The Classical Linear Regression Model in a Nutshell

Joachim Merz*

January 2016

Content

*Univ.-Prof. Dr. Joachim Merz, LEUPHANA University Lüneburg, Department of Economics, Research Institute on Professions (Forschungsinstitut Freie Berufe, FFB)), Chair ‚Statistics and Professions', CREPS (Center for Research in Entrepreneuship, Professions and Small Business Economics), IZA (Institute for the Study of Labour (Merz)), Scharnhorststr. 1, 21332 Lüneburg, Tel.: +49 4131 / 677- 2051, Fax: +49 4131 / 677- 2059, E-Mail: merz@uni.leuphana.de, (www.leuphana.de/ffb).

The Classical Linear Regression Model in a Nutshell

References: See e.g. econometric textbooks like Greene (2008), Wooldridge (2006, 2002) and Studenmund (2006); from a business and economics perspective Anderson et al. (2010); as recent German references Fahrmeir, Kneib and Lang (2009), Bauer, Fertig und Schmidt (2009), Hübler (2005) and von Auer (2003), and the regression script by Merz (2012).

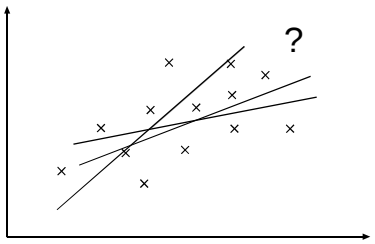
The **FFB (Forschungsinstitut Freie Berufe) e-learning moduls**: Lineare Regression - Deskriptives Modell (www.leuphana.de/ffb) (Merz and Stolze 2010a) and: Lineare Regression – Stochastisches Modell (Merz and Stolze 2010b) offer an easy respective audio/video internet based introduction to the Classical Linear Regression (CLR) model. For descriptive and inference/testing basics see Merz 2009, 2012.

1 Empirical Regression – Approach and OLS Method

General problem: Find a balancing structure/relation between a variable to be explained, y, and one or more variables as explanatory variables x.

Single Linear Regression y=f(x); Balancing structure/line in R2

General task: Find the balancing line for a one variable x explanation y=f(x) given a set of n observations (x_i, y_i) (cloud of n observations in a scatter plot).



Time series, longitudinal situation, Macro

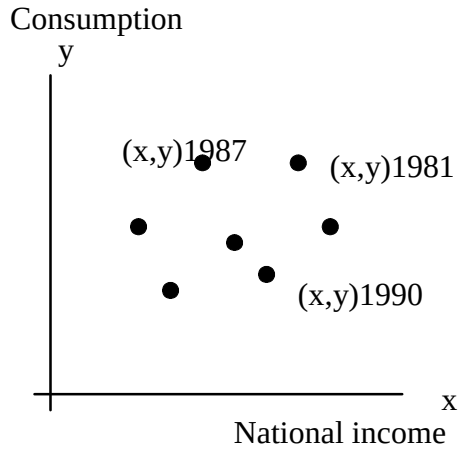
$\text{Consumption}_t = f(\text{GDP}_t) \quad t = 1,...,T$

For each year (period) there is one single pair of observation (x_t, y_t)

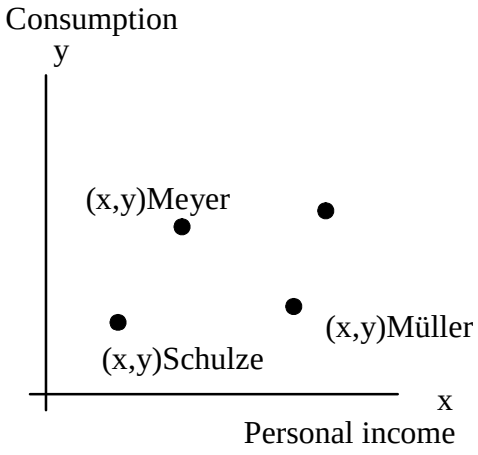
Cross section situation, Micro

$\text{consumption}_i = f(\text{income}_i) \quad i = 1,...,n$

For each microunit (person, household, firm …) there is one single pair of observation (x_i, y_i)



Longitudinal ; Macro



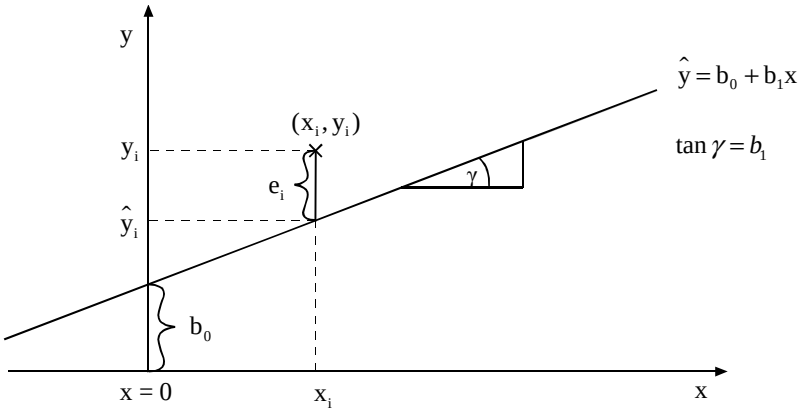
Cross-section ; Micro

Note: there is in principle no single year or microunit in a scatter plot directly visible; important is only y=f(x)

The regression problem and its solution

The balancing structure, the line of best fit in the R2 space shall be a straight line/linear slope $\hat{y}_i = b_0 + b_1$. The task then is to find the parameters b_0 and b_1 which characterize the optimal location of the balancing line.

Regression line in the R2 space



Ordinary Least Squares (OLS): Method to minimize the sum of squared deviations

$$S = \sum_{i=1}^{n} e_i^2 = e_1^2 + e_2^2 + \ldots + e_n^2 = \min! \qquad \text{with } e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$$

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = \min!$$

Partial derivations

$$\frac{\partial S}{\partial b_0} = \sum_{i=1}^{n}\left[2\left(y_i - b_0 - b_1 x_i\right)^{2-1}\cdot(-1)\right] = -2\sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_i\right) \quad \text{chain rule}$$

$$\frac{\partial S}{\partial b_1} = \sum_{i=1}^{n}\left[2\left(y_i - b_0 - b_1 x_i\right)\cdot(-x_i)\right] = -2\sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_i\right)\cdot x_i$$

Find optimum values b_0 and b_1

$$\frac{\partial S}{\partial b_0} = 0 \qquad \sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_i\right) = 0$$

$$\frac{\partial S}{\partial b_1} = 0 \sum_{i=1}^{n} x_i \cdot \left(y_i - b_0 - b_1 x_i\right) = 0$$

Normal equations

$$\sum_{i=1}^{n} b_0 + \sum_{i=1}^{n} b_1 x_i = \sum_{i=1}^{n} y_i \qquad n b_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\sum_{i=1}^{n} b_0 x_i + \sum_{i=1}^{n} b_1 x_i^2 = \sum_{i=1}^{n} y_i x_i \qquad b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

OLS single linear regression solution

$$b_0 = \frac{1}{n}\sum y_i - b_1 \cdot \frac{1}{n}\sum x_i = \bar{y} - b_1 \cdot \bar{x}$$

$$b_1 = \frac{\frac{1}{n}\sum y_i x_i - \bar{y}\cdot\bar{x}}{\frac{1}{n}\sum x_i^2 - \bar{x}^2} = \frac{s_{xy}}{s_x^2} = \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)} = r \cdot \frac{s_y}{s_x}$$

where r = Pearson correlation coefficient

$$b_0 = \frac{\sum x^2 \sum y - \sum x \sum xy}{n\sum x^2 - \sum x \sum x} = \frac{\bar{y}\frac{1}{n}\sum x^2 - \bar{x}\frac{1}{n}\sum xy}{\frac{1}{n}\sum x^2 - \bar{x}^2}$$

Example Income = b0 + b1 age

$y = b_0 + b_1 x_1 + e$ income=b_0 + $b_1$age + e

Data:

i	y_i	x_i	x_i^2	$x_i \cdot y_i$	y_i^2
1	1200	22	484	26400	1440000
2	1700	24	576	40800	2890000
3	3500	28	784	98000	12250000
4	4200	27	729	113400	17640000
5	1600	23	529	36800	2560000
6	5200	36	1296	187200	27040000
Sum	17400	160	4398	502600	63820000

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 26,67 \text{ years} \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 2900 \ (€ \text{ per month})$$

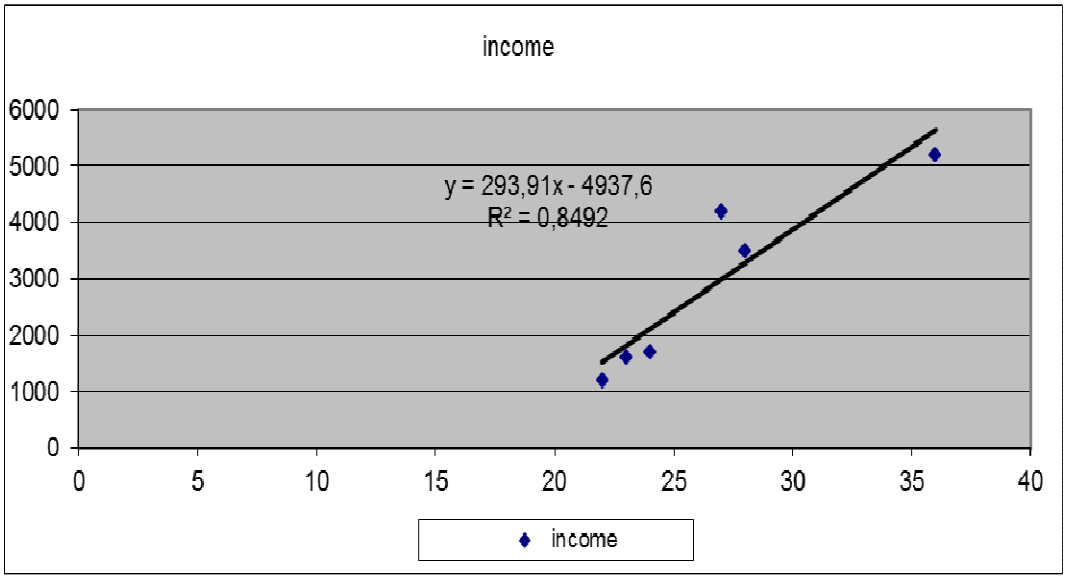
$$b_1 = \frac{\frac{1}{n}\sum y_i x_i - \overline{yx}}{\frac{1}{n}\sum x_i^2 - \bar{x}^2} = \frac{\frac{1}{6}502600 - 2900 \cdot 26,6667}{\frac{1}{6}4398 - 26,6667^2} = \frac{6433,3333}{21,8889} = 293,91$$

$$b_0 = \bar{y} - b_1 \cdot \bar{x} = 2900 - 293,908 \cdot 26,6667 = 4937,56$$

Regression line: $\hat{y}_i = b_0 + b_1 x_i$

incôme= -4937,56 + 293,91 age

Interpretation:
If age is increasing by one year (unit) then income will increase on average about 293,91 €



Multiple Linear Regression $y=f(x_1, x_2, \dots x_K)$; Balancing structure/hyperplane in R(K+1)

General task: Quantify the contribution of various variables x in explaining another variable y (using a linear equation): $y = f(x_1, x_2, \dots, x_K)$.

