☺A

OLS MULTIVARIATE: PARTIAL EFFECTS, MULTICOLLINEARITY, SPECIFICATION TESTS

Maria Elena Bontempi e.bontempi@economia.unife.it

Roberto Golinelli roberto.golinelli@unibo.it

this version: 26/09/2007[§]

1. An introduction to multivariate regression and the Frisch-Waugh-Lovell theorem

To better understand main topics about multivariate regressions, we start with a very simple database, scolari.dta. The aim of the related exercise is to measure the effect of a number of potential determinants on the scholars' reading ability at the primary education level. Therefore, in our example reading ability (labelled as *y*) is the dependent variable of the model.

Obs	individual identifier of the i-th pupil (i = 1, 2, ..., 10); therefore, N = 10
F	"female" dummy variable, f=1 if the scholar gender is female, f=0 if is male
Eta	age in years
Y	reading ability; the higher the quotient, the better the scholar's performance
Taglia	shoe size
m	= 1 – f

The simple regression model (i.e. with only one explanatory variable) is $y_i = \alpha + \beta x_i + \varepsilon_i$. The OLS estimate of the slope parameter is:

$$\hat{\beta} = \frac{C\hat{O}V(Y,X)}{V\hat{A}R(X)}$$

Associated to each estimate, $\hat{\beta}$, you have the standard error of the estimator, $s_{\hat{\beta}} = \frac{\sqrt{s^2}}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{X})^2}}$.

The multiple regression model (i.e. with more explanatory variables) is $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$, where, for simplicity, we consider only two explanatory variables. The OLS estimate of the first of the two "slope" parameters is:

$$\hat{\beta}_1 = \frac{C\hat{O}V(Y,X_1) - \hat{\beta}_2 C\hat{O}V(X_1,X_2)}{V\hat{A}R(X_1)} = \frac{C\hat{O}V(Y,X_1)V\hat{A}R(X_2) - C\hat{O}V(Y,X_2)C\hat{O}V(X_1,X_2)}{V\hat{A}R(X_1)V\hat{A}R(X_2) - \left[C\hat{O}V(X_1,X_2)\right]^2}$$

(symmetrically, we can define the OLS estimate of $\hat{\beta}_2$).

Associated to each estimate, $\hat{\beta}_1$ or $\hat{\beta}_2$, you have the standard error of the estimator. For example:

$$s_{\hat{\beta}_1} = \frac{\sqrt{s^2}}{\sqrt{\sum_{i=1}^{N}\left(x_{1i} - \overline{X}_1\right)^2\left(1 - \hat{\rho}^2_{X_1X_2}\right)}} = \frac{\sqrt{s^2}}{\sqrt{\sum_{i=1}^{N}\left(x_{1i} - \overline{X}_1\right)^2\left(1 - R^2_{X_1}\right)}},$$

where $\hat{\rho}^2_{X_1X_2}$ is the squared correlation coefficient between X_1 and X_2 which corresponds to the R-squared from regressing X_1 on all the other explanatory variables (and including an intercept).

[§] Very preliminary. Comments welcome.

Note that in any equation with more than one independent variable, the coefficients represents the change in the dependent variable Y caused by a one-unit increase in the kth independent variable X_k, holding constant the other independent variables in the equation (*ceteris paribus* condition). In the present context, β_1 is a partial regression coefficient measuring the amount of Y change due to a unit change in X_1 given X_2, i.e. holding X_2 constant.
The multiple regression model OLS estimate of β_1 parameter is equal to the corresponding simple regression model OLS estimate of β: $\hat{\beta}_1 = C\hat{O}V\,(Y, X_1) \,/\, V\hat{A}R\,(X_1)$ only if $C\hat{O}V\,(X_1, X_2)=0$, i.e. if the two explanatory variables of the multiple regression are uncorrelated.
However, if X_1 and X_2 are correlated, the β_1 estimate will depend, in addition to $C\hat{O}V\,(X_1, X_2)$, also on $C\hat{O}V\,(Y, X_2)$ e $V\hat{A}R\,(X_2)$. The same is true for the β_2 estimate.

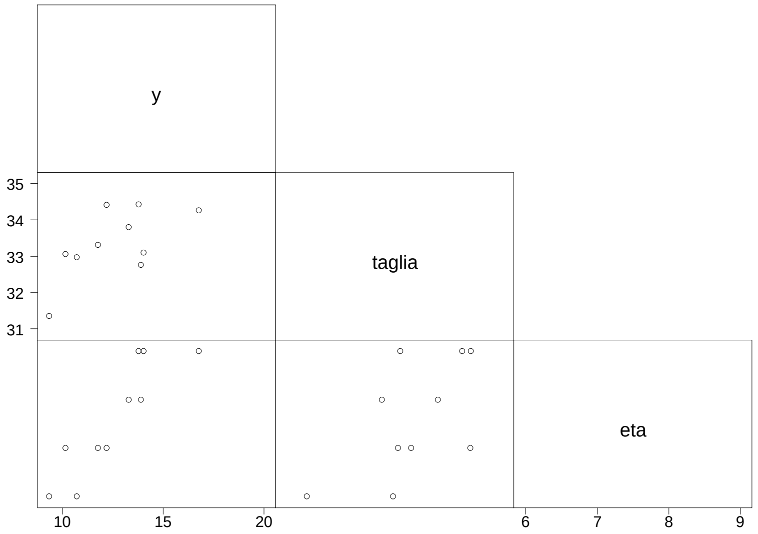
What happens if one estimates with OLS the simple (and stupid) model: $y_i = \alpha + \beta_1\, taglia_i + \varepsilon_i$?
With this model, we assume that the probably irrelevant regressor X_1 (*taglia*, shoe size) explains the reading ability, while we incorrectly exclude X_2 (*eta*, the age in years), probably the most relevant explanatory variable of the reading ability.

```
reg y taglia

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =    5.25
       Model |  17.4998438     1  17.4998438           Prob > F      =  0.0511
    Residual |  26.6498106     8  3.33122633           R-squared     =  0.3964
-------------+------------------------------           Adj R-squared =  0.3209
       Total |  44.1496545     9  4.90551716           Root MSE      =  1.8252

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      taglia |   1.485849   .6482757     2.29   0.051    -.0090772    2.980776
       _cons |  -36.95083   21.62551    -1.71   0.126    -86.81934    12.91767
------------------------------------------------------------------------------
```

The effect of the shoe size on the reading ability is surprisingly positive and quite significant. What does this mean? We need a new pedagogical interpretation? Hint: never ex post justify (i.e. on the basis of the results obtained) your findings. Instead, try to understand what could explain the result.

To do so, a useful graph is the scatterplot matrix. It is a relevant diagnostic tool for detecting non-linearity and outliers, and for revealing information about the joint distributions of the variables that would not appear from the univariate distributions. Remember that in the multiple regression, *all* the correlation between variables is important.

```
graph7 y taglia eta, matrix half label
```



From previous plots, *y* and *taglia*, and *y* and *eta* are positively related; but the main point here is that also *taglia* and *eta* are positively related (when you are a child, feet grow with time). Look at what happens if you include both shoe size and age as explanatory variables of the reading ability model with the Stata command:

```
reg y taglia eta

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  2,     7) =   14.79
       Model |  35.7021003     2  17.8510501           Prob > F      =  0.0031
    Residual |   8.4475542     7  1.20679346           R-squared     =  0.8087
-------------+------------------------------           Adj R-squared =  0.7540
       Total |  44.1496545     9  4.90551716           Root MSE      =  1.0985
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      taglia |   .4130543   .4780688     0.86   0.416    -.7173987    1.543507
         eta |   1.484458   .3822275     3.88   0.006      .580634    2.388283
       _cons |  -12.45863   14.46338    -0.86   0.418     -46.6591    21.74183
------------------------------------------------------------------------------
```

The outcome is strongly different to that we obtained with the stupid model: age exerts a positive and very significant effect on the scholars’ reading ability, while the effect of the shoe size on the reading ability is no longer significant.

In general, we can state that the omission of a relevant explanatory variable of the model (such as *eta*) leads to biased estimates of the effect of the included explanatory variable (such as *taglia*) if the omitted variable is correlated with the included variable. In our case, the omission of the age (*eta*) leads to the over-estimation of the relevance of the shoe size (*taglia*) effect on the reading ability.

The omitted variable bias is given by $\frac{C\hat{O}V(Y,X_1)}{V\hat{A}R(X_1)}-\frac{C\hat{O}V(Y,X_1)-\hat{\beta}_2 C\hat{O}V(X_1,X_2)}{V\hat{A}R(X_1)}=\frac{\hat{\beta}_2 C\hat{O}V(X_1,X_2)}{V\hat{A}R(X_1)}$.

