

**OLS BIVARIATE: ESTIMATES, REGRESSION STANDARD ERROR, t-TEST AND R<sup>2</sup>**

Maria Elena Bontempi [e.bontempi@economia.unife.it](mailto:e.bontempi@economia.unife.it)

Roberto Golinelli [roberto.golinelli@unibo.it](mailto:roberto.golinelli@unibo.it)

this version: 26/09/2007<sup>§</sup>

**1. EDA applied to multivariate data**

The aim of this note is to introduce the simple linear regression, in which we have only an explanatory variable. This variable may be continuous (as in the following case), or it may be dichotomous (0 or 1, as in section 2 of lecture\_OLS\_multivariate).

The data set and the model are the same as in lecture\_exploratory\_data\_analysis.

Before proceeding to estimates, analyse the data and review few concepts.

```
summ homic poor
```

Variable	Obs	Mean	Std. Dev.	Min	Max
homic	20	6.9055	6.12563	.55	29.98
poor	20	8.18	3.273434	3.1	14.5

The concepts at the basis of OLS are covariance and correlation.

Theoretically:

$$COV(Y, X) = E[(Y - \mu_Y)(X - \mu_X)]$$

$$\rho_{Y,X} = \frac{COV(Y, X)}{\sqrt{VAR(Y)VAR(X)}}$$

The corresponding estimators are:

$$C\hat{O}V(Y, X) = \frac{\sum_{i=1}^N (y_i - \bar{Y})(x_i - \bar{X})}{N - 1}$$

$$\hat{\rho}_{Y,X} = \frac{\sum_{i=1}^N (y_i - \bar{Y})(x_i - \bar{X})}{\sqrt{\sum_{i=1}^N (y_i - \bar{Y})^2 \sum_{i=1}^N (x_i - \bar{X})^2}}$$

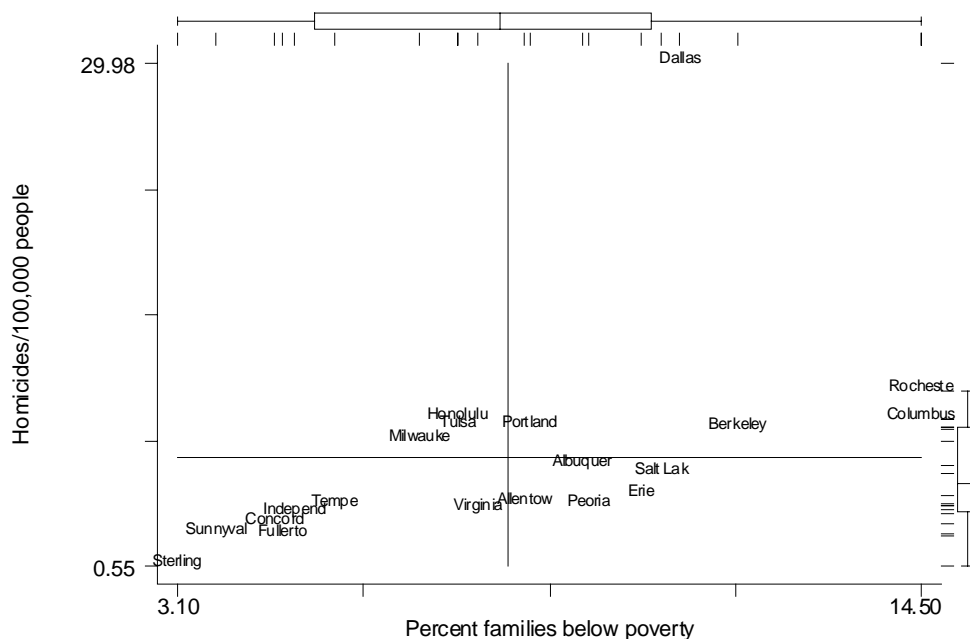
where  $\bar{Y} = \frac{\sum_{i=1}^N y_i}{N}$  (the same for X) and  $V\hat{A}R(Y) = \frac{\sum_{i=1}^N (y_i - \bar{Y})^2}{N - 1}$  (the same for X).