Given: n observations $(y_i, x_{i1}, x_{i2}, \dots, x_{iK})$ $(i = 1, \dots, n)$; n = cases

Observations: regions, countries, firms, individuals, years … any unit with a set of characteristics

Balancing hyperplane

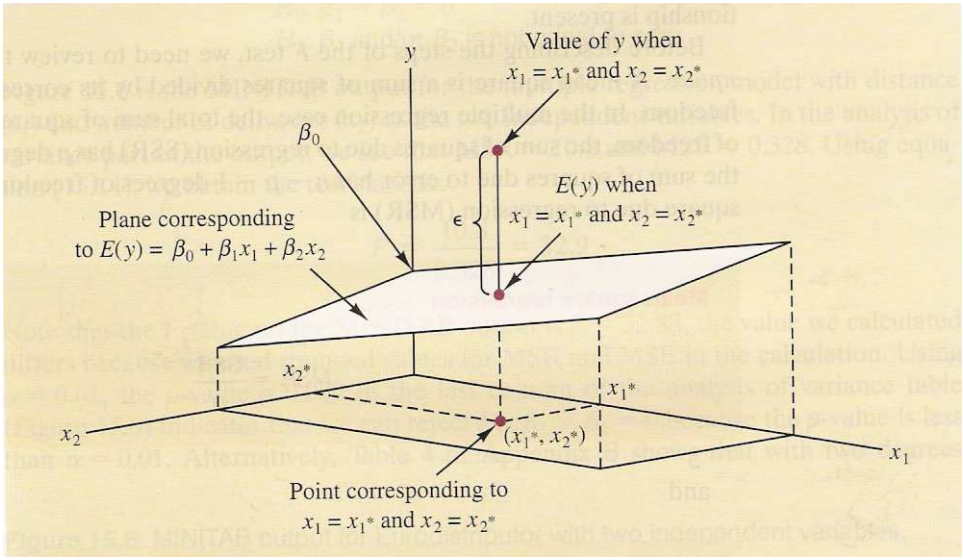
$$y_i = \underbrace{b_0(1 = x_{i0}) + b_1 x_{i1} + b_2 x_{i2} + \dots b_K x_{iK}}_{\text{Hyperplane}} + \underbrace{e_i}_{\text{Deviation}}$$

Multiple Linear Regression: Balancing line in $\mathbb{R}^{K+1}$

$$y_i = \sum_{k=0}^{K} b_k x_{ik} + e_i \qquad (i = 1, \dots, n = \text{number of cases})$$

$x_{i0} = 1$ (constant) $\qquad (k = 1, \dots, K = \text{number of variables})$

Regression hyperplane in the R3 space



Source: Anderson et al. 2010, p. 573

The regression problem and solution

Find the parameter vector b which characterizes the optimal location of the best balancing hyperplane by

Ordinary Least Squares (OLS) method as minimizing the sum of squared deviations

$$S = \sum_{i=1}^{n} e_i^2 = e_1^2 + e_2^2 + \dots + e_n^2 = \min!$$

again with $e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_1 + b_2 x_2 + \dots + b_K x_K)$:

Matrix notation

$$y = Xb + e$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1K} \\ 1 & x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nK} \end{pmatrix} \begin{pmatrix} b_0 \\ \vdots \\ b_K \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

$$y_{n\times 1} = X_{n\times(K+1)} \; b_{(K+1)\times 1} + e_{n\times 1}$$

$$S = \sum_{i=1}^{n} e_i^2 = e_1^2 + e_2^2 + \ldots + e_n^2 = \min!$$

$$S = \mathbf{e'e} = (e_1, e_2, \ldots, e_n) \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \sum_{i=1}^{n} e_i^2 = e_1^2 + e_2^2 + \ldots + e_n^2$$

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \sum_{k=0}^{K} b_k x_{ik})^2 = \sum_{i=1}^{n} (y_i - \mathbf{b'x}_i)^2 = \mathbf{e'e} = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) = \min!$$

$$S = \mathbf{e'e} = \mathbf{y'y} - 2\mathbf{b'}\underbrace{\mathbf{X'y}}_{\mathbf{a}} + \mathbf{b'}\underbrace{\mathbf{X'X}}_{\mathbf{A}}\mathbf{b}$$

$$\frac{\partial S(\mathbf{b})}{\partial \mathbf{b}} = \frac{\partial \mathbf{e'e}}{\partial \mathbf{b}} = -2\mathbf{X'y} + 2\mathbf{X'Xb}$$

$$\frac{\partial S(\hat{\mathbf{b}})}{\partial \hat{\mathbf{b}}} = \mathbf{0} \qquad \hat{\mathbf{b}} = \text{minimizing } \mathbf{b}, \text{ minimizes } S(\mathbf{b})$$

$$\boxed{\mathbf{X'X}\hat{\mathbf{b}} = \mathbf{X'y}} \qquad \text{(Normal equations)}$$

$$(\mathbf{X'X})^{-1}(\mathbf{X'X})\hat{\mathbf{b}} = (\mathbf{X'X})^{-1}\mathbf{X'y}$$

Solution:

$$\boxed{\hat{\mathbf{b}} = \mathbf{b}_{OLS} = (\mathbf{X'X})^{-1}\mathbf{X'y}}$$

Central bricks to compute $\hat{\mathbf{b}} = \mathbf{b}_{OLS}$

X'X, matrix of moments

$$\mathbf{X'X} = \overbrace{\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \ldots & x_{n1} \\ x_{12} & x_{22} & \ldots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1K} & x_{2K} & \ldots & x_{nK} \end{pmatrix}}^{X'}_{(K+1)\times n} \overbrace{\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1K} \\ 1 & x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nK} \end{pmatrix}}^{X}_{n\times(K+1)}$$

$$= \begin{pmatrix} \sum_i 1 = n & \sum_i x_{i1} & \sum_i x_{i2} & \cdots & \sum_i x_{iK} \\ \sum_i x_{i1} & \sum_i x_{i1}^2 & \sum_i x_{i1}x_{i2} & \cdots & \sum_i x_{i1}x_{iK} \\ \sum_i x_{i2} & \sum_i x_{i2}x_{i1} & \sum_i x_{i2}^2 & \cdots & \sum_i x_{i2}x_{iK} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_i x_{iK} & \sum_i x_{iK}x_{i1} & \sum_i x_{iK}x_{i2} & \cdots & \sum_i x_{iK}^2 \end{pmatrix}_{(K+1)\times(K+1)\,\text{symmetric}}$$

Matrix of moments **X'X**

- symmetric
- first element: $\sum_i 1 = n$ number of cases
- first row: $\sum_i x_{ik}$ Sum over cases of variable k
  $(k = 1,2,\ldots,K)$ K = number of explanatory variables
- Diagonal element: $\sum_i x_{ik}^2$ (sum of x squares)
- Secondary diagonal: $\sum_i x_{ik}x_{il}$ (sum of cross products)

X'y, cross products x_k and y

$$\mathbf{X'y} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \ldots & x_{n1} \\ x_{12} & x_{22} & \ldots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1K} & x_{2K} & \ldots & x_{nK} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_i y_i \\ \sum_i x_{i1}y_i \\ \sum_i x_{i2}y_i \\ \vdots \\ \sum_i x_{iK}y_i \end{pmatrix}_{(K+1)\times 1}$$

- First value: Sum of y
- Next values: Sum of cross products x_k and y

$(X'X)^{-1}$, computation of the inverse via Excel or computer programs based on various mathematical methods (triangulation, row transformation by Gauss …).

Simple calculation of a 2×2 inverse matrix (not extendable for higher dimensions)

$$A=\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad \text{Inverse:} \quad A^{-1}=\frac{1}{|A|}\begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}$$

with Determinant $|A| = a_{11}a_{22} - a_{12}a_{21}$

Example Single Regression (matrix): income = b0 + b1 age

$y = b_0 + b_1 x_1 + e$ $\qquad$ $\text{income} = b_0 + b_1 \text{age} + e$

Endogeneous variable Income (y), exogeneous variable: age (x_1), error term (e)

Solution: $b = (X'X)^{-1} X'y$

Data:

i	income y	x_0	age x_1
1	1200	1	22
2	1700	1	24
3	3500	1	28
4	4200	1	27
5	1600	1	23
6	5200	1	36

$$\Rightarrow \qquad X = \begin{pmatrix} 1 & 22 \\ 1 & 24 \\ 1 & 28 \\ 1 & 27 \\ 1 & 23 \\ 1 & 36 \end{pmatrix}_{(6\times 3)}$$

Matrix of moments:

$$(X'X) = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 22 & 24 & 28 & 27 & 23 & 36 \end{pmatrix}_{(2x6)} \begin{pmatrix} 1 & 22 \\ 1 & 24 \\ 1 & 28 \\ 1 & 27 \\ 1 & 23 \\ 1 & 36 \end{pmatrix}_{(6x2)}$$

$$= \begin{pmatrix} 6 & 160 \\ 160 & 4398 \end{pmatrix}_{(2x2)}$$

Interpretation of the (symmetric) Matrix of moments:

$$6 = \sum_i x_{i0}^2 = \sum_i (x_{i0} = 1)^2 = n$$

$$160 = \sum_i 1 \cdot x_{i1} = \sum_i x_{i1} = \sum_i 1 \cdot x_{i1} \cdot 1 = \text{Sum 'age'}$$

$$4398 = \sum_i x_{ii}^2 = \text{Sum of squared 'age'}$$

Cross products:

$$(X'y) = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 22 & 24 & 28 & 27 & 23 & 36 \end{pmatrix}_{(2\times 6)} \begin{pmatrix} 1200 \\ 1700 \\ 3500 \\ 4200 \\ 1600 \\ 5200 \end{pmatrix}_{(6\times 1)}$$

$$= \begin{pmatrix} 17400 \\ 502600 \end{pmatrix}_{(2\times 1)} \begin{matrix} \sum_i \text{income} \\ = \sum_i \text{age} \cdot \text{income} \end{matrix}$$

Vector of OLS-Regression coefficients

$$b = (X'X)^{-1} X'y$$

Calculation of the Inverse of a matrix e.g. by Gauß $(X'X) \xrightarrow{Gauß} (X'X)^{-1}$

Calculation of the Inverse of a 2×2 – Matrix by the adjoint

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad \text{Inverse:} \quad A^{-1} = \frac{1}{|A|} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}$$

where Determinant $|A| = a_{11}a_{22} - a_{12}a_{21}$

$$A = (X'X) = \begin{pmatrix} 6 & 160 \\ 160 & 4398 \end{pmatrix}$$

$$A^{-1} = \frac{1}{6 \cdot 4398 - 160 \cdot 160} \begin{pmatrix} 4398 & -160 \\ -160 & 6 \end{pmatrix}$$

$$= \begin{pmatrix} 5,5812 & -0,2030 \\ -0,2030 & 0,7614 \cdot 10^{-2} \end{pmatrix}$$

$$b = \begin{pmatrix} 5,58122 & -0,203046 \\ -0,203046 & 0,761421E-02 \end{pmatrix}_{(2\times2)} \begin{pmatrix} 17400 \\ 502600 \end{pmatrix}_{(2\times1)}$$

$$= \begin{pmatrix} -4937,56 \\ 293,91 \end{pmatrix} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$$

Regression result

$$\hat{y} = b_0 + b_1 x_1$$

incôme= -4937,56 + 293,91 age

Interpretation

On average an additional year of age yields an income of 293,91 EURO.

