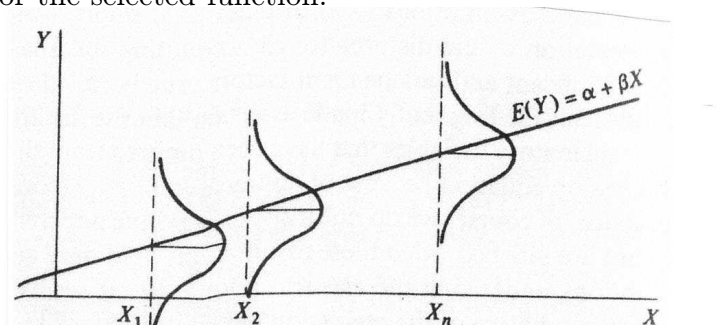
10 Simple linear regression

This chapter is concerned with relations among variables like demand and supply relations, cost functions, production functions and many others. Deterministic relations characterized as a function $y = f(x)$ are usually from the natural sciences. A relation between X and Y is *deterministic* if each element of the domain is paired off with just one element of the range. In economic situations the deterministic relations are very rare and we usually deal with stochastic (probabilistic) relations which ae more realistic for most real-world situations. A relation between X and Y is said to be *stochastic* if for each value of X there is a whole probability distribution of values of Y. Thus for any given value of X the variable Y may assume some specific value (or fall within some specific interval) with a probability smaller than one and greater than zero. Its value is effected by a random disturbance. Regression analysis is dealing with stochastic relations. It is aimed a) to determine a form of a function which may describe the dependence of Y on X and b) to estimate the parameters for the selected function.



ad a)
To determine the form of the function we may start from logical analysis (i.e. to follow some economic theory) or it could be estimated from the two-dimensional diagram (scatter plot). (This way can be used only in the case where the dependent (response) variable is a function of just one independent (explanatory, predictor) variable.)
The list of commonly used forms of regression functions follows:

- regression line: $E(Y|x) = \beta_0 + \beta_1 x$
- regression parabola: $E(Y|x) = \beta_0 + \beta_1 x + \beta_2 x^2$
- regression polynomial of degree p: $E(Y|x) = \beta_0 + \beta_1 x + \ldots + +\beta_p x^p$
- regression hyperbola: $E(Y|x) = \beta_0 + \beta_1 \frac{1}{x}$
- regression logarithmic function: $E(Y|x) = \beta_0 + \beta_1 \ln x$

Each of listed regression functions represents simple linear regression function. (The term linear regression function is used if the function is linear with respect to the parameters $\beta_0,\ \beta_1,\ \beta_2, \ldots$. It is said to be simple if the dependent variable is a function of just one independent variable. Otherwise it is said to be multiple.)
ad b)
The unknown parameters $\beta_0,\ \beta_1,\ \beta_2, \ldots$ are estimated so that to fit the data set of n pairs of observed values
$(x_1, y_1), \ldots, (x_n, y_n)$. To estimate the parameters the least square estimation is commonly used method.

10.1 Specification of the classical simple linear regression model

A model consist of the regression equation and the basic assumptions. Let us begin with the equation:

•$Y = \beta_0 + \beta_1 f_1(x) + \ldots + \beta_p f_p(x) + \varepsilon$ where:

Y is a dependent random variable which is observable
x is an independent non-stochastic variable which is observable
ε is a random error which accounts for the random factors and is unobservable
$\beta_0+\beta_1 f_1(x)+\ldots+\beta_p f_p(x)$ is a theoretic regression function with unknown parameters $\beta_0, \beta_1, \ldots, \beta_p$

For n observations the regression equation can be expressed as follows:
$y_1 = \beta_0 + \beta_1 f_1(x_1) + \ldots + \beta_p f_p(x_1) + \varepsilon_1$
$\vdots$
$y_i = \beta_0 + \beta_1 f_1(x_i) + \ldots + \beta_p f_p(x_i) + \varepsilon_i$
$\vdots$
$y_n = \beta_0 + \beta_1 f_1(x_n) + \ldots + \beta_p f_p(x_n) + \varepsilon_n$
The subscript $i = 1, \ldots, n$ refers to the ith observation. Observations on X and Y can be made over time, in which case we speak of "time-series data" or they can be made over individuals, objects, or geographical areas, in which case we speak of cross-section data.

