 |  (dropped)
_IsexXagem~1 |  -.0031427   .0207363    -0.15   0.880     -.043831    .0375455
    _Isexd_1 |  (dropped)
_IsexXeduc~1 |    .391932   .1146716     3.42   0.001     .1669259    .6169381
    _Isexd_1 |  (dropped)
_IsexXmapr~1 |  -.0005186    .024932    -0.02   0.983    -.0494397    .0484024
      __13_6 |  -.0038885   .0038209    -1.02   0.309    -.0113858    .0036088
      __14_6 |   .0004487   .0008266     0.54   0.587    -.0011732    .0020706
      __15_6 |   .0033919   .0044236     0.77   0.443     -.005288    .0120717
       _cons |   24.98069   .2579745    96.83   0.000      24.4745    25.48688
------------------------------------------------------------------------------
------------------------------------------------------------------------
Fitting and testing any interactions and any main effects not included
in interaction terms using the ratio of the mean square error of each
term and the residual mean square error to obtain an F ratio statistic
------------------------------------------------------------------------
Model summary
Number of observations used in estimation:    1089
Regression command:         regress
Dependent variable:        agekdbrn
Residual MSE:                 23.30
degrees of freedom:            1073
------------------------------------------------------------------------
      Term         |   Mean square    F ratio  df1    df2          P>F
-------------------+----------------------------------------------------
  i.bornd*i.sexd   |          0.21       0.01    1    1073      0.9241
  i.bornd*agemean  |          0.17       0.01    1    1073      0.9321
 i.bornd*educmean  |         44.87       1.93    1    1073      0.1655
 i.bornd*mapres80mean|        0.30       0.01    1    1073      0.9101
  i.sexd*agemean   |          0.54       0.02    1    1073      0.8796
  i.sexd*educmean  |        272.21      11.68    1    1073      0.0007
 i.sexd*mapres80mean|         0.01       0.00    1    1073      0.9834
 agemean*educmean  |         24.13       1.04    1    1073      0.3091
 agemean*mapres80mean|        6.87       0.29    1    1073      0.5874
 educmean*mapres80mean|      13.70       0.59    1    1073      0.4434
------------------------------------------------------------------------
```

It appears that when all twoway interactions are tested simultaneously, the only one that is statistically significant is sex by education. We could also check each two-way interaction separately to make sure we did not miss anything by testing all simultaneously:

```
. for X in var bornd sexd agemean educmean mapres80mean: for Y in var bornd sexd
agemean educmean mapres80mean: fitint reg agekdbrn bornd sexd agemean educmean
mapres80mean, twoway(Y X) factor(bornd sexd)
[output omitted]
```

Note that you should always include main effect variables in addition to the interaction, because the interaction term can only be interpreted together with that main effect. Further, if you want to explore three-way interactions, the model should also include all possible two-way interactions in addition to main terms. For example:

```
. gen bornsex=bornd*sexd
(6 missing values generated)
```

```
. gen borneduc=bornd*educmean
(13 missing values generated)
. gen educsex=educmean*sexd
(12 missing values generated)
. gen educsexborn=educmean*sexd*bornd
(13 missing values generated)
. xi: reg agekdbrn bornd sexd agemean educmean mapres80mean  bornsex borneduc educsex
educsexborn
      Source |       SS       df       MS              Number of obs =    1089
-------------+------------------------------           F(  9,  1079) =   29.48
       Model |  6152.90509     9  683.656121           Prob > F      =  0.0000
    Residual |  25019.7579  1079  23.1879128           R-squared     =  0.1974
-------------+------------------------------           Adj R-squared =  0.1907
       Total |   31172.663  1088  28.6513447           Root MSE      =  4.8154
------------------------------------------------------------------------------
    agekdbrn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       bornd |   1.779615   .9740461     1.83   0.068     -.131624    3.690854
        sexd |  -2.220267   .3134215    -7.08   0.000    -2.835252   -1.605282
     agemean |   .0594442   .0098793     6.02   0.000     .0400593     .078829
    educmean |   .4461687   .0831105     5.37   0.000     .2830922    .6092451
mapres80mean |   .0324834   .0118427     2.74   0.006     .0092461    .0557208
     bornsex |   .0946345   1.204318     0.08   0.937    -2.268437    2.457706
    borneduc |  -.4745646   .2819971    -1.68   0.093    -1.027889    .0787601
     educsex |   .3621368   .1124932     3.22   0.001     .1414065    .5828671
 educsexborn |   .4750623   .3902632     1.22   0.224    -.2906985    1.240823
       _cons |   25.00961   .2479526   100.86   0.000     24.52309    25.49614
-------------+----------------------------------------------------------------
```

But we'll focus on two-way interactions for now, and in order to explore how to interpret them, we'll review 4 examples: (1) an interaction of two dichotomous variables; (2) an interaction of a dummy variable and a continuous variable; (3) an interaction of a set of dummy variables and a continuous variable; (4) an interaction of two continuous variables.

<u>Example 1: Two dichotomous variables</u>