Single Regression (ET) $\text{income} = b_0 + b_1 \cdot \text{age}$

```
DATA LISTING (Current sample)
Observation      INCOME       AGE          SEX
      1         1200.0      22.000       1.0000
      2         1700.0      24.000       1.0000
      3         3500.0      28.000       1.0000
      4         4200.0      27.000       .00000
      5         1600.0      23.000       1.0000
      6         5200.0      36.000       .00000

     1. Matrix -> XSX=XDOT(ONE,AGE)

      <<<< XSX      >>>>   COLUMN
                1             2
 ROW   1   6.00000       160.000
 ROW   2   160.000       4398.00

     2. Matrix -> XSX=XDOT(X)
```

```
     <<<< XSX      >>>>   COLUMN
               1            2
ROW   1   6.00000       160.000
ROW   2   160.000       4398.00


    1. Matrix -> XSY=XDOT(X,INCOME)

     <<<< XSY      >>>>   COLUMN
               1
ROW   1   17400.0
ROW   2   502600.

    1. Matrix -> XSXINV=GINV(XSX)

     <<<< XSXINV   >>>>   COLUMN
               1            2
ROW   1   5.58122      -.203046
ROW   2  -.203046       .761421E-02



    1. Matrix -> BOLS=XSXINV|XSY

     <<<< BOLS     >>>>   COLUMN
               1
ROW   1  -4937.56
ROW   2   293.909



    2. Matrix -> BET=XLSQ(X,INCOME)

     <<<< BET      >>>>   COLUMN
               1
ROW   1  -4937.56
ROW   2   293.909
```

Regression by the ET macro command: regress lhs rhs

```
================================================================================
Ordinary    Least Squares
Dependent Variable         INCOME      Number of Observations  6
Mean of Dep. Variable    2900.0000     Std. Dev. of Dep. Var.   1634.625339
Std. Error of Regr.       709.7758     Sum of Squared Residuals .201513E+07
R - squared                 .84917     Adjusted R - squared          .81146
F( 1,   4)                 22.5194     Prob. Value for F             .00900
================================================================================
Variable  Coefficient  Std. Error t-ratio Prob|t|>x  Mean of X Std.Dev.of
--------------------------------------------------------------------------------
Constant  -4937.56       1677.     -2.945  .04220
AGE        293.909       61.93      4.745  .00900     26.66667   5.12510
```

Example Multiple Regression: income = b0 + b1 age + b2 sex

Income should be explained by age and sex (gender) given a fictive sample of n = 6 persons as microunits.

Data:

i	income y	$x_0 = 1$	age x_1	sex x_2
1	1200	1	22	1
2	1700	1	24	1
3	3500	1	28	1
4	4200	1	27	0
5	1600	1	23	1
6	5200	1	36	0

K = 2 number of explanatory variables (age and sex); sex is a dummy variable (0=male, 1=female)

$$(X'X) = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 22 & 24 & 28 & 27 & 23 & 36 \\ 1 & 1 & 1 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 22 & 1 \\ 1 & 24 & 1 \\ 1 & 28 & 1 \\ 1 & 27 & 0 \\ 1 & 23 & 1 \\ 1 & 36 & 0 \end{pmatrix} = \begin{pmatrix} 6 & 160 & 4 \\ 160 & 4398 & 97 \\ 4 & 97 & 4 \end{pmatrix}$$

(3x6) (6x3) (3x3)

$$6 = \sum_i x_{i0}^2 = \sum_i (x_{i0} = 1)^2 = n$$

$$160 = \sum_i 1 \cdot x_{i1} = \sum_i x_{i1} = \text{Sum of ages}$$

$$97 = \sum_i x_{i1} x_{i2} = \text{Sum of cross products (age} \cdot \text{sex)}$$

$$4 = \sum_i x_{i2}^2 = \text{Sum of sex}^2$$

Cross products

$$(X'y) = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 22 & 24 & 28 & 27 & 23 & 36 \\ 1 & 1 & 1 & 0 & 1 & 0 \end{pmatrix}_{(3,6)} \begin{pmatrix} 1200 \\ 1700 \\ 3500 \\ 4200 \\ 1600 \\ 5200 \end{pmatrix}_{(6,1)} = \begin{pmatrix} 17400 \\ 502600 \\ 8000 \end{pmatrix}_{(3,1)} = \begin{matrix} \sum_i \text{income} \\ \sum_i \text{age} \cdot \text{income} \\ \sum_i \text{sex} \cdot \text{income} \end{matrix}$$

Vector of the OLS estimated regression coefficients

$$b = (X'X)^{-1} X'y \qquad (X'X) \xrightarrow{Gauß} (X'X)^{-1} \qquad \text{symmetric}$$

$$b = \begin{pmatrix} 16,7000 & -0,514286 & -4,22857 \\ -0,514286 & 0,163265E-01 & 0,118367 \\ -4,22857 & 0,118367 & 1,60816 \end{pmatrix}_{(3,3)} \begin{pmatrix} 17400 \\ 502600 \\ 8000 \end{pmatrix}_{(3,1)}$$

$$= \begin{pmatrix} -1728,57 \\ 204,082 \\ -1220,41 \end{pmatrix} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} \qquad \hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

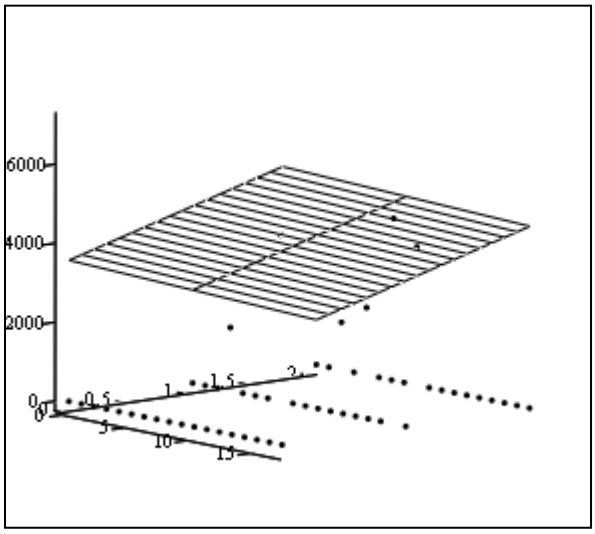
incôme = -1728,57 + 204,08 · age - 1220,41 · sex

Interpretation

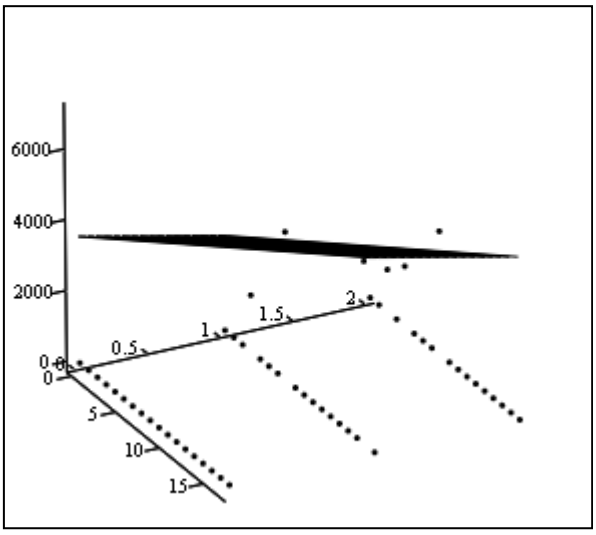
- Averaged an additional year of age results in additional 204,08 EURO.
- Women on average will have 1220,41 EURO less than men.

Regression hyperplane in R3

income=- 1728.57 + 204.08 age - 1220.41 sex

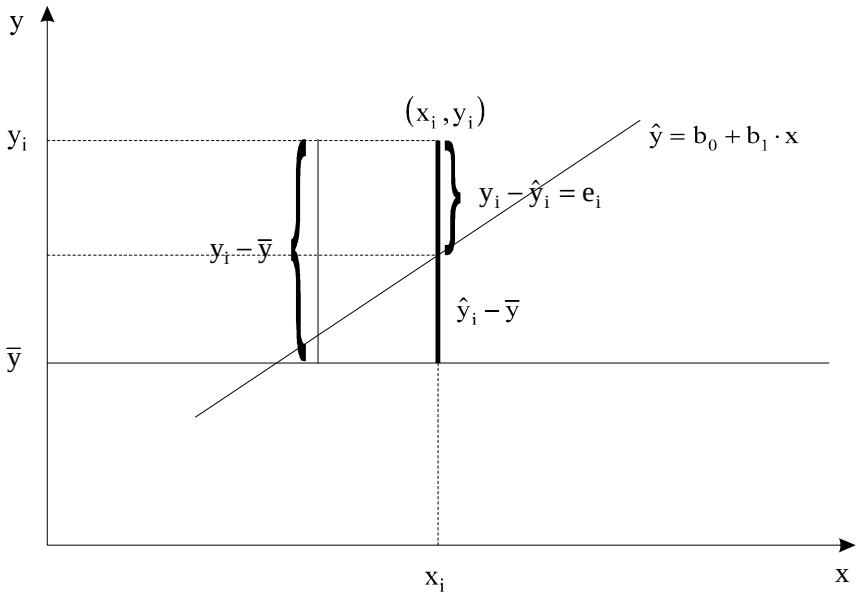


coul, B



coul, B

2 Classical Linear Regression Model – Goodness of fit



$$\underbrace{(y_i - \bar{y})}_{\text{deviation to be explained (total deviation)}} = \underbrace{(\hat{y}_i - \bar{y})}_{\text{explained deviation by regression}} + \underbrace{(y_i - \hat{y}_i)}_{\text{not explained deviation}} \quad (i = 1, \ldots, n)$$

Estimation by OLS results in

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{Sum of squared total deviations}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{Sum of squared explained deviations}} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{Sum of squared residual deviations}}$$

$$SQT = SQE + SQR$$

Coefficient of determination

$$B = r^2 = R^2 = \frac{SQE}{SQT} = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2} = \frac{1/n\sum(\hat{y} - \bar{y})^2}{1/n\sum(y - \bar{y})^2}$$

$$= \frac{SQT - SQR}{SQT} = 1 - \frac{SQR}{SQT} = 1 - \frac{\mathbf{e'e}}{\mathbf{y'y} - n\bar{y}^2} = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \bar{y})^2}$$

Adjusted R^2 : panelizes a regression with many var iables for mod el comparisons :

$$\bar{R}^2 = 1 - \frac{n-1}{n-K-1}(1 - R^2) \qquad (\text{adjusted } R^2)$$

Coefficient of Determination: Proportion explained variance of observed variance

$0 \leq R^2 \leq 1$ (‚the nearer 1, the better')