Its sign can be summarised as:

	$C\hat{O}V(X_1,X_2)>0$	$C\hat{O}V(X_1,X_2)<0$
$\hat{\beta}_2>0$	Positive (upward bias)	Negative (downward bias)
$\hat{\beta}_2<0$	Negative (downward bias)	Positive (upward bias)

Be careful during the model specification phase, because of the risk of omitting model's relevant explanatory variables.

As said above, in a regression model with more explanatory variables, the β_1 parameter measures the effect of X_1 on Y given the level of X_2, i.e. accounting for the effect that, at the same time, X_2 exerts on Y and on X_1. The working of such "statistical adjustment" implicit in the multivariate model can be better understood through a sequence of partial simple regressions aiming at filtering out the effect that X_2 exerts on both Y and X_1 (this is the so called Frisch-Waugh-Lovell theorem[1]).

Filter-out the effect of X_2 (*eta*) on Y with the following partial regression:

```
reg y eta
```

```
      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =   29.78
       Model |  34.8012212     1  34.8012212           Prob > F      =  0.0006
    Residual |  9.34843325     8  1.16855416           R-squared     =  0.7883
-------------+------------------------------           Adj R-squared =  0.7618
       Total |  44.1496545     9  4.90551716           Root MSE      =   1.081

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         eta |   1.675276   .3069825     5.46   0.001     .9673733    2.383179
       _cons |  -.1348808   2.357977    -0.06   0.956    -5.572386    5.302624
------------------------------------------------------------------------------
```

and save the filtered values of Y (i.e. the residuals of previous partial regression)

```
predict ydepx2, resid
```

Symmetrically, filter-out the effect of X_2 (*eta*) on X_1 (*taglia*) with the following partial regression:

```
reg taglia eta
```

```
      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =    4.01
       Model |   2.6463364     1   2.6463364           Prob > F      =  0.0802
    Residual |  5.28022182     8  .660027727           R-squared     =  0.3339
-------------+------------------------------           Adj R-squared =  0.2506
       Total |  7.92655822     9  .880728691           Root MSE      =  .81242

------------------------------------------------------------------------------
      taglia |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         eta |   .4619678    .230712     2.00   0.080    -.0700551    .9939907
       _cons |   29.83566   1.772133    16.84   0.000     25.74912    33.92221
------------------------------------------------------------------------------
```

and, as before, save the filtered values of X_1 (i.e. the residuals of previous partial regression)

```
predict x1depx2, resid
```

[1] Frisch R. and Waugh F. V. (1933) Partial Time Regressions as Compared with Individual Trends, Econometrica, 1, 387-401; Lovell M. C. (1963) Seasonal Adjustment of Economic Time Series, Journal of the American Statistical Association, 58, 993-1010.

Finally, run the simple regression using the filtered variables. This leads to the same estimation result for the effect exerted by shoe size on the reading ability (measured by the β_1 parameter) as we obtained above using the multivariate model.

```
. reg ydepx2 x1depx2

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =    0.85
       Model |  .900879032     1  .900879032           Prob > F      =  0.3827
    Residual |  8.44755426     8  1.05594428           R-squared     =  0.0964
-------------+------------------------------           Adj R-squared = -0.0166
       Total |  9.34843329     9  1.03871481           Root MSE      =  1.0276
------------------------------------------------------------------------------
      ydepx2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     x1depx2 |   .4130543   .4471924     0.92   0.383    -.6181732    1.444282
       _cons |  -7.57e-10    .324953    -0.00   1.000    -.7493429    .7493429
------------------------------------------------------------------------------
```

(a question for you: why the OLS estimate of the constant in the regression with filtered data is so close to zero?)

In this regression with filtered values, the effect of the age (though not explicitly listed in the model's regressors) is accounted for thanks to the filtering procedure of Y e X_1 values from the X_2 effects that we accomplished through the partial regressions. Hence, the reported outcome supports the idea that the effect of the shoe size (*taglia*, i.e. X_1) on the reading ability (Y) is not significant if we appropriately use a model that accounts for the probably most relevant explanatory variable: the student's age (*eta*, i.e. X_2).

2. A comparison between simple and multiple regression

Now we know enough to go back to the `Urban.dta` case, and to further assess the determinants of the homicide rate in USA using more explanatory variable.

In the lecture_OLS_bivariate we used a model with only one explanatory variable, i.e. we ran the simple regression:

```
reg homic poor
      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  1,    18) =    6.14
       Model |  181.370325     1  181.370325           Prob > F      =  0.0233
    Residual |  531.573154    18  29.5318419           R-squared     =  0.2544
-------------+------------------------------           Adj R-squared =  0.2130
       Total |  712.943479    19   37.523341           Root MSE      =  5.4343

------------------------------------------------------------------------------
       homic |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        poor |   .9438495   .3808596     2.48   0.023     .1436932    1.744006
       _cons |  -.8151891   3.344025    -0.24   0.810    -7.840726    6.210348
------------------------------------------------------------------------------
```

The inclusion of a further explanatory variable measuring the population (*pop*) effect on homicide rate leads to the following model of multiple regression:

```
reg homic poor pop

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  2,    17) =   36.30
       Model |  577.667334     2  288.833667           Prob > F      =  0.0000
    Residual |  135.276144    17  7.95742026           R-squared     =  0.8103
-------------+------------------------------           Adj R-squared =  0.7879
       Total |  712.943479    19   37.523341           Root MSE      =  2.8209

------------------------------------------------------------------------------
       homic |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        poor |   .6562371   .2018567     3.25   0.005     .2303567    1.082117
         pop |   .0222329   .0031504     7.06   0.000      .015586    .0288797
       _cons |  -3.899011   1.790002    -2.18   0.044    -7.675584   -.1224378
------------------------------------------------------------------------------
```

The 0.656 estimate is the amount of *homic* change due to a unit change in *poor* (+1% of families below the poverty line) holding *pop* constant, remember the *ceteris paribus* condition. In the same way, 0.022 is the estimate of the amount of *homic* change due to a unit change in *pop* (one thousand more city population) holding *poor* constant. Apparently, the two OLS estimates are very different: *poor* seems to exert a stronger effect on *homic* than that of *pop*. However, in the linear regression model, estimates can differ strongly, depending on the average level of the corresponding explanatory variable (*poor* is measured as a %, *pop* is in thousands). In order to make comparable the strength of a coefficient to the coefficient for another variable we can add the option `beta` to the `regress` command.

```
. reg homic poor pop, beta

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  2,    17) =   36.30
       Model |  577.667334     2  288.833667           Prob > F      =  0.0000
    Residual |  135.276144    17  7.95742026           R-squared     =  0.8103
-------------+------------------------------           Adj R-squared =  0.7879
       Total |  712.943479    19   37.523341           Root MSE      =  2.8209
------------------------------------------------------------------------------
       homic |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
        poor |   .6562371   .2018567     3.25   0.005                 .3506821
         pop |   .0222329   .0031504     7.06   0.000                 .7612375
       _cons |  -3.899011   1.790002    -2.18   0.044                        .
------------------------------------------------------------------------------
```

This option gives the standardised regression coefficients (coefficients are measured in standard deviations instead of in the units of the variables). In this way, they may be compared because they measure in standard deviations the change of the dependent variable due to the increase in the explanatory variable. The formal relationship between usual and standardised estimates is:

$\tilde{\beta}_k = \hat{\beta}_k \dfrac{s_{x_k}}{s_y}, k = 1,..., K-1$, where $\tilde{\beta}_k$ is the standardised regression coefficient for the k^{th} explanatory variable (the constant term standardised coefficient is not computed), $\hat{\beta}_k$ is the usual OLS coefficient, s_{x_k} is the standard deviation of x_k, and s_y is the standard deviation of y.

Standardised regression coefficients usually fall within the –1/+1 range. In our case, the amplitude of the two effects is reversed. Of course, standardised coefficients make comparable the effects (parameter estimates) belonging to the same multiple regression.

Comparing simple and multiple regression model's OLS estimates, we note that the coefficient estimate associated to *poor* falls from 0.94 to 0.65. Given the significance of the *pop* effect on *homic*, the *poor* estimate obtained from the simple regression model is upwards biased because the simple model omits a relevant explanatory variable (*pop*).

The upwards bias of the *poor* parameter's estimate in the simple regression model is due to the fact that the included variable *poor* (the share of families below the poverty line) and the omitted variable *pop* are positively correlated. In the simple regression model, the included *poor* explanatory also captures the effect of the omitted *pop* on *homic* because in larger cities (with high *pop* values) both *homic* and *poor* are higher. We will further deepen this point in Section 4.

