Workshop on econometric analysis using Stata
Enugu, Nigeria, 26-31 July 2010

An introduction to
linear regressions with Stata

Abdelkrim Araar

ULAVAL and AIAE
Regression: a practical approach (overview)

We can use regressions to estimate the effect of changing one variable over another.

When running a linear regression, we are assuming a linear relationship between one variable and a set of N others: \( y = x_1 + x_2 + ... + x_N \).

In Stata, we use the command `regress`:

```plaintext
regress [dependent variable] [independent variable(s)]
regress y x
```

for a bivariate as well as for a multivariate setting:

```plaintext
regress y x1 x2 x3 ...
```

Before running a regression, it is useful to have a good rationale for what we are trying to estimate. A regression makes sense only if there is a sound theory behind it.
Regression: a practical approach (setting)

Example: Are per capita expenditures lower in households with a larger number of children, controlling for other possible factors?*

- Outcome (y) variable – Log of per capita expenditures
- Predictor (x) variables:
  - Number of children
  - Sex of household head
  - Age of household head
  - Household size
  - Rural/Urban area
  - Education level of household head

*Source: ECAMII, Cameroonian household survey with 10992 observations
Regression: variables

A useful first step is to examine the model variables to check for possible coding errors; for this, we can type:

```
use data/cameroon2001.dta
des
```

Contains data from C:\Documents and Settings\Client\Bureau\cameroon2001.dta
obse: 10,992
vars: 10
size: 296,784 (99.9% of memory free)

<table>
<thead>
<tr>
<th>variable name</th>
<th>storage</th>
<th>display</th>
<th>value</th>
<th>variable label</th>
</tr>
</thead>
<tbody>
<tr>
<td>strata</td>
<td>byte</td>
<td>%8.2f</td>
<td>s00g1</td>
<td>Stratum in which the household lives</td>
</tr>
<tr>
<td>psu</td>
<td>int</td>
<td>%8.2f</td>
<td>PSU of the household</td>
<td></td>
</tr>
<tr>
<td>sex</td>
<td>byte</td>
<td>%8.2f</td>
<td>sex</td>
<td>Sex of Household head</td>
</tr>
<tr>
<td>age</td>
<td>byte</td>
<td>%8.2f</td>
<td>age</td>
<td>Age of Household head</td>
</tr>
<tr>
<td>size</td>
<td>byte</td>
<td>%7.0f</td>
<td>size</td>
<td>Household size</td>
</tr>
<tr>
<td>nchild</td>
<td>byte</td>
<td>%8.2f</td>
<td>nchild</td>
<td>Number of children</td>
</tr>
<tr>
<td>area</td>
<td>float</td>
<td>%9.0g</td>
<td>area</td>
<td>Rural/Urban area</td>
</tr>
<tr>
<td>inst_lev</td>
<td>float</td>
<td>%15.0g</td>
<td>inst_lev</td>
<td>Education level of household head</td>
</tr>
<tr>
<td>pcexp</td>
<td>float</td>
<td>%9.0g</td>
<td>pcexp</td>
<td>Total expenditures per capita</td>
</tr>
<tr>
<td>weight</td>
<td>float</td>
<td>%8.2f</td>
<td>weight</td>
<td>Sampling weight</td>
</tr>
</tbody>
</table>

Sorted by:
```
. sum
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>strata</td>
<td>10992</td>
<td>6.202569</td>
<td>3.460040</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>psu</td>
<td>10992</td>
<td>342.8656</td>
<td>164.1245</td>
<td>1</td>
<td>612</td>
</tr>
<tr>
<td>sex</td>
<td>10992</td>
<td>1.243905</td>
<td>.4294554</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>age</td>
<td>10992</td>
<td>42.92704</td>
<td>15.06351</td>
<td>13</td>
<td>99</td>
</tr>
<tr>
<td>size</td>
<td>10992</td>
<td>5.134916</td>
<td>3.518811</td>
<td>1</td>
<td>38</td>
</tr>
<tr>
<td>nchild</td>
<td>10992</td>
<td>2.344432</td>
<td>2.557002</td>
<td>0</td>
<td>28</td>
</tr>
<tr>
<td>area</td>
<td>10992</td>
<td>.352984</td>
<td>.4779195</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>inst_lev</td>
<td>10992</td>
<td>2.222434</td>
<td>.9344423</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>pcexp</td>
<td>10992</td>
<td>433730.4</td>
<td>598837</td>
<td>17955.5</td>
<td>2.04e+07</td>
</tr>
<tr>
<td>weight</td>
<td>10992</td>
<td>283.9279</td>
<td>255.0097</td>
<td>15.71085</td>
<td>2048.463</td>
</tr>
</tbody>
</table>
Regression: what to look for

Let the regression be:

```plaintext
gen y=log(pcexp)
```

```plaintext
regress y nchild, robust
```

- **Outcome variable (y)**
- **Predictor variable (x)**
- **Robust standard errors (to control for heteroskedasticity)**
Regression: what to look for

Linear regression

Root MSE: root mean squared error is the standard deviation of the regression. The closer to zero, the better the fit.

| y       | Coef | Robust Std. Err. | t    | P>|t| | [95% Conf. Interval] |
|---------|------|------------------|------|-----|---------------------|
| nchild  | -.1168735 | .0029749 | -39.29 | 0.000 | -.1227048 to -.110422 |
| _cons   | 12.90595  | .0099843  | 1292.53 | 0.000 | 12.88638 to 12.92552 |

Number of obs = 10992  
F( 1, 10990) = 1543.44  
Prob > F = 0.0000  
R-squared = 0.1499  
Root MSE = 0.7117

This is the p-value of the overall model. It serves to test whether R2 is greater than 0. A p-value lower than 0.05 is usually needed to show a statistically significant relationship between Y and the X.

R-square shows the amount of the variance of Y explained by X. Here nchild explains about 15% of the total variance of y.

Adj R2 (not shown here) shows the same as R2 but is adjusted by the # of cases and of variables. When the # of variables is small and the # of cases is large, Adj R2 is close to R2. This provides a more accurate assessment of the association between X and Y.

y = 12.905 – 0.117 nchild + residual
Each child is estimated to decrease the log of per capita expenditures by 0.117

The t-values can serve to test the hypothesis that a coefficient is different from 0. For this, a t-value greater than 1.96 (for 95% confidence) is usually needed. t-values can be obtained by dividing the coefficient by its standard error.

Two-tail p-values help test the hypothesis that a coefficient is different from 0. For this, a p-value lower than 0.05 is usually sought. Here, the coefficient on nchild would be deemed to be statistically different from 0.
Regression with dummies (the xi prefix)

Regression with dummies: *level of instruction* entered here as a dummy variable. Dummy variables can be easily added to a regression by using “xi” and the prefix “i.” The first category is always the reference:

```stata
xi: regress y nchild i.inst_lev, robust
i.inst_lev    _Iinst_lev_1-4  (naturally coded;  _Iinst_lev_1 omitted)
```

|                | Coef.  | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|----------------|--------|-----------|-------|------|----------------------|
| nchild         | -.1006051 | .0020703  | -37.07| 0.000| -.1143114 - -.1030509|
| _Iinst_lev_2   | .1747328  | .0154707  | 11.29 | 0.000| .1444074  .2050583  |
| _Iinst_lev_3   | .5055329  | .0154676  | 32.68 | 0.000| .4752135  .5358522  |
| _Iinst_lev_4   | 1.099058  | .0282925  | 38.85 | 0.000| 1.0436    1.154517  |
| _cons          | 12.57537  | .0129787  | 968.92| 0.000| 12.54993  12.60081  |
## Regression: ANOVA table

Running a regression without the ‘robust’ option gives the ANOVA table

\[ \text{xi: regress y nchild size i.area i.inst_lei i.sex} \]

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 10992</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2615.82707</td>
<td>7</td>
<td>373.689582</td>
<td>F(7,10984) = 1043.79</td>
</tr>
<tr>
<td>Residual</td>
<td>3932.40385</td>
<td>10984</td>
<td>.358012004</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>6548.23092</td>
<td>10991</td>
<td>.595781178</td>
<td>R-squared = 0.3995</td>
</tr>
</tbody>
</table>

\[ \text{Adj R-squared} = 0.3991 \]
\[ \text{Root MSE} = .59834 \]

\[
F = \frac{MSS}{(k-1)} \frac{RSS}{n-k}
\]

\[
R^2 = \frac{MSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}
\]

\[
\text{RootMSE} = \sqrt{\frac{RSS}{(n-k)}}
\]

\[
\text{Adj} R^2 = 1 - \frac{n-1}{n-k} (1 - R^2)
\]

A = Model Sum of Squares (MSS). The closer to TSS, the better the fit.
B = Residual Sum of Squares (RSS)
C = Total Sum of Squares (TSS)
D = Average Model Sum of Squares = MSS/(k-1) where \( k \) = # predictors
E = Average Residual Sum of Squares = RSS/(n –k) where \( n \) = # of observations
F = Average Total Sum of Squares = TSS/(n-1)

\( R^2 \) shows the share of observed variance that is explained by the model, (Here equal to 40%).
The \( F \)-statistic, \( F(7,10984) \), tests whether \( R^2 \) is different from zero.
Root MSE shows the average distance of the estimator from its mean, (Here, about 60%).
Regression: estto/esttab

To show the models side-by-side, one can use the commands estto and esttab:

```
. gen y=log(pcexp)
. xi: regress y nchild, robust
. eststo model1
. xi: regress y nchild size i.area i.inst_lev
. eststo model2
. xi: regress y nchild size i.area i.inst_lev i.sex
. eststo model3
. esttab, r2 ar2 se scalar(rmse)
```

<table>
<thead>
<tr>
<th></th>
<th>(1) y</th>
<th>(2) y</th>
<th>(3) y</th>
</tr>
</thead>
<tbody>
<tr>
<td>nchild</td>
<td>-0.117***</td>
<td>-0.0264***</td>
<td>-0.0264***</td>
</tr>
<tr>
<td></td>
<td>(0.00297)</td>
<td>(0.00410)</td>
<td>(0.00410)</td>
</tr>
<tr>
<td>size</td>
<td>-0.0738***</td>
<td>-0.0744***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00298)</td>
<td>(0.00299)</td>
<td></td>
</tr>
<tr>
<td>_Iarea_1</td>
<td>-0.435***</td>
<td>-0.438***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0128)</td>
<td>(0.0128)</td>
<td></td>
</tr>
<tr>
<td>_Iinst_lev_2</td>
<td>0.107***</td>
<td>0.104***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0151)</td>
<td>(0.0152)</td>
<td></td>
</tr>
<tr>
<td>_Iinst_lev_3</td>
<td>0.354***</td>
<td>0.350***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0154)</td>
<td>(0.0156)</td>
<td></td>
</tr>
<tr>
<td>_Iinst_lev_4</td>
<td>0.908***</td>
<td>0.901***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0239)</td>
<td>(0.0243)</td>
<td></td>
</tr>
<tr>
<td>_Isex_2</td>
<td>-0.0266</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0137)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>_cons</td>
<td>12.91***</td>
<td>13.00***</td>
<td>13.02***</td>
</tr>
<tr>
<td></td>
<td>(0.00999)</td>
<td>(0.0161)</td>
<td>(0.0175)</td>
</tr>
</tbody>
</table>

N: 10992
R-sq: 0.150
adj. R-sq: 0.150
rmse: 0.712

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001
Regression: correlation matrix

Below is a correlation matrix for all continuous variables in the model. The numbers are Pearson correlation coefficients, which range from -1 to 1: the closer to 1, the stronger the correlation. A negative value indicates an inverse relationship (when a variable one goes up, the other tends to go down).