3 Classical Linear Regression Model – Assumptions

CLR-Assumption: data generating process by the stochastic relation

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{iK} + \varepsilon_i = E\left(y_i \left| x_{i0}, \ldots, x_{iK}\right.\right) + \varepsilon_i$$

Variable to be explained: **y**

Explanatory variabels **X**

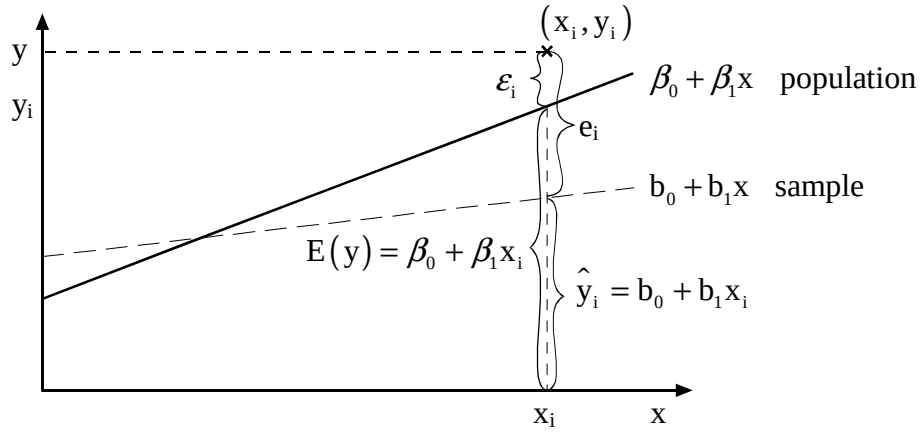
Population: β, ε

Estimation of parameter vector: **b**

Error term : ε

Population regression: $E(\mathbf{y})=\mathbf{X}\beta$

Estimation of E(**y**): $E(\hat{\mathbf{y}}) = \hat{\mathbf{y}} = \mathbf{Xb}$



Reasons for a stochastic approach:

- Not all influences are met by the model (residual ε); some influence are not ascertainable at all (e.g. insecurity about future developments)
- Measurement errors and inadequate recognition ("Adäquationsproblem") and capturing of variables
- The „true" functional relationship is not captured by the choice of functional specification
- Variability of human (nature) behaviour
- …

Model formulation CLR (Classical Linear Regression)

- n observation as a sample of a population
- conditional distribution of the dependent (left hand side, lhs) variable y given the independent (right hand side, rhs) variables $x_1, \ldots, x_K$

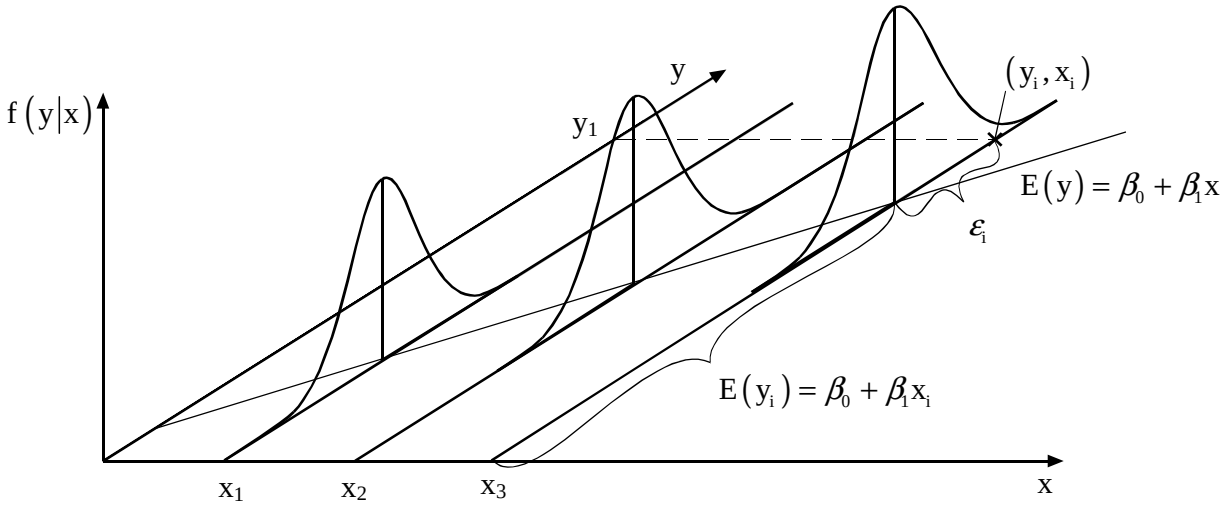
- Mean of the conditional random distribution $\bar{y}\left(\mu, E(Y)\right)$ ia a linear function of the variables $x_1, \ldots, x_K$

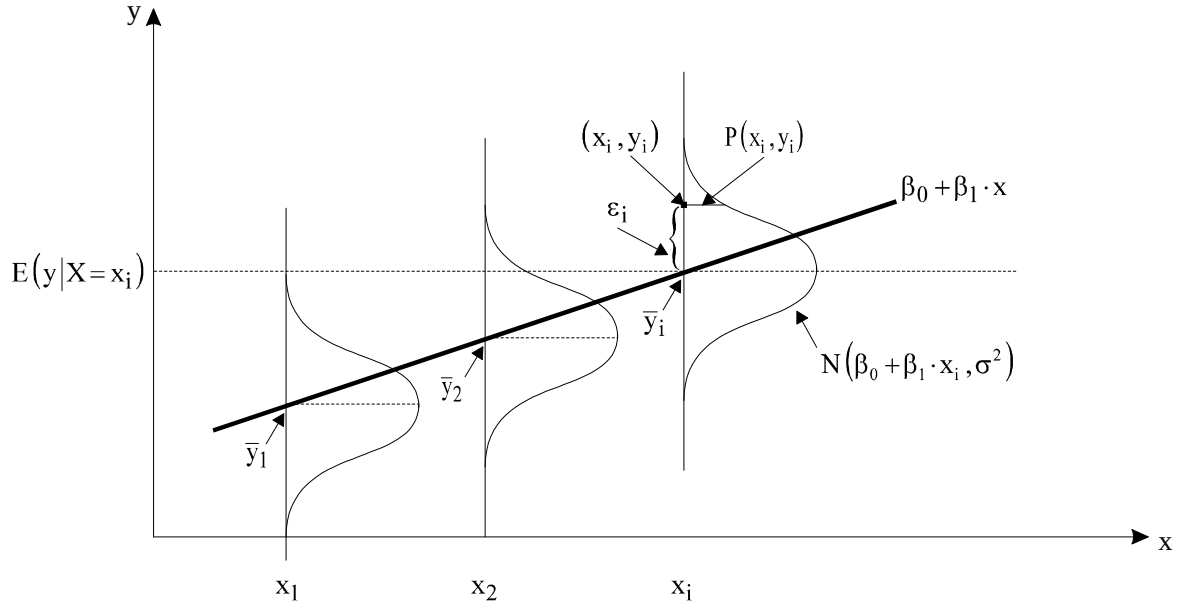
  i.e. $E(y_i|\mathbf{x_i}) = \mathbf{x_i'}\boldsymbol{\beta}$ $\quad \mathbf{x_i'} = (x_{i0}, x_{i1}, x_{i2}, \ldots, x_{iK})$ *bzw.*

  $E(\mathbf{y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$

- Variance of y is constant; y is (via ε) stochastic
- The y-values are not correlated in repeated sampling; the independent variables x_K are not stochastic and in each sample the same ("fixed in repeated samples").
- Identical residual variances σ_ε^2 for all n observations

Population regression in R2 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$





Stochastic Regression Model

Assumptions about the underlying 'data-generating process'

A1: $\mathbf{y = X\beta + \varepsilon}$ **Functional form**

Linearity: $y_i = \beta_0(1 = x_{i0}) + \beta_1 x_{i1} + \beta_2 x_{i2} + ...\beta_K x_{iK} + \varepsilon_i$

$y_i = \beta' x_i + \varepsilon_i$

specifies a linear relationship between y and $x_1, x_2, \ldots, x_K$

Linearising of nonlinear relations:

$f(y_i) = \beta_0 + \beta_1 g_1(x_1) + \beta_2 g_2(x_2) + \ldots + \beta_K g_K(x_K) + \varepsilon_i$

e.g. $g(x) = x^2$ or $g(x) = 1/x$ or $g(x) = e^x$ like $\text{income}_i = \beta_0 + \beta \text{age} + \beta_2 \text{age}^2 + \varepsilon_i$

Further transformations to a linear regression function

Model	Structural form	Linear form	Slope dy/dx	Elasticity* (dy/dx)x/y
Linear	$y = b_0 + b_1 x$	$y = b_0 + b_1 x$	b_1	$b_1 (x/y)$
Quadratic	$y = b_0 + b_1 x^2$	$y = b_0 + b_1 (x^2)$	$b_1 (2x)$	$b_1 (2x^2/y)$
Reciproke	$y = b_0 + b_1 1/x$	$y = b_0 + b_1 (1/x)$	$-b_1 (1/x^2)$	$-b_1 [1/(xy)]$
Log-linear	$y = \exp(b_0 + b_1 \ln x)$	$\ln y = b_0 + b_1 \ln x$	$b_1 (y/x)$	b_1
Log-lin	$y = \exp(b_0 + b_1 x)$	$\ln y = b_0 + b_1 x$	$b_1 (y)$	$b_1 (x)$
Lin-log	$\exp(y) = \exp(b_0 + b_1 \ln x)$	$y = b_0 + b_1 \ln x$	$b_1 (1/x)$	$b_1 (1/y)$

* Elasticity often measured at the means of y and x.

$\ln x$ $\quad y = b_0 x^{\beta} + \varepsilon_i \rightarrow \ln y = b_0 + b_1 \ln x$

$1/x$ $\quad y = b_0 + b_1 (1/x)$

e^x $\quad y = b_0 + b_1 (e^x)$

Interpretation of log-models (Wooldridge 2009, 46)

- If the model is $\ln(y) = b_0 + b_1 \ln(x) + u$

  b_1 is the elasticity of y with respect to x

- If the model is $\ln(y) = b_0 + b_1 x + u$

  b_1 is approximately the percentage change in y given a 1 unit change in x

- If the model is $y = b_0 + b_1 \ln(x) + u$

  b_1 is approximately the change in y for a 100 percent change in x

A2: $\mathbf{E(\varepsilon) = 0}$ **Expectation of the residuals is 0**

$E(\varepsilon_i \mid x_{j0}, x_{j1}, ..., x_{jK}) = 0$ The expected value of the disturbances at observation i is not a function of the independent variables observed at any observation, including this one.