```
. reg  agekdbrn educ bornd##sexd mapres80 age
      Source |       SS       df       MS              Number of obs =    1089
-------------+------------------------------           F(  6,  1082) =   40.91
       Model |  5764.17997     6  960.696662           Prob > F      =  0.0000
    Residual |   25408.483  1082  23.4828863           R-squared     =  0.1849
-------------+------------------------------           Adj R-squared =  0.1804
       Total |   31172.663  1088  28.6513447           Root MSE      =  4.8459
------------------------------------------------------------------------------
    agekdbrn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .6165377   .0561537    10.98   0.000     .5063552    .7267202
     1.bornd |   1.358118   .9670434     1.40   0.160    -.5393752     3.25561
      1.sexd |  -2.251548   .3152298    -7.14   0.000    -2.870079   -1.633017
             |
  bornd#sexd |
        1 1  |   .4964787   1.201596     0.41   0.680    -1.861244    2.854201
             |
    mapres80 |   .0333659   .0118846     2.81   0.005     .0100464    .0566855
         age |   .0584314   .0099322     5.88   0.000     .0389428      .07792
       _cons |   12.73045   .9671152    13.16   0.000     10.83281    14.62808
------------------------------------------------------------------------------
```

The interaction is not statistically significant, but let's suppose it would be. Then we can first interpret the two main effects: the foreign born men have children 1.4 years later than the native born men, and the native born women have children 2.3 years earlier than the native born men.

To interpret the interaction term, we need to focus on one variable as our main variable and the other will be used as a moderator. We can do it both ways.

Nativity status as the main variable:
The effect of being foreign born is 1.4 for men (i.e., the foreign born men have children 1.4 years later than the native born men), but for women, it is 1.4+0.5=1.9 (that is, the foreign born women have children 1.9 years later than the native born women).

Gender as the main variable:
The effect of gender is -2.3 for the native born (i.e., the native born women have children 2.3 years earlier than the native born men), but for the foreign born, it is -2.3 +.5=-1.8 (that is, the foreign born women have children 1.8 years earlier than the foreign born men).

The only time when we would use both main effects and an interaction is when we wanted to compare across gender and nativity status at the same time: that is, the foreign born women have children 0.4 of a year earlier than the native born men: 1.4-2.3+0.5=-0.4

Although it doesn't make sense to examine an interaction of two dummy variables graphically, we can use "adjust" command to help us interpret this interaction:

```
. xi: qui reg  agekdbrn educ i.bornd*sexd mapres80 age
. adjust educ mapres80 age if e(sample), by(sexd bornd)
-------------------------------------------------------------------------------
Dependent variable: agekdbrn     Command: regress
   Variables left as is: _Ibornd_1, _IborXsexd_1
 Covariates set to mean: educ = 13.316804, mapres80 = 39.440773, age = 46.125805
---------------------------------------------------------------------------------------
----------------------
          |      bornd
     sexd |       0        1
----------+-----------------
        0 | 24.9519    26.31
        1 | 22.7004   24.555
----------------------------
     Key:  Linear Prediction
```

These are the predicted values of agekdbrn given average values of education, age, and mother's occupational prestige.

<u>Example 2: A dummy variable and a continuous variable</u>

```
. reg  agekdbrn bornd##c.educmean sexd mapres80 age
      Source |       SS       df       MS              Number of obs =    1089
-------------+------------------------------           F(  6,  1082) =   41.17
       Model |   5793.5421     6  965.590349           Prob > F      =  0.0000
    Residual |  25379.1209  1082  23.4557494           R-squared     =  0.1859
-------------+------------------------------           Adj R-squared =  0.1813
       Total |   31172.663  1088  28.6513447           Root MSE      =  4.8431
----------------------------------------------------------------------------------
        agekdbrn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------+----------------------------------------------------------------
         1.bornd |   1.716336   .5764944     2.98   0.003      .585162    2.847509
        educmean |   .6352486    .058401    10.88   0.000     .5206565    .7498407
                 |
bornd#c.educmean |
```

```
           1  |  -.2323323    .194782    -1.19   0.233    -.6145255    .1498609
              |
         sexd |  -2.229199   .3044525    -7.32   0.000    -2.826583   -1.631814
     mapres80 |   .0334778   .0118729     2.82   0.005     .0101813    .0567743
          age |   .0587405   .0099263     5.92   0.000     .0392636    .0782175
        _cons |   20.93833   .7498733    27.92   0.000     19.46696     22.4097
------------------------------------------------------------------------------
```
Again, no significant interaction, but for practice, we'll interpret the results.

Education as the main variable, nativity status as the moderator:
Among the native born individuals, a one year increase in education is associated with a 0.6 of a year increase in the age of having kids. Among the foreign born individuals, a one year increase in education is associated with a (.63-.23)=0.4 of a year increase in the age of having kids.

Nativity status as the main variable, education as the moderator:
Among those with average education (13.4 years), the foreign born have kids 1.7 years later than the native born. Among those with education one unit above average (14.4 years), the foreign born have kids 1.5 years later than the native born (1.7+1*(-0.2)). Among those with education one unit below average (12.4 years), the foreign born have kids 1.9 years later than the native born (1.7 + (-1*(-0.2))). We could also look at those whose education is 4 years below average (9.4 years); for them, the foreign born have kids 2.5 years later than the native born (1.7 + (-4*(-0.2))).

We could estimate this model in a different way to see separately the effects of education in the native born and the foreign born groups; that will also allow us to see if the effect is significant in each of the groups:

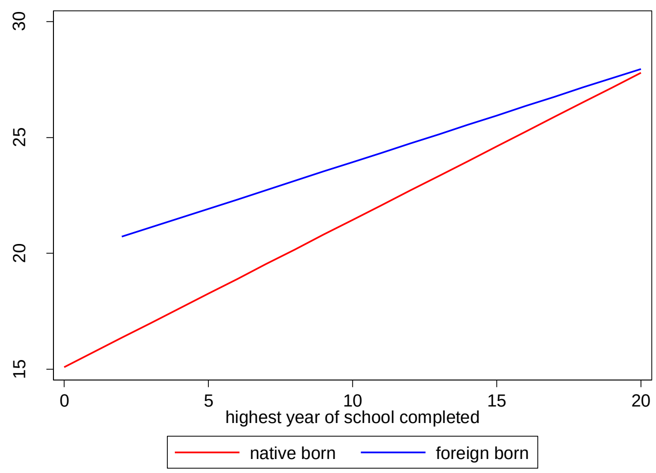
```
. gen educfb=educmean*bornd
(13 missing values generated)
. gen educnb=educmean
(12 missing values generated)
. replace educnb=0 if bornd==1
(256 real changes made)

. reg agekdbrn bornd educfb educnb sexd mapres80 age
      Source |       SS       df       MS              Number of obs =    1089
-------------+------------------------------           F(  6,  1082) =   41.17
       Model |   5793.5421     6  965.590349           Prob > F      =  0.0000
    Residual |  25379.1209  1082  23.4557494           R-squared     =  0.1859
-------------+------------------------------           Adj R-squared =  0.1813
       Total |   31172.663  1088  28.6513447           Root MSE      =  4.8431
------------------------------------------------------------------------------
    agekdbrn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       bornd |   1.716336   .5764944     2.98   0.003      .585162    2.847509
      educfb |   .4029163   .1871522     2.15   0.032     .0356939    .7701387
      educnb |   .6352486    .058401    10.88   0.000     .5206565    .7498407
        sexd |  -2.229199   .3044525    -7.32   0.000    -2.826583   -1.631814
    mapres80 |   .0334778   .0118729     2.82   0.005     .0101813    .0567743
         age |   .0587405   .0099263     5.92   0.000     .0392636    .0782175
       _cons |   20.93833   .7498733    27.92   0.000     19.46696     22.4097
------------------------------------------------------------------------------
```

This way we can see that the effect of education is significant in both groups.
Finally, we can again examine this interaction graphically.

