Data analysis

Dmitry Yanter

DBTC 2016

Cases

- Performance of trading

  E.g. compare PnL across different sectors

- Performance of products

  E.g. build pricing adjustment models based on market events

- Hedging

  E.g. what is the best product to hedge with

Model characteristics

- Results make economic sense

  Model input variables make sense and its result is easy to make interpretation of.

- Maintainable

  If market changes its behavior model can be adjusted in controllable way, several versions on the model can be run for different times periods.

- Actionable

  Model results give an idea of what should be done to accomplish the goal.

Why linear methods?

- Almost everyone has intuitive understanding of linear models, no sophisticated math background is required
- “Cheap” to support: model can be tested and/or adjustment proposed with use of standard software
- Linear models results are easy to put into statement like “1% change in parameter X will cause a% change in Y”

Simple Linear regression

$$y = X\beta + \varepsilon$$

where y is vector of observed values, X is a matrix of variables.

Classic assumption is that errors ε are independent and have normal distribution.

Columns of X are called *independent variables, input variables, exogenous variables, regressors* etc.

y is called *dependent variable, response variable, endogenous variable, regressand* etc.

OLS

$$\beta = (X^T X)^{-1} X^T y$$

This solution can be interpreted as

- Linear fitting
- Solution of overdetermined system

In other words, this formula make sense event if the classic assumption is not applicable.

OLS Estimators

Fitted values

$$\hat{y} = X\beta$$

Residuals

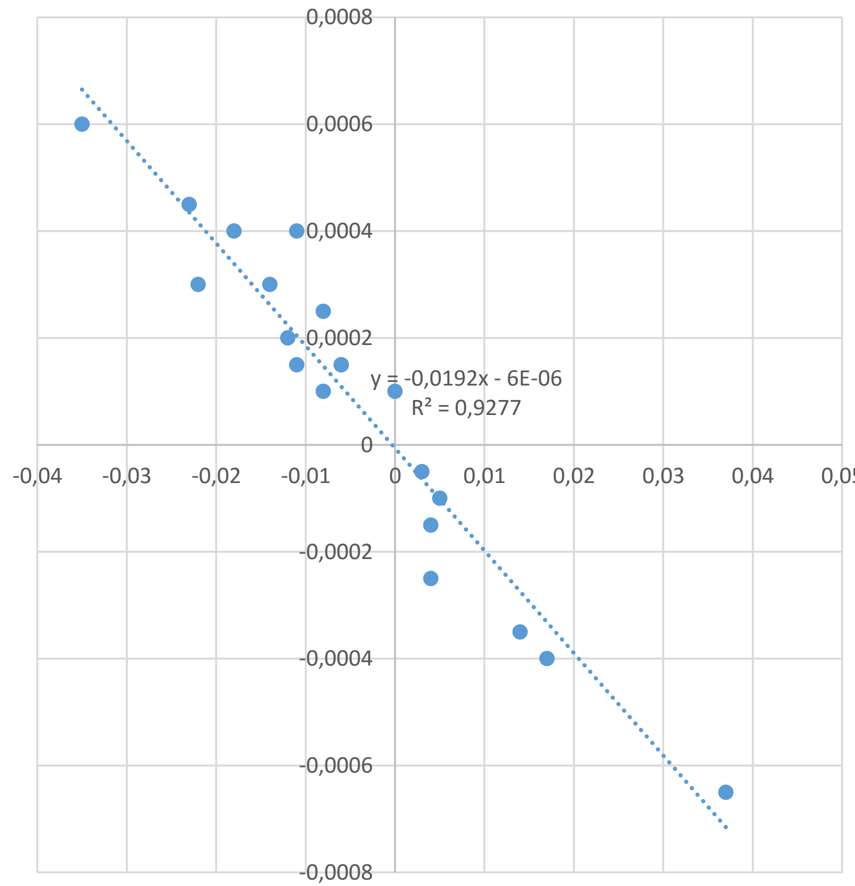
$$\hat{\varepsilon} = y - \hat{y}$$

Coefficient of determination

$$R^2 = \frac{\|\hat{y} - \bar{y}\|^2}{\|y - \bar{y}\|^2}$$

Bond over Bond Futures

DE 2Y Benchmark				Euro Schatz		
	Yield				Price	
29.02.2016	-0,57	-0,023		29.02.2016	111,945	0,00045
27.02.2016	-0,547	-0,005		26.02.2016	111,9	0,0001
26.02.2016	-0,542	-0,008		25.02.2016	111,89	0,0003
25.02.2016	-0,534	-0,014		24.02.2016	111,86	-0,0001
24.02.2016	-0,52	0,005		23.02.2016	111,87	-0,0002
23.02.2016	-0,525	0,004		22.02.2016	111,895	1E-04
22.02.2016	-0,529	0		19.02.2016	111,885	0,0002
19.02.2016	-0,529	-0,012		18.02.2016	111,865	0,00015
18.02.2016	-0,517	-0,011		17.02.2016	111,85	-5E-05
17.02.2016	-0,506	0,003		16.02.2016	111,855	-0,0002
16.02.2016	-0,509	0,004		15.02.2016	111,87	0,00015
15.02.2016	-0,513	-0,006		12.02.2016	111,855	-0,0006
14.02.2016	-0,507	-0,009		11.02.2016	111,92	0,0006
12.02.2016	-0,498	0,037		10.02.2016	111,86	-0,0004
11.02.2016	-0,535	-0,035		09.02.2016	111,9	0,0004
10.02.2016	-0,5	0,017		08.02.2016	111,86	0,0003
09.02.2016	-0,517	-0,011		05.02.2016	111,83	0,00015
08.02.2016	-0,506	-0,022		04.02.2016	111,815	-0,0003
05.02.2016	-0,484	-0,006		03.02.2016	111,85	0,0004
04.02.2016	-0,478	0,014		02.02.2016	111,81	0,00025
03.02.2016	-0,492	-0,018		01.02.2016	111,785	
02.02.2016	-0,474	-0,008				
01.02.2016	-0,466					



Returns

Durbin–Watson statistics

$$DW = \frac{\sum (e_t - e_{t-1})^2}{\sum e_t^2}$$

DW=0	ρ = 1
DW=2	ρ = 0
DW=4	ρ = -1

Confidence intervals for betas

Variance

$$s^2 = \frac{\|\mathrm{e}\|^2}{n-p}$$

Beta Variance

$$s_\beta^2 = s^2(X^T X)^{-1}$$

$$\frac{\hat{\beta} - \beta}{S_\beta} \sim t_{n-p}$$

Independent variable is significant if its confidence interval doesn’t include zero.