•Assumptions for the random error $\varepsilon_i,\ i = 1, \ldots, n$ are

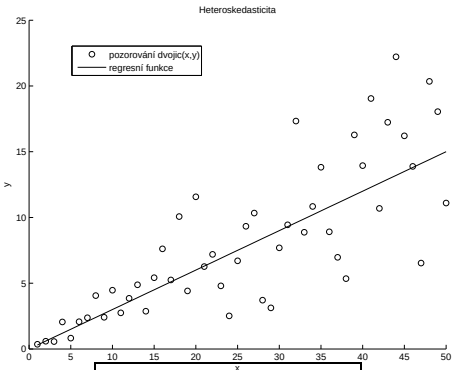
a) $E(\varepsilon_i) = 0$ [zero mean for errors which are not systematic]

b) $D(\varepsilon_i) = \sigma^2 > 0$ [each observation is done with the equal precision]

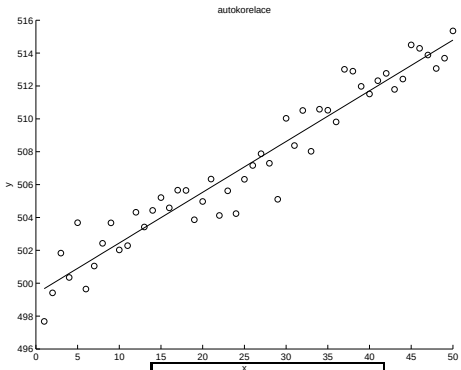
c) $C(\varepsilon_i, \varepsilon_j) = 0$ pro $i \neq j$ [there is no linear relationship between the errors]

d) $\varepsilon_i \sim N(0, \sigma^2)$ [errors are normally distributed]

Violations of some basic assumptions are shown in following pictures. In the first one there is a violation of the assumption b) and we speak of heteroskedasticity of random errors; in the second picture there is a violation of the assumption c) and then we speak of autocorrelation of random errors.

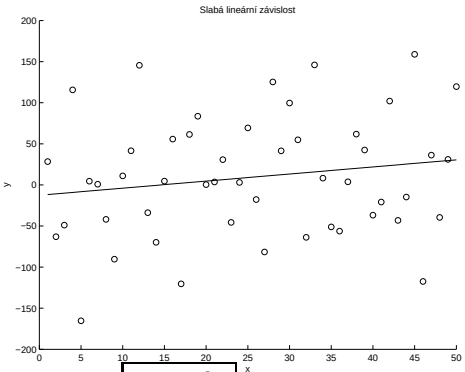


Heteroskedasticity

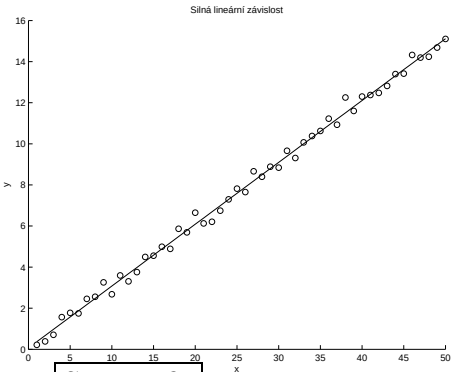


Autocorrelation

The following pictures perform low and strong linear dependence under basic assumptions:



Low fit



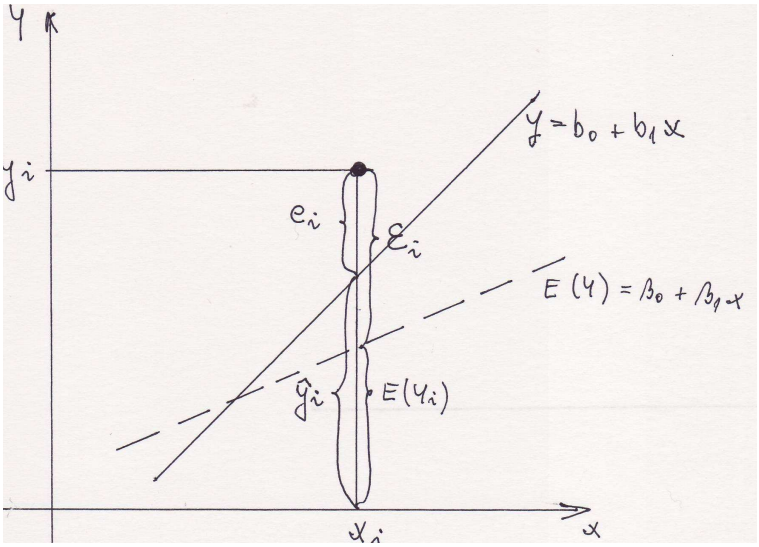
Strong fit

Since the mathematical form of the relation is specified the unknown parameters $\beta_0, \beta_1 \ldots, \beta_p$ should be estimated.