The scatterplot provides a basic tools for visualising the joint distribution of two variables. Usually the dependent variable goes on the y-axis, while the explanatory variable(s) on the x-axis. It is relevant as a diagnostic tool searching for non-linearities and outliers; it is a useful screening tool revealing information in the joint distributions of the variables that would not be apparent from

<sup>§</sup> Very preliminary. Comments welcome.

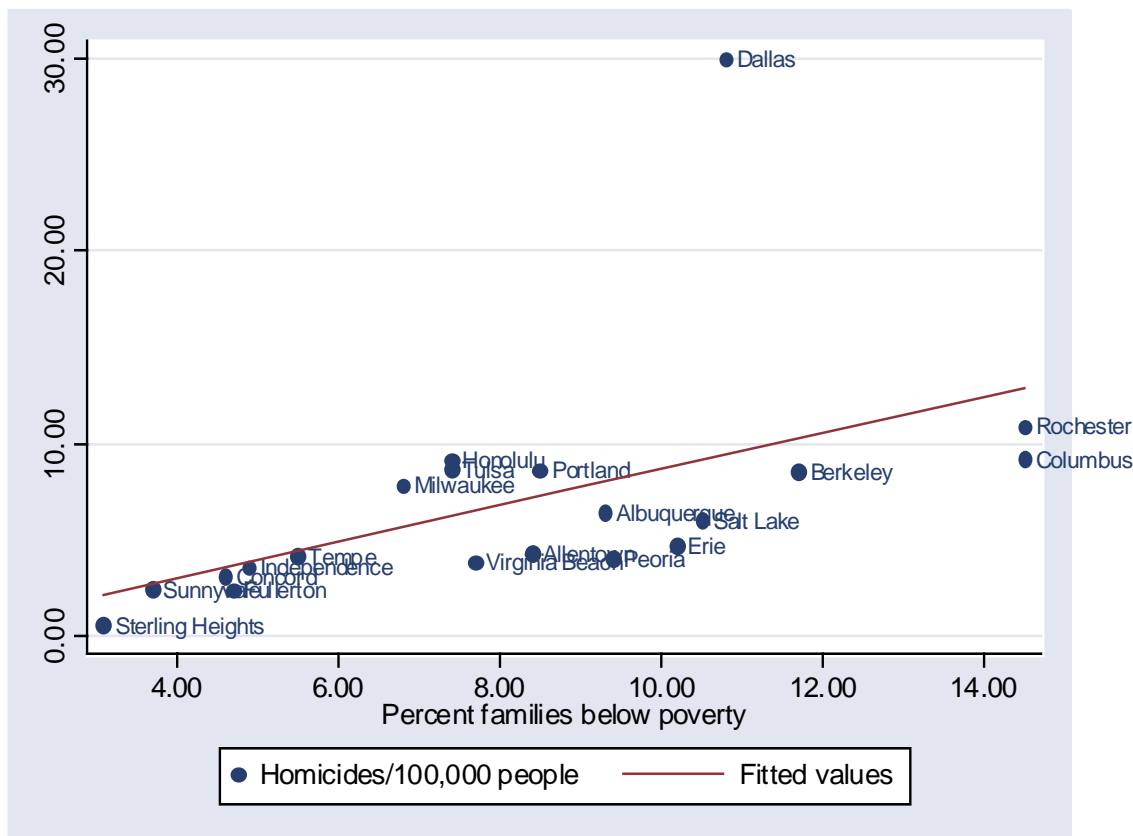
examining univariate distributions. Especially in multiple regression, all the correlations between dependent and explanatory variables are important.

```
. graph7 homic poor, xline(8.18) yline(6.9055) s([city]) oneway twoway box
```



The idea of the OLS regression is to find the best straight line to summarise the trend of points in a scatterplot (the straight line that best fits the data). Such line shows how the mean of Y changes at changing levels of X.

```
. twoway (scatter homic poor, mlabel(city)) (lfit homic poor)
```