```
pwcorr y size size2 age age2 nchild nchild2, star(0.001) sig
```

<table>
<thead>
<tr>
<th></th>
<th>y</th>
<th>size</th>
<th>size2</th>
<th>age</th>
<th>age2</th>
<th>nchild</th>
<th>nchild2</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>size</td>
<td>-0.4059*</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>size2</td>
<td>-0.2871*</td>
<td>0.9069*</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>-0.2029*</td>
<td>0.2295*</td>
<td>0.1878*</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>age2</td>
<td>-0.1869*</td>
<td>0.1704*</td>
<td>0.1486*</td>
<td>0.9818*</td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nchild</td>
<td>-0.3872*</td>
<td>0.8374*</td>
<td>0.7368*</td>
<td>0.1859*</td>
<td>0.1161*</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nchild2</td>
<td>-0.2615*</td>
<td>0.7215*</td>
<td>0.7839*</td>
<td>0.1530*</td>
<td>0.1105*</td>
<td>0.8627*</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
</tr>
</tbody>
</table>
Regression: graph matrix command produces a graphical representation of the correlation matrix by presenting a series of scatter plots for all variables. Type:

graph matrix y size size2 age age2 nchild nchild2, half maxis(ylabel(none) xlabel(none))
There seems to be a curvilinear relationship between nchild and $y$. To capture this, we can add a squared variable, in this case $nchild^2$. 
Regression: searching for a relationship

In a bivariate context (with only one independent variable), a scatter plot is sufficient to show the explanatory power of an independent variable.

To assess the contribution of a dependent variable in a multivariate context, other methods are needed. One is to draw a scatter plot to show the relationship between two series of residuals:

1. e1: come from regressing \( x_i \) on the other independent variables: provides evidence for a distinct explanatory power for \( x_i \)
2. e2: come from regressing \( y \) on the other independent variables (without the \( x_i \)): shows what cannot be explained by the other independent variables

\[
\text{qui } xi \text{ regress } y \text{ nchild nchild2 i.area i.inst_level size size2 i.sex age}
\]

\[
\text{avplot age}
\]
Partial residual plot

When performing a linear regression with a single independent variable, a scatter plot of the response variable against the independent variable provides a good indication of the nature of the relationship.

If there is more than one independent variable, things become more complicated. Although it can still be useful to generate scatter plots of the response variable against each of the independent variables, this does not take into account the effect of the other independent variables in the model.

Partial residual plots attempt to show the relationship between a given independent variable and the response variable given that other independent variables are also in the model. Partial residual plots are formed as:

\[ \text{Residuals} + B_i X_i \text{ versus } X_i \]

where
\[ \text{Res} = \text{residuals from the full model} \]
\[ B_i = \text{regression coefficient from the } i\text{th independent variable in the full model} \]
\[ X_i = \text{the } i\text{th independent variable} \]
Partial residual plot

\texttt{xr: regress y nchild i.area i.inst_lev size i.sex}
\texttt{cprplot nchild}
Regression: functional form/linearity

The augmented partial residual (APR) plot is a graphical display of regression diagnosis. The APR plot is the plot of

$$ e_{y|x} x_j^2 + x_j b_j + x_j^2 b_{jj} \text{ against } x_j $$

where $b_j$ and $b_{jj}$ are the least squares estimates from model

$$ y = X \beta + \beta_j x_j + \beta_{jj} x_j^2 + \varepsilon. $$

This plot was suggested by Mallows (1986) to explore whether or not a transformation of $x_j$ is needed in the linear multiple regression model.

*Note* that the APR plot does not just intend to detect the need for a quadratic term in the regression model. The introduced quadratic term is really a truncated version of the potential nonlinear form of $x_j$. 
Regression: functional form/linearity

The command acprplot (augmented component-plus-residual plot) provides another graphical way to examine the relationship between variables. It does provide a good testing for linearity. Run this command after running a regression

```
regress y nchild age /* Notice we do not include nchild2 */
acprplot nchild, lowess
acprplot age, lowess
```

Form more details see [http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter2/statareg2.htm](http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter2/statareg2.htm), and/or type help acprplot and help lowess.
Regression: getting predicted values

How good the model is will depend on how well it predicts $Y$, the linearity of the model and the behavior of the residuals.

There are two ways to generate the predicted values of $Y$(usually called $Y_{hat}$) given the model:

**Option A**, using `generate` after running the regression:

```
generate y_hat = _b[_cons] + _b[age]*age + _b[nchild]*nchild + ...```

**Option B**, using `predict` immediately after running the regression:

```
predict y_hat
label variable y_hat "y predicted"```
Regression: observed vs. predicted values

For a quick assessment of the model run a scatter plot

twoway (scatter y y_hat) (lowess y y_hat) (line y_hat y_hat), legend(order( 1 "Scatter" 2 "Smoothed link " 3 "45 line"))

We should expect a 45 degree pattern in the data. Y-axis is the observed data and x-axis the predicted data (Yhat).
Regression: testing for homoskedasticity

An important assumption is that the variance in the residuals has to be homoskedastic or constant.

Residuals cannot varied with values of X (i.e. fitted values of Y since Y=Xb).

A definition: “The error term [e] is homoskedastic if the variance of the conditional distribution of [ei] given Xi [var(ei|Xi)], is constant for i=1...n, and in particular does not depend on x; otherwise, the error term is heteroskedastic”

When plotting residuals vs. predicted values (Yhat) we should not observe any pattern at all.

In Stata we do this using `rvfplot` right after running the regression, it will automatically draw a scatter plot between residuals and predicted values. `rvfplot, yline(0)` Residuals seem to slightly expand at higher levels of Yhat.
Regression: testing for homoskedasticity

rvfplot, yline(0)
Regression: testing for heteroskedasticity

Non-graphical way to detect heteroskedasticity is the Breusch-Pagan test. The null hypothesis is that residuals are homoskedastic.