2.1. F-test to verify if the overall model is significant

In lecture_OLS_bivariate we introduced the t-test for the significance of one parameter. Now we focus on the F-test for the joint significance of more than one parameter. In the present context (two explanatory variables), the null hypothesis to be tested is $H_0: \beta_1=\beta_2=0$ versus H_1: at least one parameter is $\neq 0$.

With reference to the multivariate regression output above, we have that F(2,17)=36.30, with Prob>F (probability value, or simply P-value) = 0.0000. Since this P-value is smaller than the probability of 5% (the chosen significance level), the risk of a type I error (reject the true H_0) is very low and smaller than 5%; for this, the null hypothesis is rejected: the two parameters are jointly different to zero in the population.

The F-test of the regression helps assessing whether our regression model has some interesting features. In fact, under H_0 , none of the included regressors is useful to explain the dependent variable. Under the null hypothesis ($\beta_1=\beta_2=0$) the restricted model is $y_i = \alpha + \varepsilon_i$ (a constant term plus the unpredictable noise); the not rejection of the null implies that any model's regressor is able to predict, at least in part, the behaviour of the dependent variable.

The actual calculation of the F-test can be articulated in the following steps.

(1) Estimate the unrestricted model (see the multiple regression above) and save the sum of squares of the residuals (RSS_{NV}= 135.276144).

(2) Estimate the restricted model (with only the intercept) and save the sum of squares of the residuals (RSS_V= 712.943479).

```
. reg homic

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  0,    19) =    0.00
       Model |           0     0           .           Prob > F      =       .
    Residual |  712.943479    19   37.523341           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared =  0.0000
       Total |  712.943479    19   37.523341           Root MSE      =  6.1256

------------------------------------------------------------------------------
       homic |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |     6.9055   1.369732     5.04   0.000     4.038617    9.772383
------------------------------------------------------------------------------
```

(remember that in this model the OLS estimate of the intercept = 6.9055 is the sample mean of Y)

(3) Obtain the test statistic $\hat{F} = \frac{(RSS_V - RSS_{NV})/V}{RSS_{NV}/(N-K)} = \frac{RSS_V - RSS_{NV}}{RSS_{NV}} \times \frac{N-K}{V}$, distributed as an F(V, N – K), where V is the number of restricted parameters (V=K-1=2)[2], N is the number of observations (N=20), K is the number of estimated parameters in the unrestricted model (K=3).

In our example, we have a F(2, 17) test statistic equal to:

$\frac{712.943479 - 135.276144}{135.276144} \times \frac{20-3}{2}$ = 36.2974 (as in the upper panel of the regression output).

In order to have the critical value at 5% (i.e. 0.05) of the F distribution we can use the command:

```
. display invFtail(2,17,.05)
3.5915306
```

In order to have the P-value of a given F test statistic we can use the command:

```
. display Ftail(2,17,36.2974)
7.318e-07
```

3. Multicollinearity

Multicollinearity refers to any linear relationship amongst explanatory variables in a multiple regression model. It can affect two or more of them. The original definition referred to an exact linear relationship, but later it was extended to mean a nearly perfect relationship. The correlation can be negative or positive.

In the presence of multicollinearity, the estimate of one explanatory variable's impact on the dependent variable (while controlling for the others) tends to be less precise than if regressors were uncorrelated with one another. In other terms, if the correlation coefficient between X_1 and X_2 is close to ±1, OLS estimator can hardly distinguish the effect of X_1 on Y (measured by β_1) from the corresponding effect measured by β_2. Multicollinearity may either depend on the nature of the data or on the issue analysed (see e.g. the Cobb-Douglas production function).

Simply put, if nominally "different" measures (model’s regressors) actually quantify the same phenomenon (doesn't need to be same phenomenon - just presence of correlation is enough - positive or negative. Think about positive or negative interaction between variables.) to a significant degree -- i.e., wherein the variables are accorded different names and perhaps employ different numeric measurement scales but correlate highly with each other -- they are redundant.

A principal danger of such data redundancy is that of overfitting in regression analysis models. The best regression models are those in which the explanatory variables each correlate highly with the dependent variable but correlate at most only minimally with each other.

The following Norman Swanson’s sentence about multicollinearity can summarise all previous points:

[2] Of course, the constant term must be excluded in this computation.

“Many economic variables have the property that they are correlated. This is not surprising, given the natural links between almost all facets of economic activity within any given economy. However, this feature of most economic data suggests that within the context of regression, not only are the regressors (or independent or explanatory variables) related to the dependent variable in a regression model (which is what we want, as we're trying to "explain" our dependent variable using our independent variables), but the independent variables are also correlated with one another. When the independent variables are correlated with one another (and this can be checked by simply running a regression of one of the independent variables on another, say, and checking whether the multiple coefficient of determination from this regression is high), then we have what is termed "multicollinearity". If the multicollinearity is severe (i.e. if the R squared value from the simple regression just mentioned is close to unity), then the precision of the estimated slope coefficients in the model is very poor. As precision is simply variance, high multicollinearity implies high slope estimator variance. This in turn implies low t-statistics, since the denominator of the standard t-statistic. In fact, H_0: $\beta = 0$ is assessed on the basis of $t = \frac{\hat{\beta}}{s_\beta}$ (where s_β is the standard error of the estimator of the slope coefficient). These low t-statistics in turn may result in a failure to reject the null hypothesis that some particular slope coefficient is zero, in turn implies that the particular independent variable is not a useful explanatory variable. However, the variable may actually have a lot of explanatory power, and we may simply be fooled into believing the variable is irrelevant because we observe low t-statistics which are simply an artefact of the multicollinearity in our regression model. This problem is often signalled when our regression has a high R^2 value, but very low slope coefficient t-statistics.”

One simple remedy to multicollinearity is to omit one of the variables that are highly multicollinear, as the informational content of this variable is essentially the same as that of other variable(s), anyway.[3]

It should be noted that if multicollinearity is particularly severe, then OLS regression may not even work, and you may get an error message from your software when attempting to run OLS; therefore, in order to diagnose this problem, simply discard one of your regressors and attempt again to run the regression. Differently, and more pragmatically, in the context of perfect collinearity Stata automatically arbitrarily drops one or more regressors till the explanatory variables matrix is no longer singular, and runs a restricted model.
Referring to this, Appendix A1 below analyses the classical case of the dummy variable (collinearity) trap.
Qualitative explanatory variables (i.e. the use of the dummies) are particularly useful for the Chow parameter constancy test (see lecture_nonlinearity_Chow), and in the context of panel models specification.

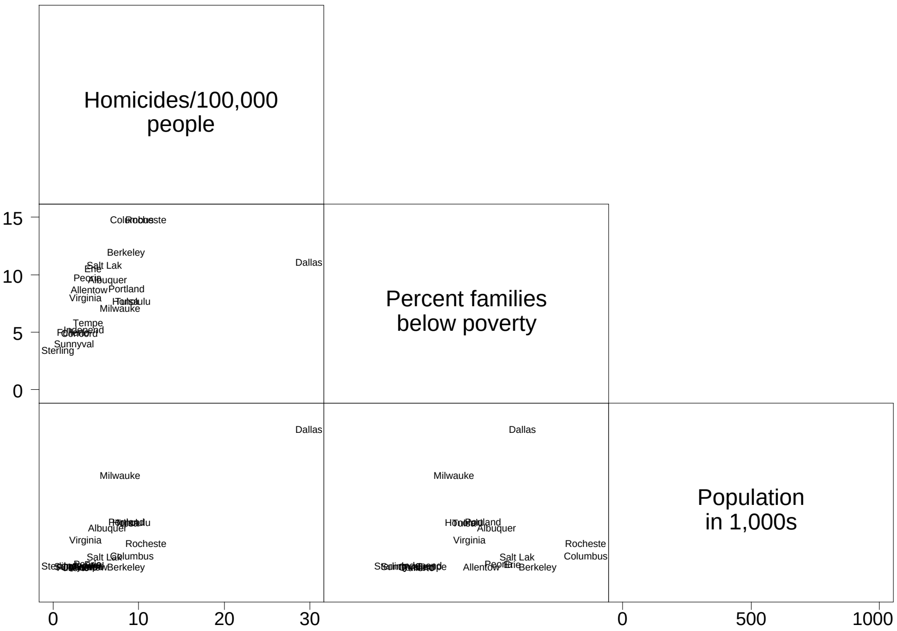
Are our previous regressions about the homicide rate in USA affected by multicollinearity? From the Swanson’s words, the symptoms of multicollinearity are: *(a)* Large changes in the estimated regression coefficients when a predictor variable is added or deleted. *(b)* Non significant t-statistics of X_1 and X_2 parameter estimates, while the corresponding joint F is significant (i.e. t and F tests apparently are not coherent).