```
. adjust sexd mapres80 age if e(sample), gen(pred1)
```

```
--------------------------------------------------------------------------------
Dependent variable: agekdbrn     Command: regress
       Created variable: pred1
   Variables left as is: bornd, educfb, educnb
 Covariates set to mean: sexd = .62442607, mapres80 = 39.440773, age = 46.125805
--------------------------------------------------------------------------------
----------------------
      All |         xb
----------+-----------
          |    23.6648
----------------------
     Key:  xb  =  Linear Prediction
```

```
. twoway (line pred1 educ if bornd==0, sort color(red) legend(label(1 "native born")))
(line pred1 educ if bornd==1, sort color(blue) legend(label(2 "foreign born"))
ytitle("Respondent's Age When 1st Child Was Born"))
```

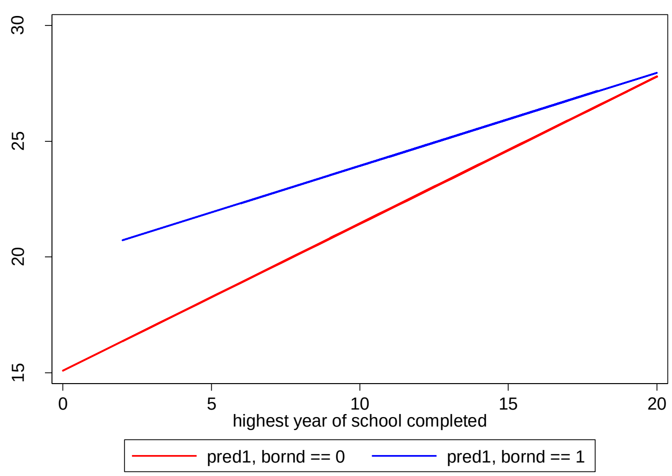


Alternatively, we could split pred1 into two variables (or if needed more):

```
.separate pred1, by(bornd)
```

This would generate two variables, pred10 and pred11, which we can graph:

```
.line pred10 pred11 educ, lcolor(red blue) sort
```



<u>Example 3: A set of dummy variables and a continuous variable</u>

```
. reg  agekdbrn bornd marital##c.educmean sexd mapres80 age

      Source |       SS       df       MS              Number of obs =    1089
-------------+------------------------------           F( 13,  1075) =   21.96
       Model |  6540.34346    13  503.103343           Prob > F      =  0.0000
    Residual |  24632.3195  1075  22.9137856           R-squared     =  0.2098
-------------+------------------------------           Adj R-squared =  0.2003
       Total |   31172.663  1088  28.6513447           Root MSE      =  4.7868

------------------------------------------------------------------------------------
          agekdbrn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
             bornd |   1.536577   .5729824     2.68   0.007     .4122865    2.660868
                   |
           marital |
          widowed  |  -.8946254    .626208    -1.43   0.153    -2.123354    .3341031
         divorced  |  -.9166076   .3889825    -2.36   0.019    -1.679859   -.1533567
        separated  |  -1.944692   .7095625    -2.74   0.006    -3.336977   -.5524077
    never married  |   -2.55648   .5380556    -4.75   0.000    -3.612238   -1.500722
                   |
          educmean |   .6467199   .0727279     8.89   0.000      .504015    .7894247
                   |
marital#c.educmean |
          widowed  |  -.3294696    .167311    -1.97   0.049    -.6577629   -.0011764
         divorced  |   .0213546    .151949     0.14   0.888    -.2767956    .3195049
        separated  |  -.0935184   .2455722    -0.38   0.703    -.5753736    .3883368
    never married  |   -.527267   .2268917    -2.32   0.020    -.9724677   -.0820662
                   |
              sexd |  -2.028997   .3066702    -6.62   0.000    -2.630737   -1.427257
          mapres80 |   .0292701   .0118022     2.48   0.013     .0061121    .0524282
               age |   .0435388   .0117499     3.71   0.000     .0204835    .0665942
             _cons |   22.24782   .8245124    26.98   0.000     20.62999    23.86566
------------------------------------------------------------------------------------
```

To test whether the set of interactions is jointly significant:

```
. mat list e(b)

e(b)[1,16]
                          1b.            2.            3.            4.            5.
          bornd      marital       marital       marital       marital       marital
y1    1.5365772            0    -.8946254    -.91660762    -1.9446921    -2.5564798

                 1b.marital#   2.marital#    3.marital#    4.marital#    5.marital#
       educmean  co.educmean   c.educmean    c.educmean    c.educmean    c.educmean
y1    .64671988            0    -.32946962    .02135465    -.09351841    -.52726698


           sexd     mapres80          age         _cons
y1   -2.0289968    .02927015    .04353882    22.247823

. test 2.marital#c.educmean 3.marital#c.educmean 4.marital#c.educmean
5.marital#c.educmean

 ( 1)  2.marital#c.educmean = 0
 ( 2)  3.marital#c.educmean = 0
 ( 3)  4.marital#c.educmean = 0
 ( 4)  5.marital#c.educmean = 0

       F(  4,  1075) =    2.22
            Prob > F =    0.0653
```

We cannot reject the null hypothesis, so we conclude that jointly these interaction effects are not statistically significant (they do not add significantly to the amount of variance explained by the model; although it is possible that with fewer groups, the overall significance test would change). If we were to explore these interaction terms, however, we would want to get the estimates of separate slopes of education by marital status:

```
. tab marital, gen(mardummy)
      marital |
       status |      Freq.     Percent        Cum.
--------------+-----------------------------------
      married |      1,269       45.90       45.90
      widowed |        247        8.93       54.83
     divorced |        445       16.09       70.92
    separated |         96        3.47       74.39
never married |        708       25.61      100.00
--------------+-----------------------------------
        Total |      2,765      100.00

. for num 1/5: gen educmarX=educmean*mardummyX
->  gen educmar1=educmean*mardummy1
(12 missing values generated)
->  gen educmar2=educmean*mardummy2
(12 missing values generated)
->  gen educmar3=educmean*mardummy3
(12 missing values generated)
->  gen educmar4=educmean*mardummy4
(12 missing values generated)
->  gen educmar5=educmean*mardummy5
(12 missing values generated)

. reg  agekdbrn bornd i.marital educmar1-educmar5 sexd mapres80 age
      Source |       SS       df       MS              Number of obs =    1089
-------------+------------------------------           F( 13,  1075) =   21.96
       Model |  6540.34346    13  503.103343           Prob > F      =  0.0000
    Residual |  24632.3195  1075  22.9137856           R-squared     =  0.2098
-------------+------------------------------           Adj R-squared =  0.2003
       Total |   31172.663  1088  28.6513447           Root MSE      =  4.7868
------------------------------------------------------------------------------
      agekdbrn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------+--------------------------------------------------------------
         bornd |   1.536577   .5729824     2.68   0.007     .4122865    2.660868
               |
       marital |
      widowed  |  -.8946254    .626208    -1.43   0.153    -2.123354    .3341031
     divorced  |  -.9166076   .3889825    -2.36   0.019    -1.679859   -.1533567
    separated  |  -1.944692   .7095625    -2.74   0.006    -3.336977   -.5524077
never married  |   -2.55648   .5380556    -4.75   0.000    -3.612238   -1.500722
               |
      educmar1 |   .6467199   .0727279     8.89   0.000      .504015    .7894247
      educmar2 |   .3172503   .1522423     2.08   0.037     .0185245     .615976
      educmar3 |   .6680745   .1348759     4.95   0.000     .4034246    .9327244
      educmar4 |   .5532015   .2360602     2.34   0.019     .0900105    1.016392
      educmar5 |   .1194529   .2155296     0.55   0.580    -.3034536    .5423594
          sexd |  -2.028997   .3066702    -6.62   0.000    -2.630737   -1.427257
      mapres80 |   .0292701   .0118022     2.48   0.013     .0061121    .0524282
           age |   .0435388   .0117499     3.71   0.000     .0204835    .0665942
         _cons |   22.24782   .8245124    26.98   0.000     20.62999    23.86566
------------------------------------------------------------------------------
```

It appears that education has a statistically significant effect on age of parenthood in all groups except for the never married.

Example 4: Two continuous variables

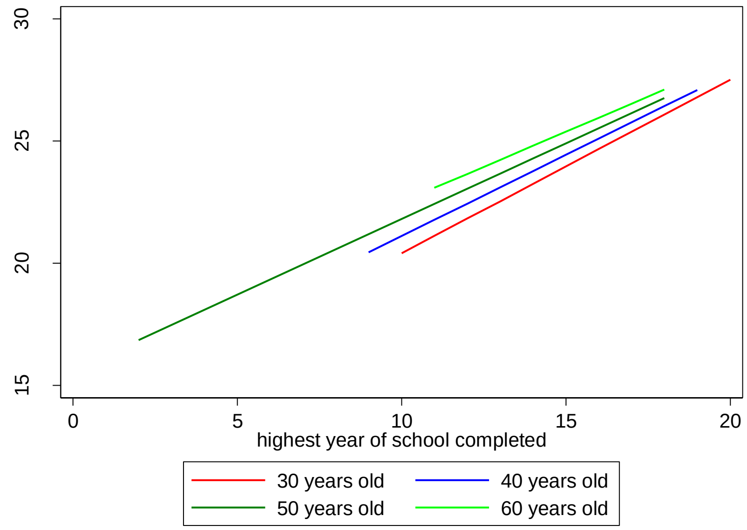
Both variables should be mean centered:

```
. reg  agekdbrn bornd c.educmean##c.agemean sexd mapres80
      Source |       SS       df       MS              Number of obs =    1089
-------------+------------------------------           F(  6,  1082) =   41.24
       Model |  5801.57307     6  966.928846           Prob > F      =  0.0000
    Residual |  25371.0899  1082  23.4483271           R-squared     =  0.1861
-------------+------------------------------           Adj R-squared =  0.1816
       Total |   31172.663  1088  28.6513447           Root MSE      =  4.8423
--------------------------------------------------------------------------------------
            agekdbrn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------------+----------------------------------------------------------------
               bornd |   1.679599   .5755567     2.92   0.004     .5502651    2.808932
            educmean |   .6362385   .0581443    10.94   0.000     .5221503    .7503268
             agemean |    .054804   .0102529     5.35   0.000     .0346862    .0749219
                     |
c.educmean#c.agemean |  -.0045353   .0034131    -1.33   0.184    -.0112324    .0021618
                     |
                sexd |  -2.232587   .3044578    -7.33   0.000    -2.829982   -1.635193
            mapres80 |   .0335181   .0118711     2.82   0.005      .010225    .0568111
               _cons |   23.64786     .52946    44.66   0.000     22.60897    24.68674
--------------------------------------------------------------------------------------
```

The interaction term is not significant. But if it were, to interpret it, we would pick one variable that's primary and the other one will serve as the moderator variable. E.g. if education is primary:
For agemean=0 (age at its mean, 46 y.o.), the effect of education is educmean coefficient, .6362385
For agemean=20 (age is at mean+20, i.e. 66 y.o.), the effect of education is
. di .6362385 + 20*-.0045353
.5455325
For agemean=-20 (age=26 y.o.), the effect of education is
. di .6362385 - 20*-.0045353
.7269445

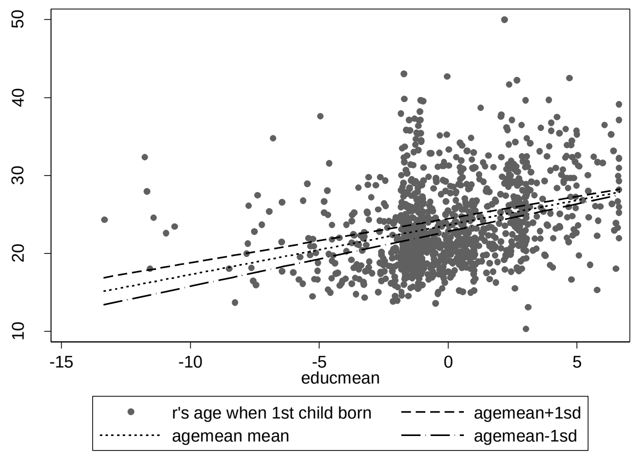
We can do the same thing graphically -- focus on one of the continuous variables and then graph it at various levels of the other one. E.g., we'll see how the effect of education varies by age:

```
. gen educage=educmean*agemean
(24 missing values generated)

. qui reg  agekdbrn bornd educmean agemean eudcage sexd mapres80

. qui adjust bornd sexd mapres80 if e(sample), gen(pred2)

. twoway (line pred2 educ if age==30, sort color(red) legend(label(1 "30 years old")))
(line pred2 educ if age==40, sort color(blue) legend(label(2 "40 years old")))  (line
pred2 educ if age==50, sort color(green) legend(label(3 "50 years old")))  (line pred2
educ if age==60, sort color(lime) legend(label(4 "60 years old")) ytitle("Respondent's
Age When 1st Child Was Born"))
```



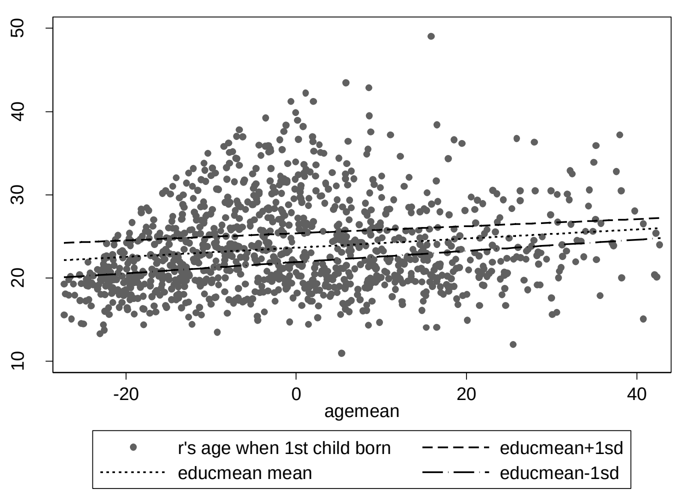
Here we can see that the higher one's age, the later they had their first child, but the effect of education becomes a little bit smaller with age (e.g. with age, the intercept becomes larger but the slope of education becomes smaller).We could have done it other way around – graph how agekdbrn is related to age for educational levels of, say, educ=10, 12, 14, 16, and 20.
There is also a user-written command that allows to automatically generate such a graph for three values – mean, mean+sd, mean-sd:

```
. net search sslope
Click on: sslope from http://fmwww.bc.edu/RePEc/bocode/s

. sslope agekdbrn bornd educmean sexd mapres80 agemean educage, i(educmean agemean
educage) graph
      Source |       SS       df       MS              Number of obs =    1089
-------------+------------------------------           F(  6,  1082) =   41.24
       Model |  5801.57308     6  966.928846           Prob > F      =  0.0000
    Residual |  25371.0899  1082  23.4483271           R-squared     =  0.1861
-------------+------------------------------           Adj R-squared =  0.1816
       Total |   31172.663  1088  28.6513447           Root MSE      =  4.8423
------------------------------------------------------------------------------
    agekdbrn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       bornd |   1.679599   .5755567     2.92   0.004     .5502651    2.808932
    educmean |   .6362385   .0581443    10.94   0.000     .5221503    .7503268
        sexd |  -2.232587   .3044578    -7.33   0.000    -2.829982   -1.635193
    mapres80 |   .0335181   .0118711     2.82   0.005      .010225    .0568111
     agemean |    .054804   .0102529     5.35   0.000     .0346862    .0749219
     educage |  -.0045353   .0034131    -1.33   0.184    -.0112324    .0021618
       _cons |   23.64786     .52946    44.66   0.000     22.60897    24.68674
-----------------------------------------------------------------
       Simple slope of agekdbrn on educmean at agemean  +/- 1sd
-----------------------------------------------------------------
    agemean |      Coef.   Std. Err.      t     P>|t|
------------+----------------------------------------------------
       High |   .5678996    .066709     8.51    0.000
       Mean |   .6362385   .0581443    10.94    0.000
        Low |   .7045775   .0871861     8.08    0.000
------------+----------------------------------------------------
```



Note that this gives us significance tests for the slope estimates at three levels of the moderator variable. If we reverse how we list the two main effect variables in the i() option of this command, we get:

```
. sslope agekdbrn bornd educmean sexd mapres80 agemean educage, i(agemean educmean
educage) graph
------------------------------------------------------------------
       Simple slope of agekdbrn on agemean at educmean  +/- 1sd
------------------------------------------------------------------
   educmean |       Coef.    Std. Err.       t     P>|t|
------------+-----------------------------------------------------
       High |    .0424724    .0154784      2.74   0.006
       Mean |     .054804    .0102529      5.35   0.000
        Low |    .0671357    .0119546      5.62   0.000
------------+-----------------------------------------------------
```



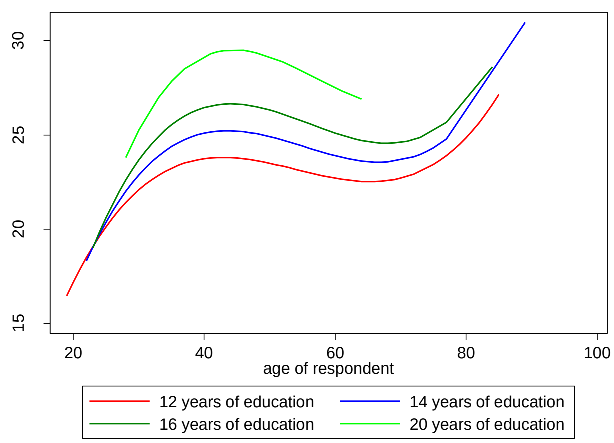
Finally, let's consider a more complicated case when we have a curvilinear relationship of age with agekdbrn and an interaction between age and education; we will right away create interaction terms to be able to use adjust command for graphs:

```
. gen agemean2=agemean^2
(14 missing values generated)
. gen agemean3=agemean^3
(14 missing values generated)
. gen educage2=educmean*agemean2
(24 missing values generated)
```

```
. gen educage3=educmean*agemean3
(24 missing values generated)
. reg  agekdbrn bornd sexd mapres80 educmean agemean agemean2 agemean3 educage educage2
educage3
      Source |       SS       df       MS              Number of obs =    1089
-------------+------------------------------           F( 10,  1078) =   35.55
       Model |  7731.43912    10  773.143912           Prob > F      =  0.0000
    Residual |  23441.2239  1078  21.7451056           R-squared     =  0.2480
-------------+------------------------------           Adj R-squared =  0.2410
       Total |   31172.663  1088  28.6513447           Root MSE      =  4.6632
------------------------------------------------------------------------------
    agekdbrn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       bornd |   1.278985     .556004     2.30   0.022     .1880122    2.369958
        sexd |  -2.113086    .2941837    -7.18   0.000    -2.690323   -1.535848
    mapres80 |   .0355671    .0114369     3.11   0.002     .0131259    .0580082
    educmean |   .7185734    .0759774     9.46   0.000      .569493    .8676538
     agemean |  -.0445573    .0182216    -2.45   0.015     -.080311   -.0088036
    agemean2 |  -.0064784    .0007326    -8.84   0.000    -.0079158    -.005041
    agemean3 |   .0002514    .0000327     7.69   0.000     .0001873    .0003155
     educage |  -.0001007     .005545    -0.02   0.986     -.010981    .0107796
    educage2 |  -.0008988    .0003225    -2.79   0.005    -.0015315   -.0002661
    educage3 |   .0000198    9.75e-06     2.03   0.042     6.87e-07     .000039
       _cons |   24.53094    .5244201    46.78   0.000     23.50194    25.55994
------------------------------------------------------------------------------
```

Indeed, significant interactions with the squared term and the cubed term.

```
. qui adjust bornd sexd mapres80 if e(sample), gen(pred3)
. twoway (line pred3 age if educ==12, sort color(red) legend(label(1 "12 years of
education"))) (line pred3 age if educ==14, sort color(blue) legend(label(2 "14 years
of education"))) (line pred3 age if educ==16, sort color(green) legend(label(3 "16
years of education"))) (line pred3 age if educ==20, sort color(lime) legend(label(4
"20 years of education")) ytitle("Respondent's Age When 1st Child Was Born"))
```



5. Multicollinearity

Our real life concern about the multicollinearity is that independent variables are highly (but not perfectly) correlated. Need to distinguish from perfect multicollinearity -- two or more independent variables are linearly related – in practice, this usually happens only if we make a mistake in including the variables; Stata will resolve this by omitting one of those variables and will tell you it did it. It can also happen when the number of variables exceeds the number of observations.

Perfect multicollinearity violates regression assumptions -- no unique solution for regression coefficients.

High, but not perfect, multicollinearity is what we most commonly deal with. High multicollinearity does not explicitly violate the regression assumptions - it is not a problem if we use regression only for prediction (and therefore are only interested in predicted values of Y our model generates). But it is a problem when we want to use regression for explanation (which is typically the case in social sciences) – in this case, we are interested in values and significance levels of regression coefficients. High degree of multicollinearity results in imprecise estimates of the unique effects of independent variables.

First, we can inspect the correlations among the variables; it allows us to see whether there are any high correlations, but does not provide a direct indication of multicollinearity:

```
. corr  educ born sex mapres80
(obs=1615)
             |     educ     born      sex mapres80
-------------+------------------------------------
        educ |   1.0000
        born |   0.0182   1.0000
         sex |   0.0066   0.0205   1.0000
    mapres80 |   0.2861   0.0169  -0.0423   1.0000
```

Variance Inflation Factors are a better tool to diagnose multicollinearity problems. These indicate how much the variance of a given coefficient estimate increases because of correlations of a certain variable with the other variables in the model. E.g. VIF of 4 means that the variance is 4 times higher than it could be, and the standard error is twice as high as it could be.

```
. reg agekdbrn educ born sex mapres80
      Source |       SS       df       MS              Number of obs =    1091
-------------+------------------------------           F(  4,  1086) =   51.24
       Model |  4954.03533     4  1238.50883           Prob > F      =  0.0000
    Residual |  26251.1232  1086   24.172305           R-squared     =  0.1588
-------------+------------------------------           Adj R-squared =  0.1557
       Total |  31205.1586  1090  28.6285858           Root MSE      =  4.9165