10.2 The least square estimators of regression parameters and the notation

$b_0, b_1, \ldots, b_p$	estimators of regression parameters $\beta_0, \beta_1, \ldots, \beta_p$
$b_0 + b_1 f_1(x) + \ldots + b_p f_p(x)$	empirical (sample) regression function
$\hat{y}_i = b_0 + b_1 f_1(x_i) + \ldots + b_p f_p(x_i)$	regression estimate of the ith value of the random variable Y
$e_i = y_i - \hat{y}_i$	ith residual
$S_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$	residual sum of squares
$s^2 = \frac{S_E}{n-p-1}$	estimator of the variance σ^2
$S_R = \sum_{i=1}^{n} (\hat{y}_i - m_2)^2; \quad m_2 = \frac{1}{n} \sum_{i=1}^{n} y_i$	regression sum of squares
$S_T = \sum_{i=1}^{n} (y_i - m_2)^2;$	total sum of squares [It holds: $S_T = S_R + S_E$]
$ID^2 = \frac{S_R}{S_T} = 1 - \frac{S_E}{S_T}$	coefficient of determination [$ID^2 \in (0, 1)$]

[Coefficient of determination is a measure of "goodness of fit"; it is simply the proportion of the variation of Y that can be attributed to the variation of X and describes how well the sample regression function fits the observed data. A zero value of ID^2 indicates the poorest and a unit value the best fit that can be attained.]



10.3 The method of least squares

The purpose of the least-square method is to find estimators $b_0, b_1, \ldots, b_p$ of regression parameters $\beta_0, \beta_1, \ldots, \beta_p$ so that the sum of squares of residuals is as little as possible. (The regression estimates fit the data "best".) ThusTedy

$$S(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - \beta_0 + \beta_1 f_1(x_i) + \ldots + \beta_p f_p(x_i)]^2 \rightarrow \ min$$

Thus we have to minimize the function $S(\beta_0, \beta_1, \ldots, \beta_p)$, which is dependent only on unknown parameters of the regression model. The procedure how to do it follows:

1. Differentiate $S(\beta_0, \beta_1, \ldots, \beta_p)$ with respect to each regression parameter.
2. Equate each derivatives to zero. This leads to the system of n equations of n unknown variables. These equations are generally known as the *least squares normal equations.*
3. Solving the least squares normal equations we obtain wanted estimators $b_0, b_1, \ldots, b_p$ of regression parameters $\beta_0, \beta_1, \ldots, \beta_p$.

Then the least squares normal equations have the form:

$$\begin{array}{lllllllll}
\beta_0 \sum 1 & + & \beta_1 \sum f_1 & + & \beta_2 \sum f_2 & + & \ldots & + & \beta_p \sum f_p & = & \sum y_i \\
\beta_0 \sum f_1 & + & \beta_1 \sum f_1^2 & + & \beta_2 \sum f_1 f_2 & + & \ldots & + & \beta_p \sum f_1 f_p & = & \sum y_i f_1 \\
\vdots & & & & & & & & & & \vdots \\
\beta_0 \sum f_p & + & \beta_1 \sum f_p f_1 & + & \beta_2 \sum f_p f_2 & + & \ldots & + & \beta_p \sum f_p^2 & = & \sum y_i f_p
\end{array}$$

where the symbol $\sum$ states for $\sum_{i=1}^{n}$ and te symbol $\sum f_j$ states for $\sum_{i=1}^{n} f_j(x_i)$.

$\beta_0, \beta_1, \ldots, \beta_p$, which solve the least squares normal equations are denoted as $b_0, b_1, \ldots, b_p$.

Example 10.4

Considering regression line find the estimators b_0, b_1 of the parameters β_0, β_1. (And the basic assumptions of the classical regression model are satisfied.)

Solution

The estimates $b_0,\ b_1$ can be obtained from the least squares normal equations :

$$\begin{array}{lllll}
b_0 \sum_{i=1}^{n} 1 & + & b_1 \sum_{i=1}^{n} x_i & = & \sum_{i=1}^{n} y_i \\
b_0 \sum_{i=1}^{n} x_i & + & b_1 \sum_{i=1}^{n} x_i^2 & = & \sum_{i=1}^{n} y_i x_i
\end{array} .$$

The solution is

$$b_0 = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i x_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad b_1 = \frac{n \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Thus the estimated sample regression line is $\hat{y} = b_0 + b_1 x$.
(Notice that $b_0,\ b_1$ are random variables; they are depending on realizations $(x_i,\ y_i)$,while parameters $\beta_0,\ \beta_1$ are constants.)