In addition, the formal detection-tolerance or the variation inflation factor (VIF) can be useful ways, though with caveats, to assess for the presence of multicollinearity.

[3] Another common solution (but only with time series observations) is to difference or log difference the data. This often removes much of the multicollinearity among regressors, particularly since the multicollinearity may have arisen because the regressors were all trending upwards or downwards over time; this problem will be discussed in the lecture on time-series and nonstationarity.

Before inspecting regression results, collinearity problems in multiple regression models can be detected by using the scatterplot matrix and the correlation matrix. Correlation coefficient between two regressors greater than 0.8 might entail multicollinearity problems.

```
graph7 homic poor pop, matrix half label s([city])
```



The positive (low) correlation between *poor* and *pop* is due to the inclusion of Dallas among the sample observations.

```
. corr homic poor pop
(obs=20)

             |    homic     poor      pop
-------------+---------------------------
       homic |   1.0000
        poor |   0.5044   1.0000
         pop |   0.8320   0.2019   1.0000
```

Note that the correlation between *poor* and *pop* is reduced if Dallas is eliminated from the analysis.

```
corr poor pop if city!="Dallas"
(obs=19)

             |     poor      pop
-------------+------------------
        poor |   1.0000
         pop |   0.0946   1.0000
```

```
. pwcorr homic poor pop, star(.05)

             |    homic     poor      pop
-------------+---------------------------
       homic |   1.0000
        poor |   0.5044*  1.0000
         pop |   0.8320*  0.2019   1.0000
```

Differently from `corr, pwcorr` displays a star for the correlations significantly different from zero at the 5% (or 1% or 10%) level. `pwcorr` also gives the number of observations used in the correlation (option `obs`). Finally, `corr` and `pwcorr` differ in the way in which missing data is handled. The command `corr` drops an observation if any variable has a missing value; i.e. it uses list-wise or case-wise deletion. `pwcorr` uses pair-wise deletion; i.e. the observation is dropped only if there is a missing value for the pair of variables being correlated.

There are formal *criteria* that can be used to detect whether the collinearity is harmful. However, which one to apply in any particular case will depend on the nature of the problem. As suggested by Leamer[4], in general all the criteria are merely complains that things are not ideal. Therefore, the standard errors and the coherence between t and F test outcomes probably offer more information about how serious multicollinearity is. In any case, in what follow, we will see how to obtain in Stata a number of alternative multicollinearity formal-detectors.

```
. qui reg homic poor pop
```

```
. vif
```

```
    Variable |       VIF       1/VIF
-------------+----------------------
        poor |      1.04    0.959236
         pop |      1.04    0.959236
-------------+----------------------
    Mean VIF |      1.04
```

The `vif` command after `regress` compute the variance inflation factor[5],

VIF($\hat{\beta}_k$)= $\dfrac{1}{1-R_k^2}$

Tolerance = 1/VIF

where R_k^2 is the R-squared from regressing X_k on all the other explanatory variables (and including an intercept). Remember that, from the theorem of Frisch-Waugh-Lovell, the variance of the estimator of the k^{th} parameter can be expressed as $V\hat{A}R(\hat{\beta}_k)=\dfrac{s^2}{\sum_{i=1}^{N}\left(x_{ki}-\overline{X}_k\right)^2\left(1-R_k^2\right)}=\dfrac{s^2}{TSS_k\left(1-R_k^2\right)}$,

where TSS_k indicates the total sample variation in X_k, and R_k^2 is the R-squared from regressing X_k on all the other explanatory variables (and including an intercept).

Therefore, the estimator variance can be high because the residuals' variance is high (in the numerator above), or because the explanatory variable variability is low and because of multicollinearity (high R-squared); the latter two cases affect the denominator above.

However, even if a low R_k^2 is better, a high R_k^2 is neither necessary nor sufficient to get multicollinearity causing high standard errors and making difficult to disentangle the separate effects of each of the explanatory variables on the dependent variable. As a rule of thumb, a variable with a VIF values greater than 10 and a tolerance value lower than 0.1 suggests further investigation because it may be considered as an almost linear combination of other explanatory variables.

[4] Leamer E. E. (1983) Model Choice and Specification Analysis, in Z. Griliches and M. D. Intrilligator (eds.) Handbook of Econometrics, vol. 1, Amsterdam: North-Holland pp. 286-330.

[5] William A. R. and Watts D. G. (1978) Meaningful Multicollinearity Measures, Technometrics, 20, 407-412. Farrar D. E. and Glauber R. R. (1967) Multicollinearity in Regression Analysis: the Problem Revisited, Review of Economics and Statistics, 49, 92-107.

The procedure `collin.ado`, written by Philip B. Ender, UCLA Department of Education, does not require `regress` before, and computes several collinearity diagnostic measures: VIF, tolerance, eigenvalues, condition index, and R-squared.

```
. collin poor pop

  Collinearity Diagnostics

                          SQRT                                Cond        R-
  Variable        VIF     VIF     Tolerance   Eigenval        Index     Squared
-------------------------------------------------------------------------
      poor       1.04     1.02     0.9592      1.2019      1.0000      0.0408
       pop       1.04     1.02     0.9592      0.7981      1.2272      0.0408
-------------------------------------------------------------------------
  Mean VIF       1.04               Condition Number      1.2272
                 Determinant of correlation matrix       0.9592
```

The condition number[6] measures the closeness to singularity of the matrix (**X'X**) i.e. how much explanatory variables are near to be almost exactly linearly dependent. It is defined as the square root of the ratio of the largest to the smallest eigenvalue of the matrix (**X'X**) and it measures the sensitivity of the regression estimates to small changes in the data.[7] As a rule of thumb, conditions index greater than 30 indicate strong collinearity.

The high R-squared of auxiliary regressions (of each of the explanatory variables upon the other K-1 regressors) indicates a near-linear dependence among the column of **X**, but does not allow the separation of the interrelationships among the covariates.

Some solutions to multicollinearity are dropping variables, adding information (for example, using extraneous estimates or getting more data), using variables transformations (such as ratios or first differences), using purely statistical devices like ridge regressions and principal components.[8]

In conclusion, the correlation between *poor* and *pop* is not so high to create multicollinearity problems. Moreover, also note from previous regression results that t and F tests are coherent.

4. Summary findings about the determinants of the homicide rate in USA

From previous results using alternative models to explain the USA homicide rate, we noted that though multicollinearity does not seem to affect our output results, the positive correlation between *poor* and *pop*, both displaying a positive and significant effect on *homic*, can explain an upward bias in the estimated parameter of *poor* in the simple bivariate regression model.

All these results may be nested in a larger experiment involving the use of other variables included in **`urban.dta`**. In particular we will report estimation results of alternative models for *homic* (from simpler to more complex), and we will interpret the results in the light of the two issues of the

[6] Raduchel W. J. (1971) Multicollinearity Once Again, Paper 205 Harvard Institute of Economic Research, Cambridge, Mass. Belsley D. A., Kuh E., Welsch R. E. (1980) Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, New York: Wiley.

[7] On eigenvalues see Maddala G. S. (1992) Introduction to Econometrics, Maxwell Macmillan, pp. 298-303.

[8] Discussions on measures to detect multicollinearity and on solutions are: Maddala G. S. (1977) Econometrics, McGraw Hill, pp. 183-194; Maddala G. S. (1992) Introduction to Econometrics, Maxwell Macmillan, pp. 269-294; Judge G. G., Griffiths W. E., Carter Hill R., Lee Tsoung-Chao (1980) The Theory and Practice of Econometrics, Wiley, pp. 452-497 (see also the second edition, 1985, by Judge G. G., Griffiths W. E., Carter Hill R., Lütkepohl H., Lee Tsoung-Chao, pp. 896-930).

specification errors in choosing the explanatory variables, and of the multicollinearity among the explanatory variables. In doing so we will also introduce a very convenient way to tabulate regression results with Stata.