A3: $\mathbf{E(\varepsilon\varepsilon') = \sigma^2 I}$ **Homoskedasticity:** $\mathbf{Var(\varepsilon_i) = \sigma^2}$ $\forall$**i**

Zero covariance: $\mathbf{Cov(\varepsilon_i, \varepsilon_j) = 0}$ **i ≠ j**

$E(\varepsilon - E(\varepsilon))(\varepsilon - E(\varepsilon))' = E(\varepsilon\varepsilon')$ ($E(\varepsilon) = \mathbf{0}$ (A2)) is the variance-covariance matrix of the residuals

$$E\left[\begin{pmatrix}\varepsilon_1\\ \vdots\\ \varepsilon_n\end{pmatrix}(\varepsilon_1 \quad \dots \quad \varepsilon_n)\right] = E\begin{pmatrix}\varepsilon_1^2 & \varepsilon_1\varepsilon_2 & \cdots & \varepsilon_1\varepsilon_n\\ \varepsilon_2\varepsilon_1 & \varepsilon_2^2 & \cdots & \varepsilon_2\varepsilon_n\\ \vdots & \vdots & \ddots & \vdots\\ \varepsilon_n\varepsilon_1 & \varepsilon_n\varepsilon_2 & \cdots & \varepsilon_n^2\end{pmatrix}$$

$$= \begin{pmatrix}\mathrm{Var}(\varepsilon_1) & \mathrm{Cov}(\varepsilon_1\varepsilon_2) & \cdots & \mathrm{Cov}(\varepsilon_1\varepsilon_n)\\ \mathrm{Cov}(\varepsilon_2\varepsilon_1) & \mathrm{Var}(\varepsilon_2) & \cdots & \mathrm{Cov}(\varepsilon_2\varepsilon_n)\\ \vdots & \vdots & \ddots & \vdots\\ \mathrm{Cov}(\varepsilon_n\varepsilon_1) & \mathrm{Cov}(\varepsilon_n\varepsilon_2) & \cdots & \mathrm{Var}(\varepsilon_n)\end{pmatrix}$$

$$= \begin{pmatrix}\sigma^2 & 0 & \cdots & 0\\ 0 & \sigma^2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \sigma^2\end{pmatrix} = \sigma^2\begin{pmatrix}1 & 0 & \cdots & 0\\ 0 & 1 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 1\end{pmatrix} = \sigma^2\mathbf{I}$$

Implications

Variances of all ε_i (i = 1, ..., n) are always identical

Homoskedasticity $E(\varepsilon_i^2) = \sigma^2$ $\forall i \ (i = 1, \dots, n)$

Residuals are not correlated with other residuals

$E(\varepsilon_i\varepsilon_j) = 0$ $i \neq j$ $(i, j = 1, \dots, n)$

A4: X is not stochastic

In repeated samples **X** is fix but not y.

A5: $\mathbf{Rank(X) = K + 1 \leq n}$

Necessary condition to compute $(\mathbf{X'X})^{-1}$. There is only a balancing problem if Rank(X)>n (more observations than unknown regression coefficients to be computed).

A6: ε is (multivariate) normal distributed

Because the error term is the sum of many different unobserved factors affecting y, the central limit theorem can be invoked to conclude that ε has an approximate normal distribution.

This assumption is needed for testing the regression coefficients.

4 Classical Linear Regression Model – Estimation

Ordinary Least Squares Fitting Criterion: Minimizing the sum of squared residuals

$\sum_i \varepsilon_i^2 = \sum_i (y_i - \beta' \mathbf{x}_i)^2 = \min!$ $\mathbf{x}_i$ is (K+1) vector

$\text{Min } S(\beta) = \varepsilon'\varepsilon = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$

similar solution as of $= \text{Min } S(b) = e'e$ see above

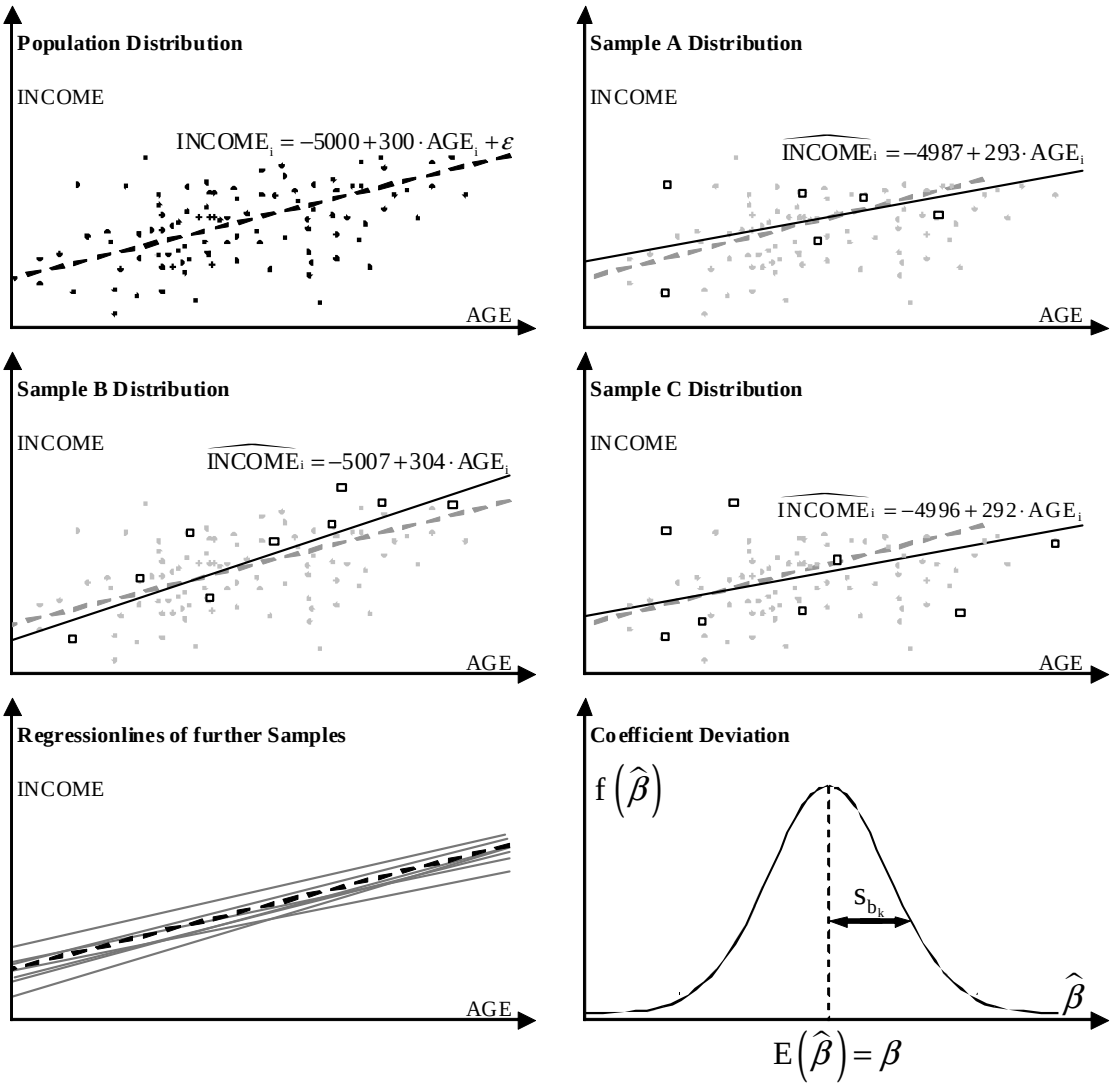
OLS estimator for β:

$$\boxed{\hat{\beta} = \mathbf{b} = \mathbf{b}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}$$

Variance-covariance matrix of b

Vector **b** is a function of y (and y is a function of ε) is stochastic; thus **b** has a variance.

For illustration: Imaging to take many samples out of a population and estimate a (different) **b** for each sample. Then we have a distribution of the many **b**s:



Desirable estimation properties of an estimator: to be unbiased and having minimum variance:

Gauss-Markov-Theorem

In the Classical Linear Regression Model the best (minimal variance) linear unbiased estimator of β (BLUE = best linear unbiased estimator) is the OLS-estimator

$$\hat{\beta} = \mathbf{b} = \mathbf{b}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

with the variance-covariance-matrix

$$\mathrm{Var}[\mathbf{b}] = \Sigma_{\mathbf{bb}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

and its unbiased estimator

$$S_{bb} = s^2(\mathbf{X}'\mathbf{X})^{-1} = \frac{\mathbf{e}'\mathbf{e}}{n-K-1}(\mathbf{X}'\mathbf{X})^{-1}$$

Exkurs: Standard error and standard error of regression

A standard error is an *estimate of the standard deviation* of an estimator (mean, variance, regression coefficient, etc.). It is the standard deviation of the sampling distribution of a statistic. For example, in the case of estimating the variability of a sample mean, the standard error of the mean is the standard deviation of the sample means over all possible samples drawn from the population.

The **standard error of the data** is the estimate of the standard deviation of the data itselves

$$se(X) = \sqrt{\text{var}(X)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

with X as a random variable of a sample. The correction (using n-1 instead of n, Bessel's correction) provides with se^2 an unbiased estimator for the variance σ^2 of the underlying population. Additionally, if n = 1, then there is no indication of deviation from the mean, and standard deviation should therefore be undefined. The term sample standard deviation is used for the corrected estimator (using n-1). Note, the population standard deviation is $\sigma = \sqrt{\text{var}(x)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

The **standard error of the mean** is the estimate of the standard deviation of the mean

$$se(\bar{X}) = se(X)/\sqrt{n} .$$

The **standard error of regression** is the estimate of the variability of the data around the regression line.

$$se(e) = \sqrt{\frac{1}{n-K-1}\sum_{i=1}^{n}e_i^2} = \sqrt{\frac{\mathbf{e'e}}{n-K-1}} .$$

Imagine a point on the regression line for a particular value of x. The standard error of regression is the estimate of the spread of the data above and below the fitted point. $se(e) = s$ is the estimate of the residual standard deviation σ .

End of Exkurs

5 Classical Linear Regression Model – Testing for Significance

F-test for overall significance, goodness of fit

For all parameters (except β_0):

$$H_0 : \beta_1 = \beta_2 = ... = \beta_K = 0$$
$$H_1 : \text{one or more of the parameters is not equal to zero}$$

Test statistic

$$F_{observed} = \frac{R^2/K}{(1-R^2)/(n-K-1)} = \frac{SQE/K}{SQR/(n-K-1)} \sim F_{\alpha}^{(K,n-K-1)}$$

Rejection rule

Critical value approach:	Reject H_0	if $F_{critical} = F_{\alpha}^{(K,n-K-1)}$	$> F_{observed}$
p-value approach:	Reject H_0	if	p-value $< \alpha$

where F is based on the F distribution with $\nu_1 = K$ and $\nu_2 = n - K - 1$ degrees of freedom.

t-test for individual significance

For any parameter β_k:

$$H_0 : \beta_k = 0$$
$$H_1 : \beta_k \neq 0$$

Test statistic

$$t_{observed} = \frac{b_k - 0}{s_{b_k}} \sim t_{\alpha}^{(n-K-1)}$$

Rejection rule

Critical value approach:	Reject H_0	if $t_{observed} = abs(t_{\alpha}^{(n-K-1)}) > t_{critical}$
p-value approach:	Reject H_0	if p-value $< \alpha$

where t is based on the t distribution with $\nu = n - K - 1$ degrees of freedom.

Note: $H_1 : \beta_k \neq 0$ requires a two-tailed test with $\alpha/2$ to get $t_{critical}$ from a t table.