```
estat hettest
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
 Ho: Constant variance
 Variables: fitted values of y

    chi2(1) = 210.62
    Prob > chi2 = 0.0000
```

In the example below we reject the null at 95% and concluded that residuals are not homogeneous. The graphical and the Breush-Pagan test suggest the presence of heteroskedasticity in our model. The problem with this is that we may have the wrong estimates of the standard errors for the coefficients and therefore their t-values.
Regression: testing for heteroskedasticity

There are two ways to deal with heteroskedasticity problem:

1. Using heteroskedasticity-robust standard errors,
2. Using weighted least squares. WLS requires knowledge of the conditional variance on which the weights are based, if this is known (rarely the case) then use WLS.

In practice it is recommended to use heteroskedasticity-robust standard errors to deal with heteroskedasticity.

By default Stata assumes homoskedastic standard errors, so if we need to adjust our model to account for heteroskedasticity, we have to use the option `robust` in the `regress` command.
Regression: Misspecification of the functional form

How do we know if we have included all variables we need to explain \( Y \)?

In Stata we can test for the misspecification form with the test with the ovtest command

```
. qui xi: regress  y nchild nchild2 i.area  i.inst_lev  size size2 i.sex. robust
. ovtest
```

Ramsey RESET test using powers of the fitted values of \( y \)
Ho: model has no omitted variables
\[
F(3, 10979) = 28.69 \\
\text{Prob} > F = 0.0000
\]

The null hypothesis is that the model is well specified and do not requires the quadratic or higher powers links, the p-value is lowest than the usual threshold of 0.05 (95% significance), so we reject the null and conclude that we need more refinements in the specification of the model.

It tests if the \( \gamma \) parameters of the following model equals to zero.

\[
y = XB + \gamma_1 \hat{y}^2 + \gamma_2 \hat{y}^3 + \gamma_3 \hat{y}^4 + e
\]
Regression: specification error

Another command to test model specification is linktest. It basically checks whether we need more variables in our model by running a new regression with the observed \( y \) against \( y_{\text{hat}} \) (y_predicted or \( X\beta \)) and \( y_{\text{hat}}^2 \) as independent variables.

The thing to look for here is the significance of \( \_\text{hatsq} \). The null hypothesis is that there is no specification error. If the p-value of \( \_\text{hatsq} \) is significant then we reject the null and conclude that our model is not correctly specified.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>KS</th>
<th>Number of obs = 10992</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2887.2455</td>
<td>2</td>
<td>1443.62275</td>
<td>F(2,10989) = 4333.25</td>
</tr>
<tr>
<td>Residual</td>
<td>3660.98542</td>
<td>10989</td>
<td>333150006</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>6548.23092</td>
<td>10991</td>
<td>595781178</td>
<td>R-squared = 0.4409</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adj R-squared = 0.4408</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Root MSE = 0.57719</td>
</tr>
</tbody>
</table>

| y         | Coef.  | Std. Err. | t     | P>|t|    | [95% Conf. Interval] |
|-----------|--------|-----------|-------|--------|---------------------|
| \_hat     | \(-0.3610154\) | \0.4065249\) | \(-0.89\) | \0.375\) | \(-1.157877\) | \(0.4358465\) |
| \_hatsq   | \0.0536618\) | \0.0160228\) | 3.35  | \0.001\) | \0.022543\) | \0.0850693\) |
| \_cons    | \8.615592\) | \2.576098\) | 3.34  | \0.001\) | \3.565976\) | \13.66521\) |
Regression: multicollinearity

An important assumption for the multiple regression model is that independent variables are not perfectly multicollinear. One regressor should not be a linear function of another.

When multicollinearity is present standard errors may be inflated. Stata will drop one of the variables to avoid a division by zero in the OLS procedure.

The Stata command to check for multicollinearity is \texttt{vif} (variance inflation factor). Right after running the regression type:

\begin{verbatim}
    . qui xi: regress y nchild nchild2 i.area i.inst_lev size size2 i.sex, robust

    . vif
\end{verbatim}

\begin{table}
\centering
\begin{tabular}{l|cc}
\hline
Variable & VIF & 1/VIF \\
\hline
    size  & 14.19 & 0.070470 \\
   size2  & 11.31 & 0.088430 \\
   nchild & 10.17 & 0.098298 \\
  nchild2 &  7.79 & 0.128449 \\