Quoting the A.H. Studenmund textbook ("Using Econometrics: a Practical Guide", Addison Wesley) Table 6.1, we can summarise the consequences of the omitted variable and the included irrelevant variable cases:

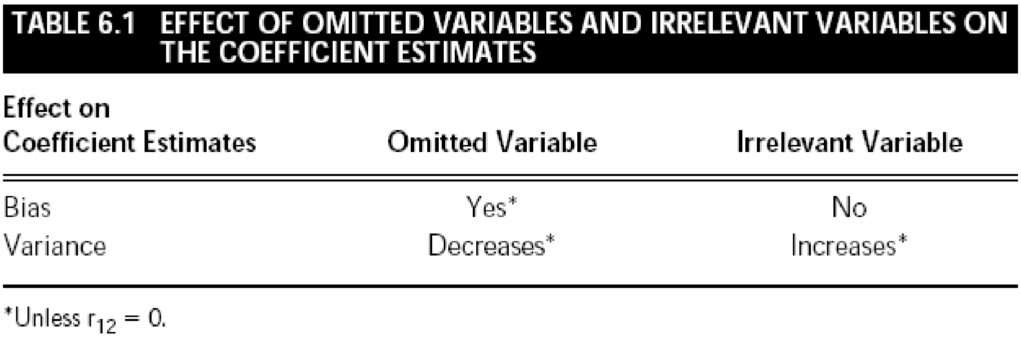


TABLE 6.1 EFFECT OF OMITTED VARIABLES AND IRRELEVANT VARIABLES ON THE COEFFICIENT ESTIMATES

Effect on Coefficient Estimates	Omitted Variable	Irrelevant Variable
Bias	Yes*	No
Variance	Decreases*	Increases*

*Unless $r_{12} = 0$.

(note: "r_{12}" is the correlation coefficient between X_1 and X_2 explanatory variables)

The major consequence of omitting a relevant independent variable from an equation is to cause bias in the regression coefficients that remain in the equation, unless model's omitted and included variables are not correlated (indeed, a very rare event).
However, it could be that, worried by the risk of omitting relevant variables, the researcher exaggerates by adding too many explanatory variables. In such circumstance, the addition of a variable to an equation where it does not belong does not cause bias, but it does increase the variances of the estimated coefficients of the included variables (unbiased but also inefficient estimates). Remembering the standard error of the estimator formula and given that $1-R^2_{X_1}$<1, it is clear why the inclusion of an additional explanatory variable to the regression implies an increase in the variance of the parameter estimator (less precision).

It is worth noting that the omitted variable model has more independent variables in the "true" (but unknown) model than in the estimated equation, while the irrelevant variable model has more independent variables in the estimated equation than in the true one. The latter point can explain the inefficiency affecting estimates of the model that includes irrelevant variables.

In the light of previous estimation results and analyses, we can assume that the "true" model for *homic* is that with two explanatory variables (*poor* and *pop*). Therefore, the simple regression model (with only *poor*) is an omitted variable model. In addition we also estimate a model where, beyond *poor* and *pop,* additional variables are included: *divorce* (divorces with ages between 15-59) and *educ* (median years education). The resulting model probably includes two irrelevant variables for the problem of interest:

```
reg homic poor pop divorce educ

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  4,    15) =   18.78
       Model |  594.266611     4  148.566653           Prob > F      =  0.0000
    Residual |  118.676868    15  7.91179119           R-squared     =  0.8335
-------------+------------------------------           Adj R-squared =  0.7892
       Total |  712.943479    19   37.523341           Root MSE      =  2.8128

------------------------------------------------------------------------------
       homic |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        poor |    .648916   .2013865     3.22   0.006     .2196709    1.078161
         pop |   .0224275   .0032685     6.86   0.000     .0154608    .0293942
     divorce |   .2278492   .2718316     0.84   0.415    -.3515461    .8072444
        educ |   .7632583   .7996625     0.95   0.355     -.941182    2.467699
       _cons |  -15.94979   10.30833    -1.55   0.143    -37.92147    6.021887
------------------------------------------------------------------------------
```

Finally, we also estimate a model where, beyond *poor* and *pop*, we also add the *inequal* (household inequality index) regressor. Given that also *poor* is similarly defined as an inequality index, we suspect that *inequal* and *pop* be quite correlated. In fact:

```
. pwcorr  poor inequal, star(.05)

             |     poor  inequal
-------------+------------------
        poor |   1.0000 
     inequal |   0.7101*  1.0000 
```

The correlation close to 0.8 lead to the assumption that the following estimate be affected by multicollinearity problems:

```
. reg homic poor pop  inequal

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  3,    16) =   24.02
       Model |  583.417226     3  194.472409           Prob > F      =  0.0000
    Residual |  129.526253    16   8.0953908           R-squared     =  0.8183
-------------+------------------------------           Adj R-squared =  0.7843
       Total |  712.943479    19   37.523341           Root MSE      =  2.8452

------------------------------------------------------------------------------
       homic |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        poor |    .490285   .2832435     1.73   0.103    -.1101644    1.090734
         pop |   .0216214   .0032594     6.63   0.000     .0147118    .0285311
     inequal |    16.1482   19.16078     0.84   0.412    -24.47084    56.76725
       _cons |  -8.851294   6.147281    -1.44   0.169    -21.88295     4.18036
------------------------------------------------------------------------------
```

In order to allow better comparisons between the four model estimates, we can give this sequence of Stata commands to obtain a convenient tabulation of the alternative estimation results.
(alternatively use the corresponding provedure: **`TABLE_summary_Urban.do`**)

```
qui reg homic poor
est store omit
qui reg homic poor pop
est store true
qui reg homic poor pop  divorce educ
```

```
est store irrel
qui reg homic poor pop   inequal
est store collin
qui reg homic pop   inequal
est store truealt
est table omit true irrel collin truealt,    /*
*/       b(%6.3f) se(%6.3f) p(%6.4f) stats(N df_r df_m r2 r2_a rmse F)

----------------------------------------------------------------
    Variable |  omit      true      irrel    collin    truealt
-------------+--------------------------------------------------
        poor |   0.944     0.656     0.649     0.490
             |   0.381     0.202     0.201     0.283
             |  0.0233    0.0047    0.0057    0.1027
         pop |             0.022     0.022     0.022     0.022
             |             0.003     0.003     0.003     0.003
             |            0.0000    0.0000    0.0000    0.0000
     divorce |                       0.228
             |                       0.272
             |                      0.4151
        educ |                       0.763
             |                       0.800
             |                      0.3550
     inequal |                                16.148    39.206
             |                                19.161    14.559
             |                                0.4118    0.0154
       _cons |  -0.815    -3.899   -15.950    -8.851   -14.046
             |   3.344     1.790    10.308     6.147     5.671
             |  0.8102    0.0438    0.1426    0.1692    0.0241
-------------+--------------------------------------------------
           N |  20.000    20.000    20.000    20.000    20.000
        df_r |  18.000    17.000    15.000    16.000    17.000
        df_m |   1.000     2.000     4.000     3.000     2.000
          r2 |   0.254     0.810     0.834     0.818     0.784
        r2_a |   0.213     0.788     0.789     0.784     0.759
        rmse |   5.434     2.821     2.813     2.845     3.008
           F |   6.142    36.297    18.778    24.023    30.907
----------------------------------------------------------------
                                             legend: b/se/p-value
```

The interpretation of the results in the light of the scheme depicted by Table 6.1 is quite straightforward, and is left as exercise for the reader. In particular, try to explain the results in the latter column, where the almost-ever significant poor parameter has been omitted.

It is just worth stressing that Table 6.1 predictions are completely met by Stata results.

5. Model's diagnostics: heteroskedasticity, outliers and leverage, robust estimators.

In this Section we deal with one of the most important issues of the econometrics practice: verifying that data we are analysing meet the assumptions underlying OLS estimator optimal (BLUE) statistical properties: unbiasedness, consistency and efficiency. If not, results can be misleading.

A number of misspecification tests and analyses deal with the assessment of the validity of the specification assumptions of the classical regression model:

- *Valid model specification*: the model should include all relevant variables, otherwise parameter estimates are biased, and it should exclude irrelevant variables, otherwise parameter estimators are not efficient; already tackled in the previous Sections 1, 2, and 4

- *Linearity and constancy* of the model's parameters; see also lecture_nonlinearity_Chow
- *Normality test*. H_0: errors are normally distributed. This is necessary to test for the significance of estimated parameters; in this Section
- *Homoskedasticity test*. H_0: the error variance is constant among different observations; introduced in this Section, the heteroskedasticity issue will be further exploited in lecture_GLS
- *Exogeneity test*. H_0: errors are not correlated with the explanatory variables; see lecture_IV

Further, other issues, though are not assumptions of OLS, can arise in EDA and can affect the estimation results:

- *Multicollinearity*: if the explanatory variables linearly related, we have problems in estimating the regression coefficients and in evaluating the results significance; in previous Section 3
- *Influential observations*: individual observations that exert a great influence on the coefficients and could be outliers (in lecture_exploratory_data_analysis and in this Section).

