Econometrics

Jerzy Mycielski

Warsaw University, Department of Economics

2009

Linear model

Hypothesis: there is a relationship between explained and explanatory variables

Example

Hypothesis: expected spendings for food in households of married workers with two children depends on their income. This hypothesis seems to be supported by data (GUS data from household budget survey).

income of household	average spendings for food
0 -1000	442
1000-1500	534
1500-2000	608
2000-2500	657
2500-3000	717
3000-	817
average	644

- Average share of expenditure for food in income: 0.33

- **Auxiliary hypothesis**: relationship between explained and explanatory variables is linear (linear model):

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \ldots + x_{Ki}\beta_K + \varepsilon_i, \text{ for } i = 1, \ldots, N$$

Example

Direction of dependence:

$$\begin{aligned} \text{expenditure}_i &= \beta_1 + \beta_2 \text{income}_i + \varepsilon_i \\ \text{income}_i &= \alpha_1 + \alpha_2 \text{expenditure}_i + \eta_i \end{aligned}$$

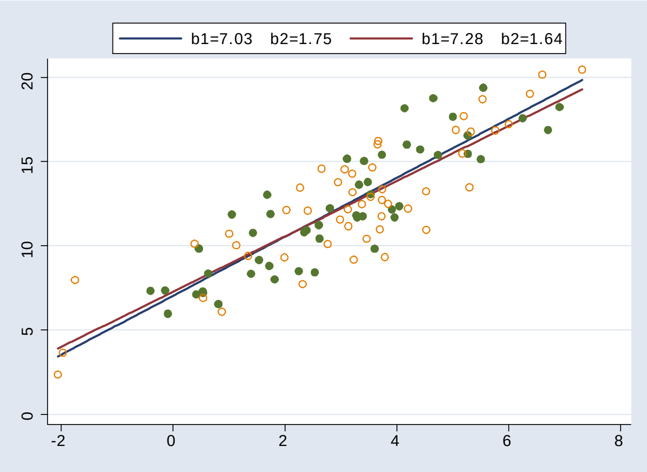
Example

It was found that there is a positive correlation between the ice-cream consumption and the number of drownings in a given day. Does it imply that after eating ice-cream it is not safe to swim?

Solution

More drownings happen in warm days as more people are swimming in such a day. For such days consumption of ice-cream is also higher.

- Estimates of parameters for two different samples for a model with *known* parameters $\beta_1 = 8$ and $\beta_2 = 1.5$ (Monte Carlo method)



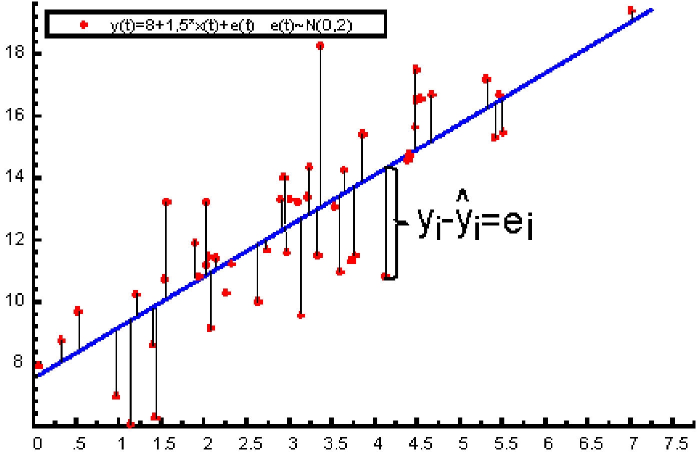
Example

(cont.) Estimates of $\boldsymbol{\beta}$ is equal to $\mathbf{b} = \begin{bmatrix} 463 \\ 0.08 \end{bmatrix}$

Fitted values and residuals for the first observation:

$$\begin{aligned} \widehat{y}_1 &= 463 + 0.08 \times 890.6 = 534.9 \\ e_1 &= 639.1 - 534.9 = 104.2 \end{aligned}$$

Residuals



y(t)=8+1,5*x(t)+e(t)   e(t)~N(0,2)
$y_i - \hat{y}_i = e_i$
18
16
14
12
10
8
0
.5
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5
6
6.5
7
7.5