Example F-test, t-test

working hours = f(wage, sex, age, hhsize) (fictive data)
CLR estimation (ET/LIMDEP) of

Observation	HOURS	WAGE	SEX	AGE	HHSIZE
1	38.000	20.000	.00000	27.000	2.0000
2	40.000	24.000	.00000	32.000	3.0000
3	42.000	28.000	.00000	35.000	4.0000
4	18.000	15.000	.00000	22.000	1.0000
5	20.000	15.000	1.0000	21.000	1.0000
6	35.000	22.000	1.0000	27.000	3.0000
7	19.000	23.000	1.0000	32.000	4.0000
8	36.000	25.000	.00000	34.000	3.0000
9	44.000	32.000	1.0000	45.000	3.0000
10	38.000	16.000	1.0000	48.000	3.0000
11	40.000	28.000	1.0000	36.000	4.0000
12	44.000	36.000	1.0000	38.000	5.0000
13	48.000	35.000	1.0000	35.000	1.0000
14	28.000	22.000	.00000	29.000	3.0000
15	43.000	42.000	1.0000	34.000	2.0000

ET-Ergebnis

```
================================================================================
Ordinary    Least Squares
Dependent Variable         HOURS      Number of Observations              15
Mean of Dep. Variable      35.5333    Std. Dev. of Dep. Var.         9.738485
Std. Error of Regr.         6.2195    Sum of Squared Residuals        386.825
R - squared                 .70866    Adjusted R - squared             .59212
F( 4,  10)                 6.0810    Prob. Value for F                .00954
================================================================================
Variable  Coefficient Std. Error  t-ratio Prob|t|>x Mean of X  Std.Dev.of X
--------------------------------------------------------------------------------
Constant  -.844767      8.059       -.105   .91859
WAGE       .733063      .2350       3.120   .01088   25.53333       8.02555
SEX       -4.29397      3.630      -1.183   .26417     .60000        .50709
```

```
AGE          .710602       .2856        2.488    .03208    33.00000        7.40656
HHSIZE      -1.14748       1.593        -.720    .48780     2.80000        1.20712
```

SPSS-Ergebnis

ANOVA[b]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	940,909	4	235,227	6,081	,010[a]
	Nicht standardisierte Residuen	386,825	10	38,682		
	Gesamt	1327,733	14			

a. Einflußvariablen : (Konstante), HHSIZE, SEX, WAGE, AGE

b. Abhängige Variable: HOURS

Koeffizienten[a]

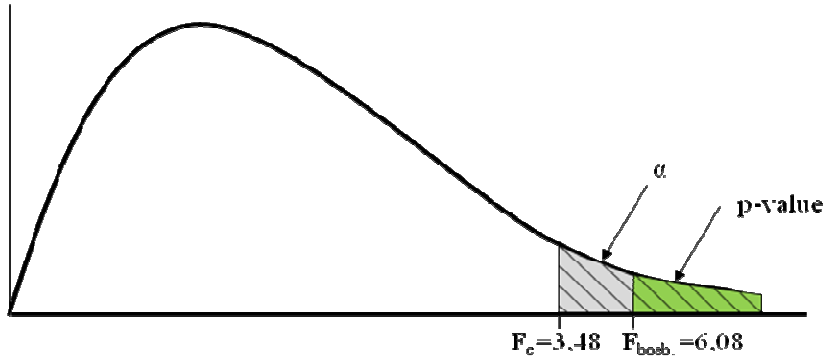
Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.
		RegressionskoeffizientB	Standardfehler	Beta		
1	(Konstante)	-,845	8,059		-,105	,919
	WAGE	,733	,235	,604	3,120	,011
	SEX	-4,294	3,630	-,224	-1,183	,264
	AGE	,711	,286	,540	2,488	,032
	HHSIZE	-1,147	1,593	-,142	-,720	,488

a. Abhängige Variable: HOURS

F-test for overall goodness of fit

$F_{observed}(4,10) = 6,081$, then the associated p-value = 0,00954 (see also the F-distribution table)

pdf



Rejection rule: $F_{observed} = F(4,10) = 6,081 > F_{critical} = 3,478$ (α=0,05, say)

p-value = 0,00954 < α=0,05

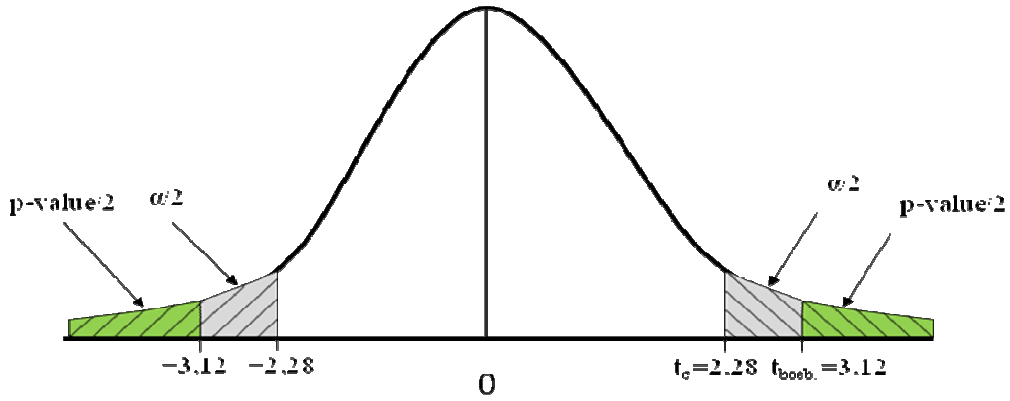
H_0 = 'no significant overall goodness of fit' is rejected in favour for H_1.

t-test for individual significance

Significance of the WAGE coefficient?

The WAGE coefficient b_1=.733063 has a calculated t-test value $t_{observed}$ = 3,120 with the associated p-value = 0,01088 (see the t distribution table)

pdf



Rejection rule: $t_{bobserved}$ = 3,120 > $t_{critical}$ = 2,2281 (ν= n-K-1=10, α=0,05, say)

p-value = 0,01088 < α=0,05 resp.

p-value/2 = 0,00544 < α/2=0,025

H_0 = 'no individual significance' is rejected in favour for H_1 (significant influence of WAGE on HOURS) because $t_{observed} < t_{critical}$ respectively p-value < α.

6 Classical Linear Regression Model – Examples

6.1 Income = f(age) (ET/LIMDEP)

Example 1: Income = f(age) (ET/LIMDEP)

Observation	INCOME	AGE	SEX
1	1200.0	22.000	1.0000
2	1700.0	24.000	1.0000
3	3500.0	28.000	1.0000
4	4200.0	27.000	.00000
5	1600.0	23.000	1.0000
6	5200.0	36.000	.00000

XSX=XDOT(ONE,AGE)

<<<< **XSX** >>>> **COLUMN**

	1	2
ROW 1	6.00000	160.000
ROW 2	160.000	4398.00

XSY=XDOT(X,INCOME)

<<<< **XSY** >>>> **COLUMN**

	1
ROW 1	17400.0
ROW 2	502600.

XSXINV=GINV(XSX)

<<<< **XSXINV** >>>> **COLUMN**

	1	2
ROW 1	5.58122	-.203046
ROW 2	-.203046	.761421E-02

Inverse of the 2*2 matrix $(\mathbf{X'X})^{-1}$

$$A=\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad \text{Inverse:} \quad A^{-1}=\frac{1}{|A|}\begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}$$

with determinant $|A|=a_{11}a_{22}-a_{12}a_{21}$

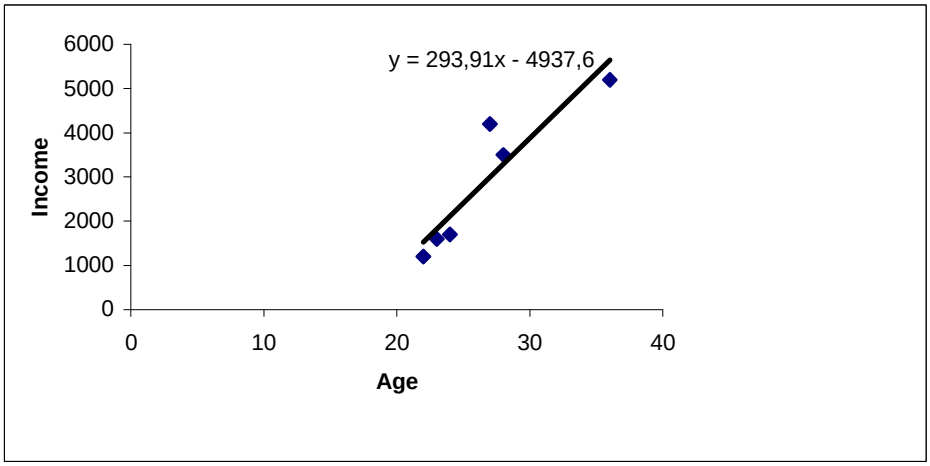
$$A=(X'X)=\begin{pmatrix} 6 & 160 \\ 160 & 4398 \end{pmatrix}$$

$$A^{-1}=\frac{1}{6\cdot 4398-160\cdot 160}\begin{pmatrix} 4398 & -160 \\ -160 & 6 \end{pmatrix}$$

$$=\begin{pmatrix} 5,5812 & -0,2030 \\ -0,2030 & 0,7614\cdot 10^{-2} \end{pmatrix}$$

Ordinary Least Squares			
Dependent Variable	INCOME	Number of Observations	6
Mean of Dep. Variable	2900.0000	Std. Dev. of Dep. Var.	1634.625339
Std. Error of Regr.	709.7758	Sum of Squared Residuals	.201513E+07
R - squared	.84917	Adjusted R - squared	.81146
F(1, 4)	22.5194	Prob. Value for F	.00900

Variable	Coefficient	Std. Error	t-ratio	Prob|t|>x	Mean of X	Std.Dev.of X
Constant	-4937.56	1677.	-2.945	.04220		
AGE	293.909	61.93	4.745	.00900	26.66667	5.12510



The **standard error of regression** is the estimate of σ, the standard deviation of the error terms, which describes the variability of the data around the regression line.

$$\hat{\sigma} = s(e) = \sqrt{\frac{1}{n-K-1}\sum_{i=1}^{n} e_i^2} = \sqrt{\frac{e'e}{n-K-1}}.$$

Imagine a point on the regression line for a particular value of x. The standard error of regression is the estimate of the spread of the data above and below the fitted point.