------------------------------------------------------------------------------
    agekdbrn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .6122718   .0569422    10.75   0.000     .5005426     .724001
        born |   1.360161   .5816506     2.34   0.020      .218875    2.501447
         sex |   -2.37973   .3075642    -7.74   0.000    -2.983218   -1.776243
    mapres80 |   .0243138   .0119552     2.03   0.042     .0008558    .0477718
       _cons |   16.95808   1.101139    15.40   0.000     14.79748    19.11868
------------------------------------------------------------------------------
. vif
    Variable |       VIF       1/VIF
-------------+----------------------
    mapres80 |      1.08    0.926124
        educ |      1.08    0.926562
        born |      1.00    0.998366
         sex |      1.00    0.999456
-------------+----------------------
    Mean VIF |      1.04
```

Different researchers advocate for different cutoff points for VIF. Some say that if any one of VIF values is larger than 4, there are some multicollinearity problems associated with that variable.

Others use cutoffs of 5 or even 10. In the example above, there are no problems with multicollinearity regardless of the cutoff we pick.

In addition, the following symptoms may indicate a multicollinearity problem:

- large changes in coefficients when adding or deleting variables
- non-significant coefficients for variables that you know are theoretically important
- coefficients with signs opposite of those you expected based on theory or previous results
- large standard errors in comparison to the coefficient size
- two (or more) large coefficients with opposite signs, possibly non-significant
- all or most coefficients are not significant even though the F-test indicates the entire regression model is significant

<u>Solutions for multicollinearity problems</u>:

1. See if you could create a meaningful scale from the variables that are highly correlated, and use that scale instead of the individual variables (i.e. several variables are reconceptualized as indicators of one underlying construct).

```
. sum mapres80 papres80
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    mapres80 |      1619    40.96912    13.63189         17         86
     papres80 |      2165    43.47206    12.40479         17         86
```

The variables have the same scales so we can add them:

```
. gen prestige=mapres80+papres80
(1519 missing values generated)
```

If the scales were different, we would first standardize each of them:

```
. egen papres80std = std(papres80)
(600 missing values generated)

. egen mapres80std = std(mapres80)
(1146 missing values generated)

. sum mapres80std papres80std
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
 mapres80std |      1619    4.12e-09           1  -1.758312   3.303348
 papres80std |      2165   -8.26e-11           1  -2.134019    3.42835

. gen prestige2=mapres80std+papres80std
(1519 missing values generated)

. pwcorr prestige prestige2

             | prestige presti~2
-------------+------------------
    prestige |   1.0000
   prestige2 |   0.9994   1.0000
```

We can now use prestige variable in subsequent OLS regressions. We might want to report a Chronbach's alpha – it indicates the reliability of the scale. It varies between 0 and 1, with 1 being perfect. Typically, alphas above .7 are considered acceptable, although some argue that those above .5 are ok.

```
. alpha mapres80 papres80
Test scale = mean(unstandardized items)
Average interitem covariance:     56.39064
Number of items in the scale:            2
Scale reliability coefficient:      0.5036
```

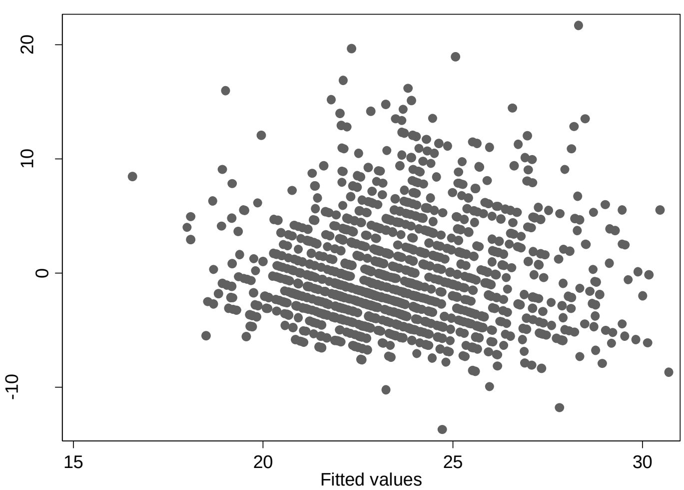
2. Consider if all variables are necessary. Try to primarily use theoretical considerations -- automated procedures such as backward or forward stepwise regression methods (available via "sw regress" command) are potentially misleading; they capitalize on minor differences among regressors and do not result in an optimal set of regressors. If not too many variables, examine all possible subsets.

3. If using highly correlated variables is absolutely necessary for correct model specification, you can use biased estimates. The idea here is that we add a small amount of bias but increase the efficiency of the estimates for those highly correlated variables. The most common method of this type is ridge regression (see http://members.iquest.net/~softrx/ for the Stata module).

6. Heteroscedasticity

The problem of heteroscedasticity commonly refers to non-constant error variance (that's opposite of homoscedasticity). We can examine this graphically as well as using formal tests. First, let's see if error variance changes across fitted values of our dependent variable:

```
. qui reg agekdbrn educ born sex mapres80 age
. rvfplot
```



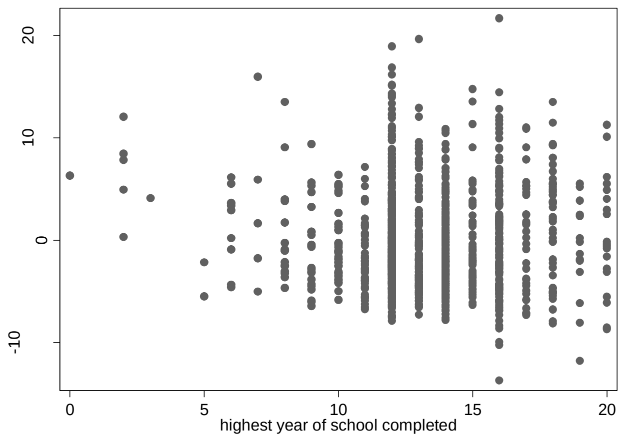
Can examine the same using a formal test:

```
. hettest
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of agekdbrn
         chi2(1)      =    21.44
         Prob > chi2  =   0.0000
```

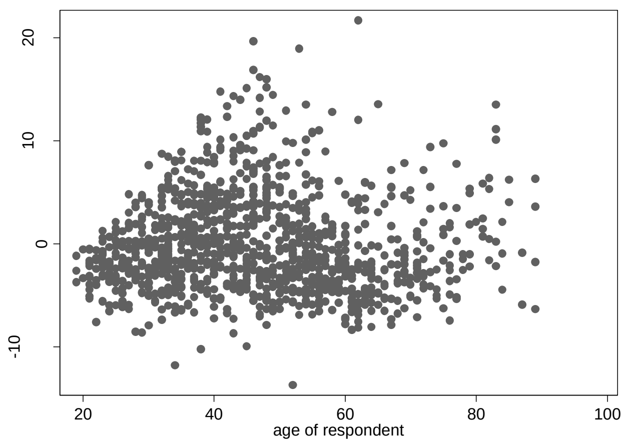
Since p<.05, we reject the null hypothesis of constant variance - the errors are heteroscedastic. Both the graph and the test indicate that the error variance is nonconsant (note the megaphone pattern).

Now let's search if there is any systematic relationship between error variance and individual regressors. First, graphical examination:

```
. rvpplot educ
```

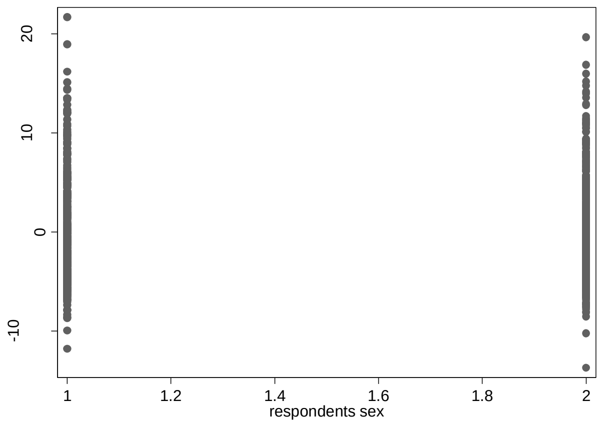



```
.rvpplot age
```



We can see the heteroscedasticity in both graphs, but it is much more severe for age. For a dummy variable, it is more difficult to examine it graphically:

```
. rvpplot sex
```



Now, let's use a formal test to examine the patterns of error variance across individual regressors:

```
. hettest, rhs mtest
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
---------------------------------------
    Variable |      chi2   df      p
-------------+-------------------------
        educ |      5.87    1   0.0154 #
        born |      0.00    1   0.9810 #
         sex |      9.19    1   0.0024 #
    mapres80 |      1.45    1   0.2279 #
         age |     10.26    1   0.0014 #
-------------+-------------------------
simultaneous |     25.78    5   0.0001
---------------------------------------
                  # unadjusted p-values
```

It looks like a number of regressors are responsible for our problems.

<u>Remedies:</u>

1. Transformations might help – it is especially important to consider the distribution of the dependent variable. As we discussed above, it is typically desirable, and can help avoid heteroscedasticity as well as non-normality problems, if the dependent variable is normally distributed. Let's examine whether the transformation we identified – reciprocal square root – would solve our heteroscedasticity problem.

```
. gen agekdbrnrr=1/(sqrt(agekdbrn))
(810 missing values generated)

. reg  agekdbrnrr educ born sex mapres80 age
      Source |       SS       df       MS              Number of obs =    1089
-------------+------------------------------           F(  6,  1082) =   48.07
       Model |   .11381105     6  .018968508           Prob > F      =  0.0000
    Residual |  .426934693  1082  .000394579           R-squared     =  0.2105
-------------+------------------------------           Adj R-squared =  0.2061
       Total |  .540745743  1088  .000497009           Root MSE      =  .01986
------------------------------------------------------------------------------
  agekdbrnrr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0024213   .0002353   -10.29   0.000    -.0028829   -.0019597
        born |  -.0070982   .0023638    -3.00   0.003    -.0117363   -.0024602
         sex |   .0095887   .0012506     7.67   0.000     .0071349    .0120425
```

```
    mapres80 |  -.0001494   .0000487    -3.07   0.002     -.000245   -.0000539
     agemean |  -.0003115   .0000434    -7.18   0.000    -.0003967   -.0002264
    agemean2 |   8.86e-06   2.29e-06     3.87   0.000     4.37e-06    .0000134
       _cons |   .2373519   .0046505    51.04   0.000      .228227    .2464769
------------------------------------------------------------------------------

. hettest
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of agekdbrnrr

         chi2(1)      =     0.35
         Prob > chi2  =   0.5566
. hettest, rhs mtest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance

---------------------------------------
    Variable |      chi2   df      p
-------------+-------------------------
        educ |      0.63    1   0.4262 #
        born |      0.26    1   0.6111 #
         sex |      0.29    1   0.5932 #
    mapres80 |      0.73    1   0.3939 #
         age |      1.71    1   0.1911 #
-------------+-------------------------
simultaneous |      3.06    5   0.6900
---------------------------------------
                  # unadjusted p-values
```

The heteroscedasticity problem has been solved. As I mentioned earlier, however, it is important to check that we did not introduce any nonlinearities by this transformation, and overall, transformations should be used sparsely - always consider ease of model interpretation as well. Also, sometimes when searching for a transformation to remedy heteroscedasticity, Box-Cox transformations can be very helpful, including the "transform both sides" (TBS) approach (see boxcox command).

2. Sometimes, dealing with outliers, influential observations, and nonlinearities might also help resolve heteroscedasticity problems. That is why I recommend testing with heteroscedasticity only after you've dealt with other problem.

3. Heteroscedasticity can also be a sign that some important factor is omitted, so you might want to rethink your model specification.

4. If nothing else works, we can obtain robust variance estimates using robust option in regress command (note that this is different from robust regression estimated by rreg!). These variance estimates do not rely on distributional assumptions and are therefore not sensitive to heteroscedasticity:

```
. reg agekdbrn educ born sex mapres80 age, robust
Linear regression                                      Number of obs =    1089
                                                       F(  5,  1083) =   47.74
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1848
                                                       Root MSE      =  4.8441
```

agekdbrn	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.6158833	.0640298	9.62	0.000	.4902467	.7415199
born	1.679078	.5756992	2.92	0.004	.5494661	2.80869
sex	-2.217823	.3143631	-7.05	0.000	-2.834653	-1.600993
mapres80	.0331945	.0122934	2.70	0.007	.009073	.0573161
age	.0582643	.0088246	6.60	0.000	.0409491	.0755795
_cons	13.27142	1.239779	10.70	0.000	10.83877	15.70406