5.1. Analysis of the residuals' distribution

Let's start by storing of the *homic*'s "true" model residuals (*res*):

```
qui reg homic poor pop
predict homichat
predict res, res
summ homic homichat res

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+-----------------------------------------------------
       homic |      20      6.9055    6.12563        .55      29.98
    homichat |      20      6.9055   5.513941   .5587064   23.28908
         res |      20    2.09e-08   2.668295  -6.877952   6.690916
```

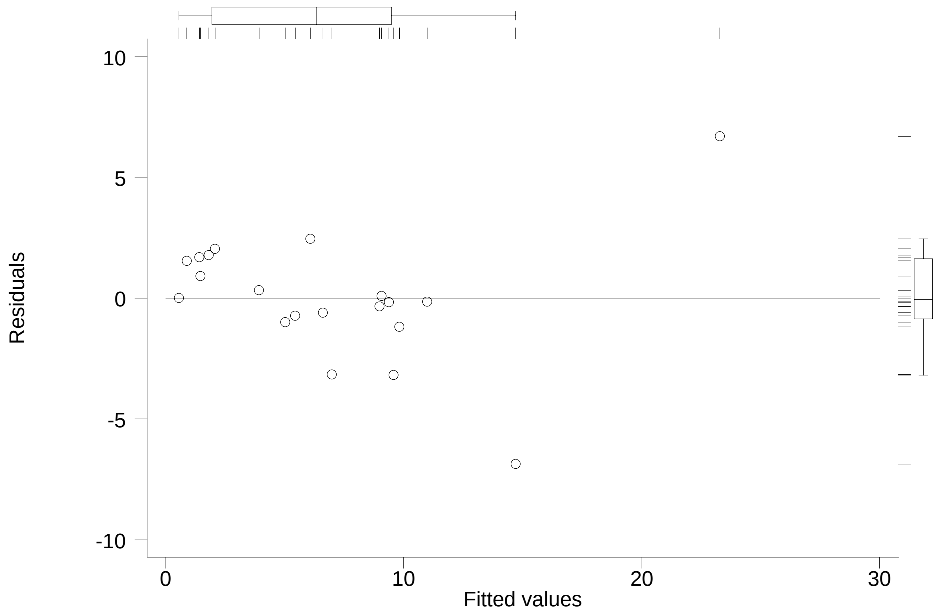
Note that *res* average is zero because actual (*homic*) and fitted (*homichat*) values have the same average.

In general, the analysis of the regression residuals is the main diagnostic tool of the model specification. Misspecification tests, based on regression residuals, must be accomplished before to test for the significance of model's parameters. They are based on residuals because *res* are the estimate of the model's (unobservable) errors.

Ideally, if the regression line summarises all patterns to be found in a scatterplot, residuals should show just random variations around this line. If some non-random pattern remains in the residuals, perhaps either the initial model specification or the estimation techniques need to be adjusted.

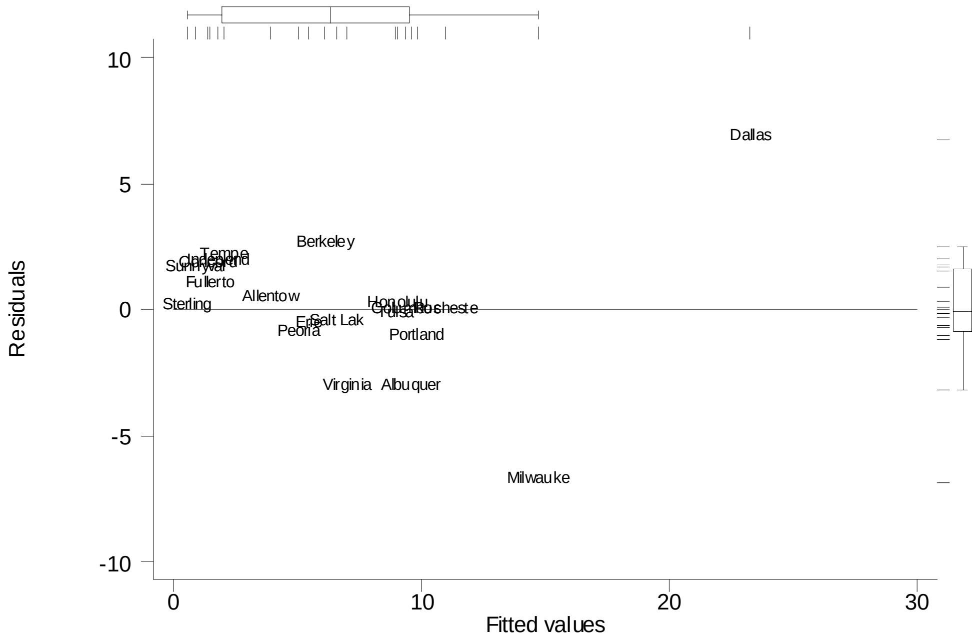
The pattern of the residuals may be depicted by the residual versus fitted graph. Since, by definition, model's residuals embody all the unexplainable factors of the dependent variable, the visual inspection of their pattern may be useful to check for weaknesses of model's explanatory ability.

```
.version 7: rvfplot ,oneway twoway box yline(0) ylabel xlabel
```



(the zero line indicates that the mean of the disturbances must be zero)

```
version 7: rvfplot ,oneway twoway box yline(0) ylabel xlabel s([city])
```



It is evident that Dallas induces a positive skew in the residuals. Dallas is also characterised by a high fit and a high positive residual: our regression prediction is most dramatically wrong in the case of Dallas (model's under-prediction). On the contrary, many cities have negative residuals (model's over-predictions): Dallas' high homicide rate pulls the regression line up. All these facts may indicate the presence of either influential observations or curvilinearity.

Moreover, residuals' dispersion above/below the zero line increases with the level of the fit: this varying spread might indicate heteroskedasticity problems.

Whatever is the true problem, we can not accept the previous multiple regression as final.

```
summ res, d

                          Residuals
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -6.877952      -6.877952
 5%    -5.034405      -3.190859
10%    -3.182166      -3.173472       Obs                  20
25%    -.8717402      -1.205127       Sum of Wgt.          20

50%    -.0794083                      Mean          -2.19e-08
                        Largest       Std. Dev.      2.668295
75%     1.612288       1.777815
90%     2.240921        2.03746       Variance       7.119797
95%     4.567649       2.444383       Skewness      -.1587214
99%     6.690915       6.690915       Kurtosis       4.948724

sktest res

                   Skewness/Kurtosis tests for Normality
                                                 ------- joint ------
    Variable |  Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------
         res |      0.722         0.032             4.76       0.0926
```

At the 5% significance level, the null hypothesis of normal errors is not rejected (P-value of the overall normality is 9.3%). Normality is useful in small samples (such ours) where we cannot refer to the asymptotic theory: in fact, thanks to normality, test statistics have fairly standard distributions (t, F, and chi-squared). Note also that the Dallas outlier induce a bit of kurtosis (P-value = 3.2%). However, the null of symmetric errors is not rejected (P-value = 72.2%).

Overall, we can suspect heteroskedastic errors and/or influential observations.

5.2. Residuals' heteroskedasticity test

The main assumption about the classical regression model errors is that they are identically and independently distributed with mean equal to zero, in symbols: $\varepsilon \sim iid(0, \sigma^2)$.

The $E(\varepsilon) = 0$ assumption is perfectly represented by OLS residuals that always sum to zero by definition, provided that model's specification includes the intercept (see `summ res` command output above).

The assumption of *independently distributed* errors (errors belonging to different observation are not related each other) is not easily checked in cross-sections, given that there is not an obvious way in which they have to be ordered (listed). In this context, an appropriate sampling design (random sampling) may prevent the insurgence of the problem. On the other side, the assessment of errors being independently distributed is crucial in time series.

With cross-section data (characterised by relevant variability) errors' heteroskedasticity is the most common problem: often the errors variance seems not constant over different observations. In this case, the assumption of identically distributed errors is no longer valid.

In general, the validity of the assumptions about the classical regression model errors is assessed by using the so called misspecification (or diagnostic) tests. They are:

(a) based on the model's residuals';

(b) under the null hypothesis there is the validity of the corresponding error specification assumption (the alternative hypothesis definition is more problematic);

(c) they must be accomplished before the parameters' significance tests. In fact, such tests can be run only after we are sure that model's errors are *well behaved*.

The (misspecification) heteroskedasticity test null hypothesis is that σ^2 (the errors variance) is the same for all the observations. What about the alternative hypothesis, H_1? Non-constant variance implies that specific variance behaviours must be assumed. A first (very simple) assumption is that the errors variance is linked to the pattern of the explanatory variables (`, rhs` option). To do so, we can use the following command that runs the Breusch-Pagan (1979) test[9]:

```
hettest, rhs

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

         chi2(2)      =    28.80
         Prob > chi2  =   0.0000

hettest

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

         chi2(1)      =    21.13
         Prob > chi2  =   0.0000
```

The null is rejected: the impression arising from the visual inspection above of residuals is formally supported.