ET:
Std. Error of Regr. = sqrt(1/(n-K-1) Sum of Squared Residuals)
Std. Dev. of Dep. Var. = sqrt(1/(n-1) Sum(Y-mean of Y)²)

e'e = **Sum of Squared Residuals = .201513E+07**
n=6, K=1, **e'e**/4 = 503782,5

Std. Error of Regr. = s(e) = sqrt(e'e/4) = sqrt(503782,5) = 709,7758

$$S_{bb} = s^2 (\mathbf{X}'\mathbf{X})^{-1} = \frac{\mathbf{e}'\mathbf{e}}{n-K-1}(\mathbf{X}'\mathbf{X})^{-1}$$

$$S_{bb} = s^2 (\mathbf{X}'\mathbf{X})^{-1} = 503782,5(\mathbf{X}'\mathbf{X})^{-1}$$

$$S_{bb} = 503782,5 \begin{pmatrix} 5,58122 & -0,203046 \\ -0,203046 & 0,761421E-02 \end{pmatrix}_{(2\times2)}$$

$$= \begin{pmatrix} 2811720,965 & -102291,02 \\ -102291,02 & 3835,91 \end{pmatrix}$$

$$s_{b_1} = \sqrt{2811720,965} = 1676,819$$

$$s_{b_2} = \sqrt{3835,91} = 61,934$$

Is exact the above result of Std. Error (standard deviation of b):

Variable	Coefficient	Std. Error	t-ratio	Prob|t|>x
Constant	**-4937.56**	***1677.***	**-2.945**	**.04220**
AGE	**293.909**	***61.93***	**4.745**	**.00900**

Std. Error of coefficient (standard deviation of coefficient) =

$$\sqrt{\text{var}(b_k)} = \sqrt{\text{Main diagonal elements of } S_{bb}}$$

6.2 Income = f(age, sex) (EXCEL)

Example 2: Income = f(age,sex) (EXCEL)

Regression module in EXCEL:
Activate in EXTRAS Add-Ins ‘Analyse Funktionen’, then click in EXTRAS ‘Analyse Funktionen’ the submodul ‘Regression’

Income=b0+b1age+b2sex

i	income	age	sex
1	1200	22	1
2	1700	24	1
3	3500	28	1
4	4200	27	0
5	1600	23	1
6	5200	36	0

sex=0 male, 1 female

AUSGABE: ZUSAMMENFASSUNG

Regressions-Statistik	
Multipler Korrelati	0,958378606
Bestimmtheitsma	0,918489552
Adjustiertes Besti	0,864149253
Standardfehler	602,4891678
Beobachtungen	6

ANOVA

	Freiheitsgrade (df)	*adratsummen (S:e*	*Quadratsumme*	*Prüfgröße (F)*	*F krit*
Regression	2	12271020,41	6135510,204	16,90254873	0,023271261
Residue	3	1088979,592	362993,1973		
Gesamt	5	13360000			

	Koeffizienten	*Standardfehler*	*t-Statistik*	*P-Wert*	*Untere 95%*	*Obere 95%*	*Untere 95,0%*	*Obere 95,0%*
Schnittpunkt	-1728,571429	2462,110151	-0,702069088	0,533204051	-9564,112132	6106,969275	-9564,112132	6106,969275
X Variable 1	204,0816327	76,98324199	2,650987765	0,076928807	-40,91363122	449,0768965	-40,91363122	449,0768965
X Variable 2	-1220,408163	764,0368613	-1,597315817	0,208480104	-3651,916731	1211,100404	-3651,916731	1211,100404

6.3 Gasoline Sales in the US-Market (ET/LIMDEP, EViews, SPSS)

Example 3: Gasoline Sales in the US-Market

The table below lists data that characterize sales of gasoline in the U.S. market from 1960 to 1986 (Greene 2008).

```
Gasoline sales in the US market from 1960 – 1982.
      G     = gasoline sales, in billions of gallons
      PG    = price index for gasoline
      Y     = per capita income
      PNC   = price index for new cars
      PUC   = price index for used cars
      PPT   = price index for public transportation
      PD    = price index for consumer durables
      PN    = price index for nondurables
      PS    = price index for services
DATA LISTING (Current sample)
Year   G       PG     Y     PNC    PUC    PPT    PD     PN     PS
1960 129.7   .925  6036  1.045   .836   .810   .444   .331   .302
1961 131.3   .914  6113  1.045   .869   .846   .448   .335   .307
1962 137.1   .919  6271  1.041   .948   .874   .457   .338   .314
1963 141.6   .918  6378  1.035   .960   .885   .463   .343   .320
1964 148.8   .914  6727  1.032  1.001   .901   .470   .347   .325
1965 155.9   .949  7027  1.009   .994   .919   .471   .353   .332
1966 164.9   .970  7280   .991   .970   .952   .475   .366   .342
1967 171.0  1.000  7513  1.000  1.000  1.000   .483   .375   .353
1968 183.4  1.014  7728  1.028  1.028  1.046   .501   .390   .368
1969 195.8  1.047  7891  1.044  1.031  1.127   .514   .409   .386
```

1970	207.4	1.056	8134	1.076	1.043	1.285	.527	.427	.407
1971	218.3	1.063	8322	1.120	1.102	1.377	.547	.442	.431
1972	226.8	1.076	8562	1.110	1.105	1.434	.555	.458	.451
1973	237.9	1.181	9042	1.111	1.176	1.448	.566	.497	.474
1974	225.8	1.599	8867	1.175	1.226	1.480	.604	.572	.513
1975	232.4	1.708	8944	1.276	1.464	1.586	.659	.615	.556
1976	241.7	1.779	9175	1.357	1.679	1.742	.695	.638	.598
1977	249.2	1.882	9381	1.429	1.828	1.824	.727	.671	.648
1978	261.3	1.963	9735	1.538	1.865	1.878	.769	.719	.698
1979	248.9	2.656	9829	1.660	2.010	2.003	.821	.800	.756
1980	226.8	3.691	9722	1.793	2.081	2.516	.892	.894	.839
1981	225.6	4.109	9769	1.902	2.569	3.120	.957	.969	.926
1982	228.8	3.894	9725	1.976	2.964	3.460	1.000	1.000	1.000
1983	239.6	3.764	9930	2.026	3.297	3.626	1.041	1.021	1.062
1984	244.7	3.707	10421	2.085	3.757	3.852	1.038	1.050	1.117
1985	245.8	3.738	10563	2.152	3.797	4.028	1.045	1.075	1.173
1986	269.4	2.921	10780	2.240	3.632	4.264	1.053	1.069	1.224

Example Gasoline Sales in the US-Market (ET/LIMDEP)

Ordinary Least Squares

Dependent Variable	G	Number of Observations	27
Mean of Dep. Variable	207.0333	Std. Dev. of Dep. Var.	43.798927
Std. Error of Regr.(1)	5.4157	Sum of Squared Residuals	586.606
R - squared	.98824	Adjusted R - squared	.98471
F(6, 20)	280.0881	Prob. Value for F	.00000

Variable	Coefficient	Std. Error	t-ratio	Prob|t|>x	Mean of X	Std.Dev.of X
Constant	-10209.5	4644.	-2.198	.03987		
YEAR	5.21657	2.401	2.172	.04203	1973.00000	7.93725
PG	-18.4347	3.695	-4.989	.00007	1.90211	1.16791
Y	.187961E-01	.9968E-02	1.886	.07396	8513.51852	1455.62903
PNC	27.4422	18.65	1.471	.15680	1.38133	.42797
PUC	-14.4631	8.007	-1.806	.08593	1.71230	.97474
PPT	-7.44493	7.654	-.973	.34235	1.86233	1.10831

Estimated V-C Matrix for Parameters

	1-ONE	2-YEAR	3-PG	4-Y	5-PNC
1-ONE	.215710E+08				
2-YEAR	-11153.3	5.76690			
3-PG	8666.22	-4.46827	13.6542		
4-Y	45.7814	-.236774°-01	.184537°-01	.993618°-04	
5-PNC	5180.41	-2.78086	-37.5415	.728855°-02	347.929
6-PUC	746.702	-.375470	7.81011	.425357°-02	-65.9089
7-PPT	10645.2	-5.48129	2.72227	.191899°-01	-28.0169

	6-PUC	7-PPT
6-PUC	64.1078	
7-PPT	-40.6463	58.5889

P r e d i c t e d V a l u e s
(* => not in estimating sample)

Observation	Observed Y	Predicted Y	Residual	95% Forecast Interval	
1	129.70	121.94	7.7582	108.659	135.225
2	131.30	128.06	3.2368	115.567	140.560
3	137.10	134.70	2.4035	122.448	146.945
4	141.60	141.52	.0774	128.933	154.112
5	148.80	152.58	-3.7783	140.582	164.574

6	155.90	162.12	-6.2246	150.287	173.962
7	164.90	171.32	-6.4169	159.481	183.153
8	171.00	179.82	-8.8156	167.956	191.675
9	183.40	188.84	-5.4362	177.078	200.594
10	195.80	196.30	-.5008	184.545	208.057
11	207.40	205.45	1.9527	193.483	217.411
12	218.30	213.74	4.5624	201.691	225.784
13	226.80	222.48	4.3166	210.171	234.796
14	237.90	233.68	4.2172	220.194	247.172
15	225.80	228.70	-2.8993	216.558	240.841
16	232.40	231.89	.5059	219.617	244.171
17	241.70	238.10	3.6045	225.979	250.212
18	249.20	244.50	4.7044	232.121	256.870
19	261.30	256.93	4.3731	243.896	269.958
20	248.90	251.46	-2.5552	237.990	264.920
21	226.80	234.38	-7.5843	220.907	247.861
22	225.60	224.22	1.3849	210.780	237.651
23	228.80	226.35	2.4454	213.436	239.273
24	239.60	233.14	6.4590	220.406	245.876
25	244.70	241.92	2.7793	228.078	255.763
26	245.80	249.18	-3.3846	236.138	262.231
27	269.40	276.59	-7.1854	261.302	291.869

Example Gasoline Sales in the US-Market (EViews)

EViews: US Gasoline Market

Dependent Variable: G
Method: Least Squares
Date: 11/12/02 Time: 18:14
Sample: 1960 1986
Included observations: 27

Variable	Coefficient	Std. Error	t-Statistic	Prob.
Y	0.029679	0.004543	6.533306	0.0000
PG	-7.601784	8.500123	-0.894315	0.3808
PUC	-45.07951	15.34882	-2.937001	0.0076
PD	-181.5067	101.8276	-1.782490	0.0885
PS	281.3091	59.43764	4.732845	0.0001

R-squared	0.953216	Mean dependent var	207.0333
Adjusted R-squared	0.944710	S.D. dependent var	43.79893
S.E. of regression	10.29879	Akaike info criterion	7.667505
Sum squared resid	2333.430	Schwarz criterion	7.907475
Log likelihood	-98.51132	F-statistic	112.0624
Durbin-Watson stat	0.628661	Prob(F-statistic)	0.000000

Example Gasoline Sales in the US-Market (SPSS)

Included/ removed variables [b]

Model	Included variables	Removed variables	Method
1	PS, Y, PG, PUC, PD [a]	,	Insert

a. All wanted variables were included

b. Dependent Variable: G

Model summary [b]

Model	R	R-squared	Adjusted R-squared	Standard error of estimator	Durbin-Watson statistic
1	,997 [a]	,993	,991	4,03858	,637

a. Independent variables: (Constant), PS, Y, PG, PUC, PD

b. Dependent Variable: G

ANOVA [b]

Model		Sum of squares	Df	Mean of squares	F	Significance
1	Regression	49534,487	5	9906,897	607,407	,000 [a]
	Residuals	342,513	21	16,310		
	Total	49877,000	26			

a. Independent variables: (Constant), PS, Y, PG, PUC, PD

b. Dependent Variable: G

Coefficients [a]

Model		Non-standardised Coefficients		Standardised Coefficients	T	Significance
		B	Standard Error	Beta		
1	(Constant)	-146,273	13,239		-11,048	,000
	Y	3,462E-02	,002	1,151	18,850	,000
	PG	-29,124	3,861	-,777	-7,544	,000
	PUC	-14,311	6,632	-,318	-2,158	,043
	PD	305,844	59,500	1,558	5,140	,000
	PS	-113,096	42,634	-,781	-2,653	,015

a. Dependent Variable: G

6.4 Return on Human Capital (Stata)

Example 4: Return on Human Capital (Stata)