_iinst_lev_3 &   1.67 & 0.597459 \\
_iinst_lev_2 &   1.53 & 0.652887 \\
_iinst_lev_4 &   1.32 & 0.758734 \\
    _Iarea_1 &   1.15 & 0.868132 \\
     _Isex_2 &   1.07 & 0.936215 \\
\hline
  Mean VIF &  5.58 &       \\
\end{tabular}
\end{table}

A \texttt{vif}> 10 or a \texttt{1/vif}< 0.10 indicates trouble.
\texttt{VIF}_i=1/(1-R^2_i): where \textit{R}^2_i is the coefficient of multiple determination of regression produced by regressing the variable \textit{x}_i against the other independent variables except \textit{x}_i.
Some Definitions

**Residual:** The difference between the predicted value (based on the regression equation) and the actual, observed value.

**Outlier:** In linear regression, an outlier is an observation with large residual. In other words, it is an observation whose dependent-variable value is unusual given its values on the predictor variables. An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem.

**Leverage:** An observation with an extreme value on a predictor variable is a point with high leverage. Leverage is a measure of how far an independent variable deviates from its mean. These leverage points can have an effect on the estimate of regression coefficients.

**Influence:** An observation is said to be influential if removing the observation substantially changes the estimate of coefficients. Influence can be thought of as the product of leverage and outlieriness.
Regression: summary of influence indicators

**Leverage**
To be redefined.

After running the regression type:

`predict lev, leverage`
Regression: summary of influence indicators

**DfFit**
Measures how much an observation influences the regression model as a whole. How much the predicted values change as a result of including and excluding a particular observation.

High influence if $|\text{DfFIT}| > 2\sqrt{k/N}$
Where $k$ is the number of parameters (including the intercept) and $N$ is the sample size.

After running the regression type:
`predict dfits if e(sample), dfits`

To generate the flag for the cutoff type:
`gen cutoffdfit = abs(dfits) > 2*sqrt((e(df_m) +1)/e(N)) & e(sample)`
Regression: summary of influence indicators

**DfBeta**
Measures the influence of each observation on the coefficient of a particular independent variable (for example, x1). An observation is influential if it has a significant effect on the coefficient.

A case is an influential outlier if $|\text{DfBeta}| > \frac{2}{\sqrt{N}}$
Where N is the sample size.

Example:

```plaintext
predict dfbeta_nchild, dfbeta(nchild)
```

/* To estimate the dfbetas for all predictors just type: */
dfbeta
/* To flag the cutoff */
gen **cutoffdfbeta** = abs(dfbeta_nchild) > 2/sqrt(e(N))
Regression: summary of influence indicators

**Cook’s distance**

Measures how much an observation influences the overall model or predicted values.

It is a summary measure of leverage and high residuals.

High influence if $D > 4/N$

Where $N$ is the sample size.

If $D > 1$, this indicates big outlier problem

In Stata after running the regression type:
predict D, cooksd
Regression: summary of influence indicators

Leverage

Measures how much an observation influences regression coefficients.

High influence if leverage $h > 2k/N$

Where $k$ is the number of parameters (including the intercept) and $N$ is the sample size.

A rule-of-thumb: Leverage goes from 0 to 1. A value closer to 1 or over 0.5 may indicate problems.

In Stata after running the regression type: predict lev, leverage
Regression: testing for normality

Another assumption of the regression model (OLS) that impact the validity of all tests (p, t and F) is that residuals behave ‘normal’.

If residuals do not follow a ‘normal’ pattern then you should check for omitted variables, model specification, linearity, functional forms. In sum, you may need to reassess your model/theory. In practice normality does not represent much of a problem when dealing with really big samples.
Regression: testing for normality

Standardize normal probability plot (pnorm) checks for non-normality in the middle range of residuals.
Quintile-normal plots (qnorm) check for non-normality in the extremes of the data (tails). It plots quintiles of residuals vs quintiles of a normal distribution. Tails are a bit off the normal.

A non-graphical test is the Shapiro-Wilk test for normality. It tests the hypothesis that the distribution is normal, in this case the null hypothesis is that the distribution of the residuals is normal. Type