With heteroskedastic errors, the OLS estimator

(a) looses the efficiency property (i.e. OLS are no longer the best estimation approach, while GLS are);

(b) the standard error of the regression is a biased estimate of the true errors variance. Therefore, parameter inference through t and F tests is no longer valid.

These two problems may be tackled in two alternative ways: the use of the GLS estimator (sub point (a) above), or the use of an estimator of the errors variance that is robust to the heteroskedasticity problem (sub point (b) above).

The `robust` option of the `reg` command specifies that the Eicker/Huber/White sandwich estimator[10] of variance is to be used instead of the traditional (and biased) OLS standard errors estimates.

In this case, the term "robust regression" indicates "regression with robust standard errors". In fact, the estimates of the regression coefficients are the same as those from the standard OLS linear regression, but the estimates of the standard errors are robust to failures to meet assumptions concerning normality and homogeneity of the residuals' variance.

[9] Breusch, T. and A. Pagan. "A Simple Test for Heteroskedasticity and Random Coefficient Variation." Econometrica, 47, 1979, 1287-1294.

[10] Eicker F. (1967) "Limit Theorems for Regressions with Unequal and Dependent Errors", in L. Le Cam and J. Neyman (eds.) Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, 1, 59-82; Huber P. J. (1973) "Robust Regression: Asymptotics, Conjectures, and Monte Carlo", The Annals of Statistics, 1, 799-821; White, H. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.", Econometrica, 48, 1980, 817-838.

```
reg homic poor pop, robust

Regression with robust standard errors                 Number of obs =      20
                                                       F(  2,    17) =   24.36
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.8103
                                                       Root MSE      =  2.8209

------------------------------------------------------------------------------
             |               Robust
       homic |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        poor |   .6562371   .1328392     4.94   0.000     .3759708    .9365034
         pop |   .0222329    .006886     3.23   0.005     .0077046    .0367611
       _cons |  -3.899011   1.484662    -2.63   0.018    -7.031375   -.7666471
------------------------------------------------------------------------------
```

Previous outcomes with traditional OLS standard errors are not reversed: both the parameters of interest are significantly different to zero.

5.3. REgression Specification Error Test, RESET

The RESET test (due to Ramsey, 1969)[11] is a general test of the model's specification. Sometimes, RESET test is defined as a linearity test because it is obtained by an auxiliary regression in which residuals are regressed against model's explanatory variables and growing powers of the fitted values. In this way, the significance of the powers in explaining the residuals is assessed. The null rejection below may suggest the inclusion of non-linear effect in the model's specification but, anyway, it probably arises from the presence of outlier observations (and residuals).

```
ovtest

Ramsey RESET test using powers of the fitted values of homic
       Ho:  model has no omitted variables
                  F(3, 14) =     17.40
                  Prob > F =      0.0001
```

5.4. Outliers

In matrix form, we know that $\hat{Y} = X\hat{\beta} = P_X Y = X(X'X)^{-1}X'Y$, where the hat or projection matrix, $P_X \equiv X(X'X)^{-1}X'$, projects the vector y upon the space spanned by the columns of X.

Expanding the vector and matrices, we have:

[11] Ramsey J. B. (1969) "Tests for Specification Errors in Classical Linear Least Squares Regression Analysis" Journal of the Royal Statistical Society, Series B, 31, 350-371.

$$\begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_i \\ \vdots \\ \hat{y}_N \end{bmatrix} = \begin{bmatrix} \frac{X_1^2}{\sum X_i^2} & \cdots & \frac{X_1 X_i}{\sum X_i^2} & \cdots & \frac{X_1 X_N}{\sum X_i^2} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ \frac{X_i X_1}{\sum X_i^2} & \cdots & \frac{X_i^2}{\sum X_i^2} & \cdots & \frac{X_i X_N}{\sum X_i^2} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ \frac{X_N X_1}{\sum X_i^2} & \cdots & \cdots & \cdots & \frac{X_N^2}{\sum X_i^2} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_N \end{bmatrix}, \text{ where } X_i = \begin{bmatrix} x_{1i} \\ \vdots \\ x_{ki} \\ \vdots \\ x_{Ki} \end{bmatrix}$$

is the (K×1) vector of explanatory variables.

A single observation that is substantially different from all other observations can be very influential for the regression results. There are three ways an observation can be unusual.

- **Outlier**: in linear regression it is an observation with **a large residual**, i.e. whose dependent variable value is unusual given its value on the explanatory variables. As show in lecture_exploratory_data_analysis, the sample is affected by at least a severe outlier. It can indicate a peculiarity of the sample or it may indicate a data entry error or other problems. This outlier cast doubts about the normality assumption. Such outlier represents a case much different from most of the data. However, since it pulls the mean and inflates the standard deviation, it may not be many standard deviations from the mean (masking). For this, it might influence coefficients without appearing as an outlier in a scatterplot, as well as not all outliers are influential. An outlier is influential only if its deletion substantially changes the estimation results (see below). In multiple regression influential cases are harder to be visualised than in bivariate regression (where the scatterplot can be used).
- **Leverage**: it is an observation with **an extreme value on an explanatory variable**. The leverage of case *i*, h_i , equals the i^{th} diagonal element of the hat or projection matrix, $P_X \equiv X(X'X)^{-1}X'$: $h_i = \frac{X_i^2}{\sum X_i^2} = X_i'(X'X)^{-1}X_i$, with *X* (N×K) matrix and *X'X* (K×K) matrix. Leverage measures how the i^{th} observation is big compared to all the other observations; theoretically it can range from 1/N to 1: magnitude depends on the sample size. The sample mean of each diagonal element is $\overline{h_i} = K/N$. A rule of thumb suggests that a high-leverage case can be identified by $h_i > 2K/N$, i.e. when its weight is twice the average weight of diagonal elements.[12]
- **Influence**: an observation is influential if removing it substantially changes the estimates of coefficients. It can be thought of **as the product of both leverage and outlierness**.

A method to identify influence is Cook's distance or Cook'D[13], which measures influence of the ith case on the model as a whole (i.e. on all K estimated regression coefficients), rather than on a specific coefficient as the measures[14].

[12] Belsley D.A., Kuh E. and Welsch R. E. (1980) Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, New York, Wiley.

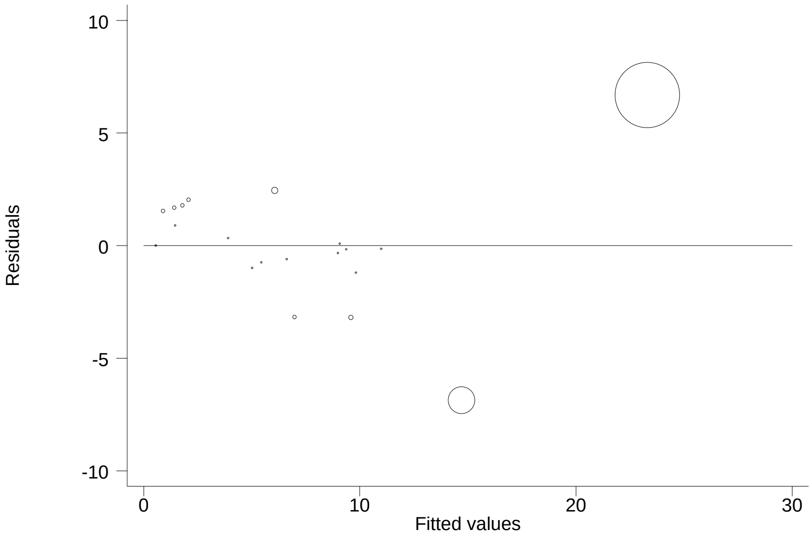
[13] Cook R. D. and Weisberg S. (1982) Residuals and Influence in Regression, New York, Chapman and Hall.

[14] Like the one suggested by suggested by Belsley D. A., Kuh E. and Welsch R. E. (1980) Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, New York, Wiley.

$D_i = \frac{z_i^2 h_i}{K(1-h_i)}$, where $z_i = \frac{\hat{\varepsilon}_i}{s_\varepsilon \sqrt{1-h_i}}$ is a standardised residuals.[15] A rule of thumb suggests that an unusually influential case can be identified by Di > 4/N.

This measure is computed and visualised by the following sequence of Stata commands:

```
qui reg homic poor pop
predict D, cooksd
graph7 res homichat [iweight=D], ylabel xlabel yline(0)
```



As suggested by all previous visual inspections and formal analyses, the biggest ball is Dallas.