Data base: Time use survey of the German Federal Statistical Office 2001/02, Time use diaries

(Merz: . use „X:\timeuse\Wha\zbe2001-02_eIJTUR wha 2009\wha_mnl_cols\zbe0102_merz stata.dta“)

Description

```
. sum wage lnwage experien woman if wage>0
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        wage |     19470    9.987384    6.254399   .3501401   96.72619
      lnwage |     19470    2.129235    .6176006  -1.049422   4.571884
    experien |     24885    25.67077      12.493          0         49
       woman |     26802    .5335423     .498883          0          1
```

Log wage Schätzung mit dummy woman

```
. regress lnwage woman experien exper2 if wage>0
      Source |       SS       df       MS              Number of obs =   18620
-------------+------------------------------           F(  3, 18616) = 1766.59
       Model |  1529.36323     3  509.787743           Prob > F      =  0.0000
    Residual |  5372.05879 18616  .288572131           R-squared     =  0.2216
-------------+------------------------------           Adj R-squared =  0.2215
       Total |  6901.42203 18619  .370665558           Root MSE      =  .53719

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       woman |  -.2635038   .0078969   -33.37   0.000    -.2789825   -.2480251
    experien |   .0686598   .0012007    57.18   0.000     .0663063    .0710132
      exper2 |  -.0011562   .0000252   -45.95   0.000    -.0012055   -.0011069
       _cons |   1.436838   .0140526   102.25   0.000     1.409293    1.464382
------------------------------------------------------------------------------
```

6.5 Daily Working Hour Arrangements (LIMDEP)

Example 5: Daily Working Hour Arrangements (LIMDEP)

Merz, Joachim, Böhm, Paul and Derik Burgert (2009), Timing and Fragmentation of Daily Working Hours Arrangements and Income Inequality – An Earnings Treatment Effects Approach with German Time Use Diary Data, in: electronic International Journal of Time Use Research, 6/2, 200-239

Data base: Time use survey of the German Federal Statistical Office 2001/02, Time use diaries

CLR-Example for daily working hours arrangement (Daily core time/ non-fragmented)

(Estimation in Merz et al. 2009 with a more sophisticated model than CLR)

```
+----------------------------------------------------------------------+
| Ordinary    least squares regression    Weighting variable = none     |
| Dep. var. = HWORK    Mean=   7.534279684    , S.D.=   2.629386890     |
| Model size: Observations =    4896, Parameters =  33, Deg.Fr.=   4863 |
| Residuals:  Sum of squares= 19358.74146    , Std.Dev.=        1.99520 |
| Fit:        R-squared=  .427974, Adjusted R-squared =          .42421 |
| Model test: F[ 32,   4863] =  113.70,    Prob value =          .00000 |
| Diagnostic: Log-L = -10312.4512, Restricted(b=0) Log-L =  -11679.8345 |
|             LogAmemiyaPrCrt.=    1.388, Akaike Info. Crt.=      4.226 |
| Autocorrel: Durbin-Watson Statistic =   1.69951,   Rho =       .15024 |
+----------------------------------------------------------------------+
+---------+-------------+---------------+--------+---------+----------+
```

Variable	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]	Mean of X
Constant	3.567611749	.44055231	8.098	.0000	
AGE	.1535422414	.20946408E-01	7.330	.0000	41.287786
AGE2100	-.1785802420	.25470817E-01	-7.011	.0000	18.078854
WOMAN	-.1881477754	.75960614E-01	-2.477	.0133	.45894608
MARRIED	-.3137882175	.94533900E-01	-3.319	.0009	.66768791
ELEMENTA	.4004220803E-01	.12289913	.326	.7446	.25776144
INTERMED	.1388616499E-01	.11407012	.122	.9031	.40257353
SUPPER	-.1653603737E-01	.12186310	-.136	.8921	.37520425
SPECVC	.1248644419	.66693555E-01	1.872	.0612	.36254085
UNIVERSI	-.5648683006E-01	.10486537	-.539	.5901	.15706699
FREIBERU	-.3620200726	.18035780	-2.007	.0447	.29207516E-01
ENTREPRE	.4617177965	.16502991	2.798	.0051	.34313725E-01
SZEITHH	-.1222957281E-01	.31428768E-03	-38.912	.0000	121.94444
ZEITKIND	-.1024598376E-01	.89770811E-03	-11.413	.0000	15.998775
PHACTHH	-.5562468704E-02	.44656589E-02	-1.246	.2129	3.3926879
PHACTCHI	-.9916299012E-02	.56620624E-02	-1.751	.0799	.71531863
PHACTCAR	.5756283161E-02	.20042920E-01	.287	.7740	.22824755
WORKDAY	2.823572997	.10441312	27.042	.0000	.91462418
WORKDIST	.2030814589E-02	.13455994E-02	1.509	.1312	25.428922
SECONJOB	-.7316559212E-01	.84749055E-01	-.863	.3880	.13582516
WAGE	-.2817667992E-01	.12971176E-01	-2.172	.0298	10.389687
WAGE2100	.3722019539E-01	.27623139E-01	1.347	.1778	1.4203866
INDUSTRY	.3236532474	.97534220E-01	3.318	.0009	.27920752
SERVICES	.1294761324	.85788969E-01	1.509	.1312	.57843137
PARTNORW	.1655870401E-02	.20153711E-02	.822	.4113	18.745347
PERSHH	.6766849025E-01	.30586785E-01	2.212	.0269	3.1921977
YOUNGKID	.1717078335	.15310445	1.122	.2621	.13786765
YKIDPWOR	.5089898187E-01	.16927709	.301	.7637	.90482026E-01
OWNHOUSE	.2779368117E-01	.64883264E-01	.428	.6684	.63378268
RESINCOM	-.7282643633E-04	.16133670E-04	-4.514	.0000	2781.6883
HHPASHH	-.2581328861E-02	.45603134E-02	-.566	.5714	1.9031454
HHPASCHI	.3701933395E-02	.52305136E-02	.708	.4791	1.6376430
OST	.6321685326	.77991805E-01	8.106	.0000	.21997549

(Note: E+nn or E-nn means multiply by 10 to + or -nn power.)

6.6 Happiness (Ferrer-i-Cabonell and Frijters 2004)

Example 6: Happiness (Ferrer-i-Carbonell and Frijters 2004)

The examples so far have shown the results as computer protocols. In the final publication only central regression results (like R2, coefficients, t-values, n) similar to the following example by Ferrer-i-Carbonell und Frijters 2004 are published.

The paper analyzes how different estimation methods (OLS; Fixed-Effects-Estimation) influences the results of Happiness/Satisfaction (General Satisfaction, GS).

Only minor differences are shown between OLS with and without controls. However larger differences are given by individuell fixed effects:

$$GS_{it} = x_{it}\beta + \varepsilon_{it}. \quad (1)$$

$$GS_{it} - GS_{it-1} = \Delta x_{it}\beta + \Delta\varepsilon_{it}, \quad (2)$$

Table 1

The Determinants of Cardinal General Satisfaction for West German Workers in the GSOEP

	Model (1)				Model (2)	
	OLS on GS		OLS on GS		Fixed-eff. OLS	
	Estimate	t-val	Estimate	t-val	Estimate	t-val
Age	−0.03	5.8	−0.05	10.0		
Age × age	0.0005	7.5	0.0007	11.3	−0.0006	6.5
ln(household income)	0.34	18.7	0.38	18.6	0.11	4.3
Number of children	−0.07	5.5	−0.05	5.2	0.01	0.9
Steady partner (1 = yes)	0.13	4.8	0.23	12.3	0.07	2.4
Subjective health	0.54	93.8	0.39	97.3	0.32	44.1
Controls	No		Yes		No	
Number of individuals	7,806		7,806		6,664	
R^2	0.25		0.26		0.09	
Number of cases	30,569		30,569		21,104	

Time-dummies were present in all estimates but are not shown. The number of individuals is lower for the fixed-effects because they require at least 2 observations per individual.
The controls for the OLS on GS contains the following variables: education, working hours, gender, and the number of adults in the household.

Source:

Ferrer-i-Carbonell A. and P. Frijters (2004), How important is Methodology for the estimates of determinants of happiness? in: The Economic Journal, Vol. 114, No. 497, 641–659.

References

Anderson, David, R., Sweeney, Dennis, J., Williams, Thomas, A., Freeman, J. and E. Shoesmith (2010), Statistics for Business and Economics, Thomson Publisher, London, United Kingdom.

Bauer, Th.K., Fertig, M. and Chr.M. Schmidt (2009), Empirische Wirtschaftsforschung – Eine Einführung, Springer-Verlag, berlin Heidelberg.

Fahrmeir, L., Kneib, Th. and St. Lang (2009), Regression – Modelle, Methoden und Anwendungen, 2. Auflage, Springer Heidelberg/New York.

Ferrer-i-Carbonell A. and P. Frijters (2004), How important is Methodology for the estimates of determinants of happiness? in: The Economic Journal, Vol. 114, No. 497, 641–659

Greene, W. (2008, 2003, 2000, 1997), Econometric Analysis, Sixth, Fifth, Fourth, Third Edition, New York/London.

Hübler, O. (2005), Einführung in die empirische Wirtschaftsforschung, Verlag Oldenbourg, Oldenbourg.

Merz, J. (2009), Statistik I - Deskription, Skriptum zur Vorlesung, 9. verbesserte Auflage, Lüneburg.

Merz, J. (2010), Statistik II - Wahrscheinlichkeitsrechnung und induktive Statistik, Skriptum zur Vorlesung, 9. verbesserte Auflage, Lüneburg.

Merz, J. (2015), Empirische Wirtschaftsforschung – Regressionsanalyse, Skriptum zur Vorlesung, Lüneburg.

Merz, J. and H. Stolze (2010), FFB e-learning: Parametertests, Lüneburg (www.leuphana.de/ffb)

Merz, J. and H. Stolze (2010a), FFB e-learning: Lineare Regression - Deskriptives Modell, Lüneburg (www.leuphana.de/ffb).

Merz, J. and H. Stolze (2010b), FFB e-learning: Lineare Regression – Stochastisches Modell, Lüneburg (www.leuphana.de/ffb).

Merz, J., Böhm, P. and D. Burgert (2009), Timing and Fragmentation of Daily Working Hours Arrangements and Income Inequality – An Earnings Treatment Effects Approach with German Time Use Diary Data, in: electronic International Journal of Time Use Research, 6/2, 200-239.

Studenmund, A.H. (2006), Using Econometrics – A Practical Guide, Fifth Edition, Pearson/Addison Wesley, Boston – Montreal.

von Auer, L. (2003), Ökonometrie – Eine Einführung, Springer-Verlag, Berlin Heidelberg.

Wooldridge, J.M. (2002), Econometric Analysis of Cross Section and Panel Data, The MIT Press, Cambridge, Mass.

Wooldridge, J.M. (2009, 2006), Introductory Econometrics, A Modern Approach, Third Edition, Thomson, South-Western, Canada.