OLS with constant and one explanatory variable

Example

variable	average	variance
q	644.18	46737
income	2262.34	1584300

Empirical covariance between variables is equal to 126211. Using derived formulas we obtain:

$$b_2 = \frac{126211}{1584300} = 0.079664$$

$$b_1 = 644.18 - 0.079664 \times 2262.34 = 463.95$$

Result of regression is usually reported with following table:

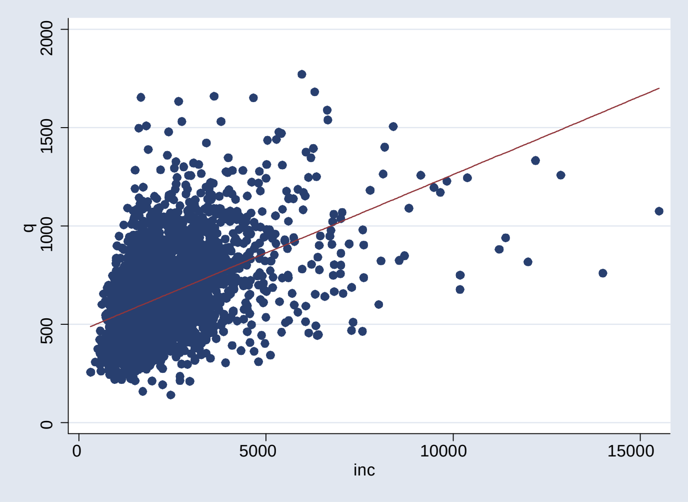
q	Coefficient
income	.079664
constant	463.95

Example

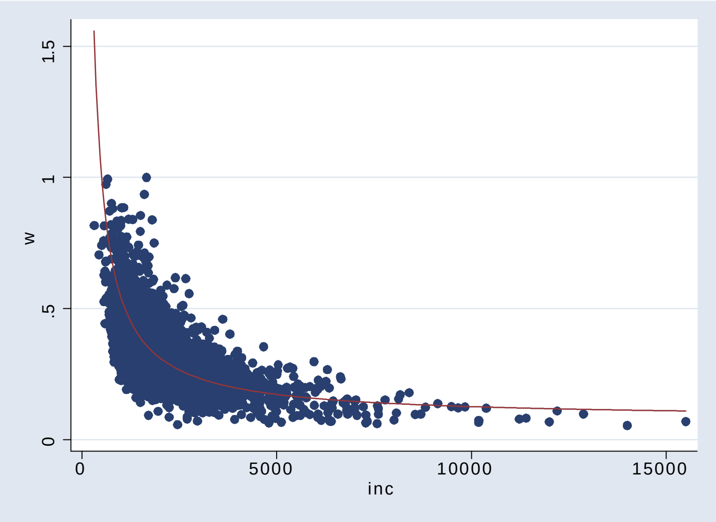
Estimation of the model explaing the expenditure for food with income of household:

$$q_i = \beta_1 + \beta_2 inc_i + \varepsilon_i$$

Dependence between expenditure for food and income (data and regression line)



Engle curve (data and estimated curve)



- $e_{inc} \approx \frac{0.8}{0.33} = 0.24$

Linear model - matrix notation

Example ((cont) In the model for food expenditure)

$$q_i = \beta_1 + \beta_2 \mathrm{inc}_i + \varepsilon_i$$

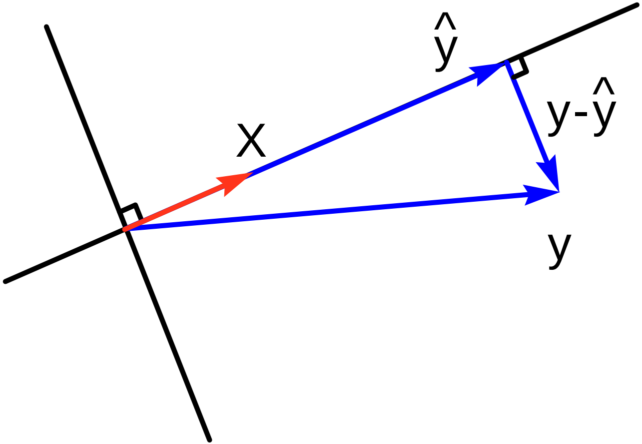
y and **X** have following form:

$$\mathbf{y} = \begin{bmatrix} 639.09 \\ 664.47 \\ 467.55 \\ \vdots \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 890.6 \\ 1 & 2300 \\ 1 & 1814.5 \\ \vdots & \vdots \end{bmatrix}$$

Notice: In model with constant first column of matrix **X** is column of ones.