The sensitiveness of the OLS estimator with respect to outlier observations bay be tackled by using a parameters' estimator that is robust to outliers. In this case, the term "robust regression" involves the robust estimation of both the regression coefficients and of the standard errors. This approach is useful in situations where the are large outliers and observations with large leverage values. It uses iteratively reweighted least squares to estimate both the regression coefficients and the standard errors. The procedure assigns weights to each of the observations. Those observations with higher leverage or influence receive lower weights.

```
rreg  homic poor pop

   Huber iteration 1:  maximum difference in weights = .07022616
   Huber iteration 2:  maximum difference in weights = .00948742
Biweight iteration 3:  maximum difference in weights = .13951425
Biweight iteration 4:  maximum difference in weights = .00351232

Robust regression estimates                           Number of obs =      19
                                                      F(  2,    16) =   25.59
                                                      Prob > F      =  0.0000
------------------------------------------------------------------------------
       homic |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        poor |   .6131628   .1100947     5.57   0.000     .3797724    .8465532
         pop |   .0098952   .0025097     3.94   0.001     .0045748    .0152157
       _cons |   -1.32179   1.047453    -1.26   0.225    -3.542292    .8987113
------------------------------------------------------------------------------
```

[15] A studentised resisual has a similar formula with standard deviation computed without the ith case.

Again, previous estimates and inferences with classical OLS methods are broadly supported. In particular, the robust outcome is quite similar to what could be obtained with classical OLS run in a sample excluding Dallas:

```
reg  homic poor pop if homic<25

      Source |       SS       df       MS              Number of obs =      19
-------------+------------------------------           F(  2,    16) =   31.77
       Model |  121.812351     2  60.9061756           Prob > F      =  0.0000
    Residual |  30.6758323    16  1.91723952           R-squared     =  0.7988
-------------+------------------------------           Adj R-squared =  0.7737
       Total |  152.488184    18  8.47156576           Root MSE      =  1.3846

------------------------------------------------------------------------------
       homic |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        poor |   .6128833   .0992558     6.17   0.000     .4024704    .8232961
         pop |   .0100326   .0022627     4.43   0.000     .0052359    .0148292
       _cons |  -1.342754   .9443303    -1.42   0.174    -3.344645    .6591368
------------------------------------------------------------------------------
```

Another way to tackle the issue of influential observations is to perform a median regression. Technically, a median regression is just one type of quantile regression, which is also known as either least absolute value (LAV) or minimum absolute deviation (MAD) model.

```
. qreg homic poor pop
Iteration  1:  WLS sum of weighted deviations =    34.9376

Iteration  1: sum of abs. weighted deviations =  34.730578
Iteration  2: sum of abs. weighted deviations =  33.493498
Iteration  3: sum of abs. weighted deviations =  33.185138
Iteration  4: sum of abs. weighted deviations =  32.705456

Median regression                                    Number of obs =        20
  Raw sum of deviations    72.27 (about 4.6999998)
  Min sum of deviations 32.70546                     Pseudo R2     =    0.5475

------------------------------------------------------------------------------
       homic |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        poor |    .574085   .1630467     3.52   0.003     .2300866    .9180835
         pop |   .0183081   .0028915     6.33   0.000     .0122075    .0244087
       _cons |  -2.215627    1.52475    -1.45   0.164    -5.432568    1.001313
------------------------------------------------------------------------------
```

Appendix A1 – A classical case of exact collinearity: *the dummy variable trap*

In the scolari.dta dataset we have, among the others, the following variables.

```
list y eta f m

             y       eta          f          m
  1.  10.71694         6          0          1
  2.  12.19278         7          0          1
  3.  13.29217         8          0          1
  4.  16.77769         9          0          1
  5.  9.344237         6          1          0
  6.  10.14903         7          1          0
  7.  11.76214         7          1          0
  8.  13.90402         8          1          0
  9.  14.03758         9          1          0
 10.  13.79559         9          1          0
```

from which is clear that *f*+*m* = 1.

In fact, if we compute the f-variable frequency distribution we have that:

```
tab f

          F |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |          4       40.00       40.00
          1 |          6       60.00      100.00
------------+-----------------------------------
      Total |         10      100.00
```

In addition, if we regress a dummy variable (such as *f*) against only the constant term and estimate it with OLS, we have the sample average of f, i.e. the female frequency distribution.

```
reg f

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  0,     9) =    0.00
       Model |        0.00     0           .           Prob > F      =       .
    Residual |        2.40     9  .266666667           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared =  0.0000
       Total |        2.40     9  .266666667           Root MSE      =   .5164

------------------------------------------------------------------------------
           f |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |         .6   .1632993     3.67   0.005     .2305913    .9694087
------------------------------------------------------------------------------
```

In order to compute separately the average of the reading ability for male and female students, we can use the command:

```
bysort f: summ y

-> f = 0
------------+--------------------------------------------------------

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           y |         4    13.24489    2.580724   10.71694   16.77769
------------+--------------------------------------------------------
```

```
-> f = 1
-------------+--------------------------------------------------------
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           y |         6    12.16543    2.067495   9.344237   14.03758
-------------+--------------------------------------------------------
```

Let’s suppose that we are interested in the estimation of a very simple model where the reading ability (*Y*) is simply explained by students genre (being qualitative, the genre is measured by *f* and *m* dicotomous dummy variables). This kind of analysis may be conducted with three alternative specifications, that we will label as model A, B and C.

```
Model A:    reg y f

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =    0.54
       Model |  2.79657197     1  2.79657197           Prob > F      =  0.4830
    Residual |  41.3530825     8  5.16913531           R-squared     =  0.0633
-------------+------------------------------           Adj R-squared = -0.0537
       Total |  44.1496545     9  4.90551716           Root MSE      =  2.2736

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           f |  -1.079462   1.467585    -0.74   0.483     -4.46372    2.304796
       _cons |   13.24489   1.136787    11.65   0.000     10.62346    15.86633
------------------------------------------------------------------------------
```

Since, as generally stated, the *_cons* parameter estimate corresponds to the sample average of *Y* when all other regressors are equal to zero, in the case above `13.24489` is the average of male students reading ability because, for a male student observation, *f* is set to zero. Simmetrically, the same reasoning can be conducted with reference to the model B regression:

```
Model B:    reg y m

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =    0.54
       Model |  2.79657197     1  2.79657197           Prob > F      =  0.4830
    Residual |  41.3530825     8  5.16913531           R-squared     =  0.0633
-------------+------------------------------           Adj R-squared = -0.0537
       Total |  44.1496545     9  4.90551716           Root MSE      =  2.2736

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           m |   1.079462   1.467585     0.74   0.483    -2.304796     4.46372
       _cons |   12.16543   .9281824    13.11   0.000     10.02504    14.30583
------------------------------------------------------------------------------
```

In both previous cases, the estimate of the dummy variable (*f* in model A, *m* in model B) measures the difference between the male and female average reading ability. If not significant, the average reading ability is not statistically different for male and female students.

Alternatively, with model C below we explicitly estimate two average reading abilities, by genre.

```
Model C:    reg y f m, noconst
```

```
      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  2,     8) =  153.77
       Model |  1589.69552     2  794.847761           Prob > F      =  0.0000
    Residual |  41.3530825     8  5.16913531           R-squared     =  0.9746
-------------+------------------------------           Adj R-squared =  0.9683
       Total |  1631.04861    10  163.104861           Root MSE      =  2.2736

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           f |   12.16543   .9281824    13.11   0.000     10.02504    14.30583
           m |   13.24489   1.136787    11.65   0.000     10.62346    15.86633
------------------------------------------------------------------------------
```

It is important to note that in model C, the estimation of both the parameters for *f* and *m* explanatory variables is subject to the restriction to zero of the regression constant. Otherwise, we wold incur in the so called dummy variable trap depending on the fact that, as noted above, the sum of f and m is a constant. Put differently, *f* and *m* explanatory variables are exacly collinear with the model intercept. In fact, if we estimate model C without the `nocons` option, we have:

`reg y f m`

```
      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =    0.54
       Model |  2.79657197     1  2.79657197           Prob > F      =  0.4830
    Residual |  41.3530825     8  5.16913531           R-squared     =  0.0633
-------------+------------------------------           Adj R-squared = -0.0537
       Total |  44.1496545     9  4.90551716           Root MSE      =  2.2736
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           f |  -1.079462   1.467585    -0.74   0.483     -4.46372    2.304796
           m |  (dropped)
       _cons |   13.24489   1.136787    11.65   0.000     10.62346    15.86633
------------------------------------------------------------------------------
```

Note that, in this case, Stata dropped *m* regressor from the model specification, leading to the same estimate of model A above.