## 2. OLS bivariate: analysis of the Stata output

OLS applied to a model with only the constant term as the explanatory variable,  $y_i = \alpha + \varepsilon_i$ , provides an estimate of the mean of the dependent variable:  $\hat{\alpha} = \bar{Y}$ .

```
reg homic
```

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

```
summ homic
```

Variable	Obs	Mean	Std. Dev.	Min	Max
homic	20	6.9055	6.12563	.55	29.98

In the simple linear regression,  $y_i = \alpha + \beta x_i + \varepsilon_i$ , estimated by OLS we know that:

$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$  is the intercept estimator and  $\hat{\beta} = \frac{C\hat{O}V(Y, X)}{\hat{V}A\hat{R}(X)}$  is the slope estimator.

```
reg homic poor
```

Source	SS	df	MS			
Model	181.370325	1	181.370325	Number of obs =	20	
Residual	531.573154	18	29.5318419	F( 1, 18) =	6.14	
Total	712.943479	19	37.523341	Prob > F =	0.0233	
				R-squared =	0.2544	
				Adj R-squared =	0.2130	
				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

Immediately after you run a regression, you can create a variable containing the predicted or fitted values,  $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$ .

```
predict homichat
```

In a similar way, you may create a variable containing the residuals,  $\hat{\varepsilon}_i = y_i - \hat{y}_i$

```
predict res, resid
```

Alternatively: `g res=homic-homichat`

**IMPORTANT:** Remember that once you run a new regression the predicted values will be based on the most recent regression.

By construction, fitted values of Y are uncorrelated with residuals:

corr res homichat

	resid homichat	
resid	1.0000	
homichat	-0.0000	1.0000

The constant estimate implies that the average homicide rate should equal  $-0.8$  in cities with 0 percent below poverty. That interpretation makes no sense, because we have no cities without poverty. Despite the constant term is important for providing simply interpretation of the regression output, the regression line may yield unreasonable results when projected beyond the X range of the data.

The slope estimate gives the direction and entity of the empirical relationship between the dependent and the explanatory variables: it implies that the average city homicide rates rise by 0.944 with each 1-point increase in the percentage of families below poverty.

The value 531.57 is the residual sum of squares,  $RSS = \sum_{i=1}^N \hat{\varepsilon}_i^2$ , from which it is computed the residual mean square (RMS) or the mean square error (MSE),  $RMS = MSE = s^2 = \frac{1}{N-K} \sum_{i=1}^N \hat{\varepsilon}_i^2 = 29.53$ , where  $N-K = 20-2 = 18$  is the residual df. The MSE is an estimate of the errors' variance. The corresponding standard deviation is  $\text{Root MSE} = \sqrt{s^2} = 5.43$ , that represents a measure of the goodness of the regression and it is usually called the standard error of the regression.

The value 181.37 is the model or explained sum of squares,  $MSS = \sum_{i=1}^N (\hat{y}_i - \bar{Y})^2$ , from which we obtain the model or explained mean square as  $MMS = \frac{1}{K-1} \sum_{i=1}^N (\hat{y}_i - \bar{Y})^2$ . It is an estimate of the variability of the fit,  $\hat{y}_i$ .

The value 712.94 is the total sum of squares,  $TSS = \sum_{i=1}^N (y_i - \bar{Y})^2$ , from which we have the total mean square,  $TMS = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{Y})^2$ . It is an estimate of the variability of  $y_i$ .

Another measure of the goodness of the regression model is the R-squared or the coefficient of determination of the regression:  $R^2 = \frac{MSS}{TSS} = 1 - \frac{RSS}{TSS}$ .

$R^2$  can also be computed as the squared correlation coefficient between the actual  $y_i$  and the fitted values,  $\hat{y}_i$ :

$$R^2 = \frac{\left(\sum y_i^* \hat{y}_i^*\right)^2}{\sum y_i^{*2} \sum \hat{y}_i^{*2}} = \hat{\rho}_{y, \hat{y}}^2,$$

where \* indicates deviations from sample means and  $\hat{\bar{y}}_i = \bar{y}_i$  because  $\bar{\varepsilon}_i = 0$ .