```
swilk res
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Shapiro-Wilk W test for normal data</th>
</tr>
</thead>
<tbody>
<tr>
<td>res</td>
<td>Obs 10992 0.98879 60.532 11.012 0.00000</td>
</tr>
</tbody>
</table>

The null hypothesis is that the distribution of the residuals is normal, here the p-value is 0.00 we reject the null (at 95%). Some users prefer to assess normality visually rather than by statistical > testing.
Regression: joint test (F-test)

To test whether two coefficients or more are jointly different from 0 use the command: test.

```
. qui xi: regress  y nchild nchild2 i.area  i.inst_lev  size size2 i.sex age
. test nchild==nchild2 +size2

( 1)  nchild - nchild2 - size2 = 0

    F(  1, 10981) =  21.22
    Prob > F =    0.0000
```

The p-value is 0.0000, we reject the null and conclude that:

\[_b(Nchild) \neq _b(nchild2) + _b(size2)\]
Some references

Regression diagnostics: A checklist
http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter2/statareg2.htm

Logistic regression diagnostics: A checklist

Times series diagnostics: A checklist (pdf)

Times series: dfueller test for unit roots (for R and Stata)

Panel data tests: heteroskedasticity and autocorrelation
http://www.stata.com/support/faqs/stat/panel.html
http://www.stata.com/support/faqs/stat/xt.html
http://dss.princeton.edu/online_help/analysis/panel.htm

Data Analysis: Annotated Output
http://www.ats.ucla.edu/stat/AnnotatedOutput/default.htm

Data Analysis Examples
http://www.ats.ucla.edu/stat/dae/

Regression with Stata
http://www.ats.ucla.edu/STAT/stata/webbooks/reg/default.htm

Regression
http://www.ats.ucla.edu/stat/stata/topics/regression.htm

How to interpret dummy variables in a regression

How to create dummies
http://www.stata.com/support/faqs/data/dummy.html
http://www.ats.ucla.edu/stat/stata/faq/dummy.htm

Logit output: what are the odds ratios?
http://www.ats.ucla.edu/stat/stata/library/odds_ratio_logistic.htm
Topics in Statistics (links)
What statistical analysis should I use?
http://www.ats.ucla.edu/stat/mult_pkg/whatstat/default.htm
Statnotes: Topics in Multivariate Analysis, by G. David Garson
http://www2.chass.ncsu.edu/garson/pa765/statnote.htm
Elementary Concepts in Statistics
http://www.statsoft.com/textbook/stathome.html
Introductory Statistics: Concepts, Models, and Applications
http://www.indiana.edu/~statmath/stat/all/ttest/
Statistical Data Analysis
http://math.nicholls.edu/badie/statdataanalysis.html
Stata Library. Graph Examples (some may not work with STATA 10)
http://www.ats.ucla.edu/STAT/stata/library/GraphExamples/default.htm
Comparing Group Means: The T-test and One-way ANOVA Using STATA, SAS, and SPSS
http://www.indiana.edu/~statmath/stat/all/ttest/

Useful links / Recommended books
• DSS Online Training Section http://dss.princeton.edu/training/
• UCLA Resources to learn and use STATA http://www.ats.ucla.edu/stat/stata/
• DSS help-sheets for STATA http://dss/online_help/stats_packages/stata/stata.htm
• Introduction to Stata (PDF), Christopher F. Baum, Boston College, USA. “A 67-page description of Stata, its key features and benefits, and other useful information.” http://fmwww.bc.edu/GStat/docs/StataIntro.pdf
• STATA FAQ website http://stata.com/support/faqs/

Books
• Unifying Political Methodology: The Likelihood Theory of Statistical Inference/ Gary King, Cambridge University Press, 1989
• Statistical Analysis: an interdisciplinary introduction to univariate & multivariate methods / Sam Kachigan, New York : Radius Press, c1986
• Statistics with Stata (updated for version 9) /Lawrence Hamilton, Thomson Books/Cole, 2006