Geometry of OLS

- $\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}}$ is perpendicular to $\mathbf{X}$!



Decomposition of the total sum of square

- One of the measures of variation is the sum of squares of variable around its mean
- We define following sum of squares:
  - **T**otal **S**um of **S**quares

$$TSS = (\mathbf{y} - \overline{\mathbf{y}})' (\mathbf{y} - \overline{\mathbf{y}}) = \sum_{i=1}^{N} (y_i - \overline{y})^2$$

    where $\overline{\mathbf{y}} = \mathbf{l}\overline{y}$
  - **E**xplained **S**um of **S**quares

$$ESS = \left(\widehat{\mathbf{y}} - \overline{\widehat{\mathbf{y}}}\right)' \left(\widehat{\mathbf{y}} - \overline{\widehat{\mathbf{y}}}\right) = \sum_{i=1}^{N} \left(\widehat{y}_i - \overline{\widehat{y}}\right)^2$$

    where $\overline{\widehat{\mathbf{y}}} = \mathbf{l}\overline{\widehat{y}}$.
  - **R**esidual **S**um of **S**quares

$$RSS = \mathbf{e}'\mathbf{e} = \sum_{i=1}^{N} e_i^2$$

Decomposition of the total sum of squares in model with constant

$$\underset{\text{Total variation}}{TSS} = \underset{\text{Explained variation}}{ESS} + \underset{\text{Unexpained variation}}{RSS}$$

- Total variation can be decomposed into the part which can be expained with the model and the part which cannot be expained with the model

Measures of fit: R2

Source	SS	df	MS
Model	33631553	1	33631553
Residual	122705284	3344	36694.164
Total	156336837	3345	46737.4701

Number of obs	=	3346
F(1, 397)	=	916.54
Prob > F	=	0.0000
R-squared	=	0.2151
Adj R-squared	=	0.2149
Root MSE	=	191.56

q	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
inc	.0796627	.0026314	30.27	0.000	.0745034	.0848219
_cons	463.9612	6.812147	68.11	0.000	450.6048	477.3176

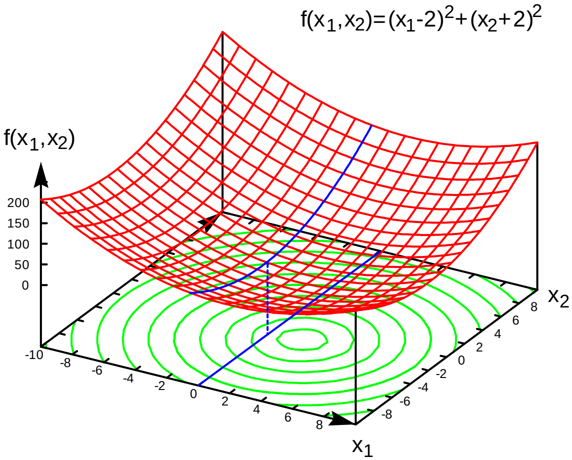
Measures of fit: R2

$$\underbrace{156336837}_{TSS} = \underbrace{33631553}_{ESS} + \underbrace{122705284}_{RSS}$$

$$R^2 = \frac{ESS}{TSS} = \frac{33631553}{156336837} = 0.2151$$

Constrained maximisation

- result of minimization of $f(x_1, x_2)$ for unconstrained x_1, x_2 and unconstrained x_2 but $x_1 = 0$



Measures of fit cont.

- Totaly random variable z_i was added to regression
- R^2

Model	RSS	R^2
$q_i = \beta_1 + \beta_2 \text{inc}_i + \varepsilon_i$	122705284	.2151
$q_i = \beta_1 + \beta_2 \text{inc}_i + \beta_3 z_i + \varepsilon_i$	122694775	.2152

- Adjusted measure:

$$\overline{R}^2 = 1 - \frac{N-1}{N-K}\left(1 - R^2\right)$$

- In previous model

Model	R^2	K	$\overline{R}^2$
$q_i = \beta_1 + \beta_2 \text{inc}_i + \varepsilon_i$	.2151	2	.2149
$q_i = \beta_1 + \beta_2 \text{inc}_i + \beta_3 z_i + \varepsilon_i$	.2152	3	.2147