Pay attention that, in comparing different model specification, you have to use a measure corrected for the df, i.e. the  $AdjR^2 = \bar{R}^2 = 1 - \frac{RSS / (N - K)}{TSS / (N - 1)}$ .

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

Inference makes use of t (Student) and F (Fisher) test.

The t test verifies the significance of each single parameter estimate. It is based on the two hypotheses,  $H_0: \beta=0$  versus  $H_1: \beta \neq 0$  (thus, a two-sided alternative), and it uses the test statistic

$$\hat{t} = \frac{\hat{\beta} - 0}{s_{\hat{\beta}}}$$

The null in a two-tailed test is rejected if the value of the reported test statistic is higher than the critical values of the t distribution,  $t_{(N-K),5\%}$ , for a given sample dimension (N-K) and significance level or size (5%) i.e. the probability of a Type I error (reject  $H_0$  true). Also 1% or 10% may be used.

Critical values may be obtained as:

```
display invttail(18, .05/2)
2.100922
```

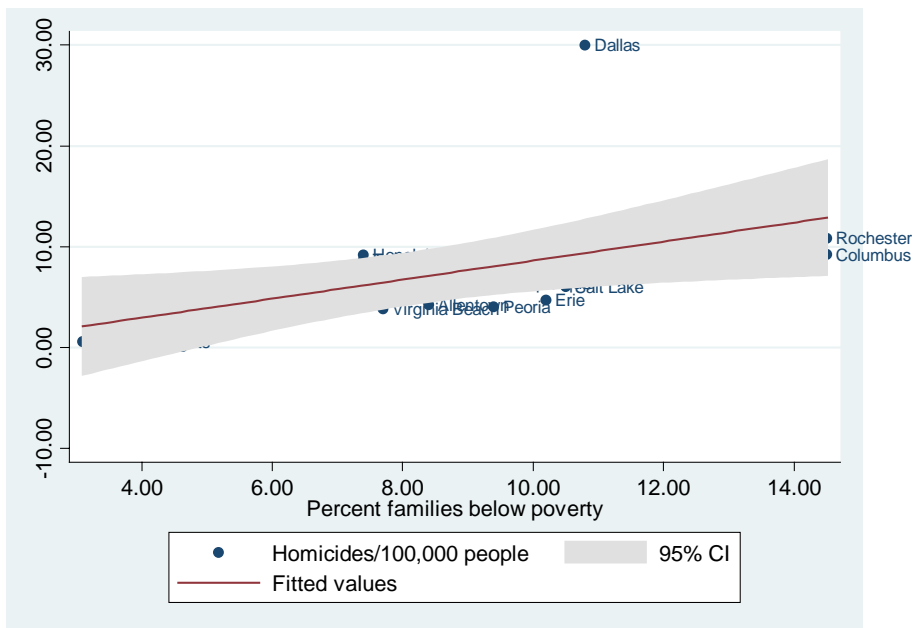
$P > |t|$  is the P-value, i.e. the estimated probability of a Type I error associated to the test statistic: the null is rejected if the P-value is lower than the chosen size (5%). In doing so, we make an error less than 5 times over 100.

If you know the value of the test-statistic, you may obtain the associated P-value as:

```
display ttail(18, .9438495/.3808596)*2
.02334034
```

The 95% confidence interval of the  $\hat{\beta}$  estimator is  $\hat{\beta} \pm t_{(N-K),5\%} s_{\hat{\beta}}$ . It represent another way to evaluate the significance of the estimate: this last is different from zero if the confidence interval does not contain the 0 value. Graphically:

```
. twoway (scatter homic poor, mlabel(city)) (lfitci homic poor)
```



Finally, the F-test statistic verify the significance of the model (constant excluded).

In the simple linear regression it corresponds to the square of the t-test. The F test may be reproduced by the command:

```
. test poor
( 1) poor = 0
      F( 1, 18) = 6.14
      Prob > F = 0.0233
```

OR

```
. testparm poor
( 1) poor = 0
      F( 1, 18) = 6.14
      Prob > F = 0.0233
```