10.5 The matrix notation of classical linear regression model and its solution

•Model: $y_i = \beta_0 + \beta_1 f_1(x_i) + \ldots + \beta_p f_p(x_i) + \varepsilon_i, \quad i = 1, \ldots, n$

can be expressed in matrix notation: $\boxed{\mathbf{y} = \mathbf{X}\beta + \varepsilon}$, thus

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & f_1(x_1) & \ldots & f_p(x_1) \\ 1 & f_1(x_2) & \ldots & f_p(x_2 1) \\ \vdots & & & \vdots \\ 1 & f_1(x_n) & \ldots & f_p(x_n) \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

A notation:

$\mathbf{y}$	a column vector of observed values of dependent random variable Y
$\mathbf{X}$	a matrix of observed values of regressors
	[we assume the rank: $h(X) = p + 1 < n$, the columns of $\mathbf{X}$ are linear independent]
β	a column vector of regression parameters
ε	a column vector of residuals

•The assumptions of the model can be rewritten as follows: $\boxed{\varepsilon \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})}$

As it was written in 10.3 the estimators $b_0, b_1, \ldots, b_p$ of regression parameters $\beta_0, \beta_1, \ldots, \beta_p$ can be obtained by solving least square normal equations. These can be expressed in matrix notation as follows:

$\mathbf{X'X}\beta = \mathbf{X'y}$	the least square normal equations
$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y}$	the lest square estimators (LSE or frequently OLS - ordinary least squares)
$\hat{\mathbf{y}} = \mathbf{Xb}$	a vector of regression estimators
$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$	a vector of residuals

10.6 Properties of the least square estimators $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y}$

1. the estimator $\mathbf{b}$ is linear; it is a linear combination of the random vector $\mathbf{y}$
2. the estimator $\mathbf{b}$ is unbiased; it is true that $E(\mathbf{b}) = \beta$
3. the estimator $\mathbf{b}$ has a variance-covariance matrix $var\ (\mathbf{b}) = \sigma^2(\mathbf{X'X})^{-1}$
4. the estimator $\mathbf{b}$ is normally distributed with mean vector β and variance-covariance matrix $var\ (\mathbf{b}) = \sigma^2(\mathbf{X'X})^{-1}$, thus $\mathbf{b} \sim N_{p+1}(\beta, \sigma^2(\mathbf{X'X})^{-1})$; the normality follows from $\varepsilon \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$ and the first property.
5. the estimator $\mathbf{b}$ is the best linear unbiased estimator of the vector β. (BLUE)

Remark 10.7
The last property is known as *Gauss-Markov theorem.*
The "best" means that if $\mathbf{b}^*$ is any other linear unbiased estimator, then $var\ (\mathbf{b}) \leq var\ (\mathbf{b}^*)$. [$var\ (\mathbf{b}^*) - var\ (\mathbf{b})$ is a positive semi-definite matrix.]

As we know the distribution of the vector of estimators $\mathbf{b}$ we may follow with statistical inferences about regression parameters β. But the parameter σ, which is involved in variance-covariance matrix $var\ (\mathbf{b})$, is unknown. Thus we have to obtain its estimator and consequently estimators of variances of elements $\mathbf{b}$.

Variance-covariance matrix has the form $var\ \mathbf{b} = \begin{pmatrix} var(b_0) & cov(b_0,b_1) & \dots & cov(b_0,b_p) \\ cov(b_1,b_0) & var(b_1) & \dots & cov(b_1,b_p) \\ \vdots & & \ddots & \vdots \\ cov(b_p,b_0) & cov(b_p,b_1) & \dots & var(b_p) \end{pmatrix} =$

$\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

Thus the variances $D(b_j),\ j = 0,1,\dots,p$ are represented by diagonal elements of the matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.
Let us recall (from 10.2) that $s^2 = \frac{S_E}{n-p-1}$ is an unbiased estimator of the parameter σ^2. Thus the matrix $s^2(\mathbf{X}'\mathbf{X})^{-1}$ estimates variance-covariance matrix $var(\mathbf{b})$ and its diagonal elements estimate the variances $D(b_j)$. The following notation is used :

v_{jj}	$j-$th diagonal element of the matrix $(\mathbf{X}'\mathbf{X})^{-1}$
$s_{b_j} = s \cdot \sqrt{v_{jj}}$	a standard error of b_j

10.8 The confidence intervals for the regression parameters
The statistic $T_j = \frac{b_j - \beta_j}{s_{b_j}}$ follows t-distribution $t(n-p-1)$ for $j = 0,1,\dots,p$.
Thus considering $100(1-\alpha)\%$ confidence interval for β_j its limits are calculated as follows: $b_j \pm s_{b_j} t_{1-\alpha/2}(n-p-1)$

10.9 Test of significance of single parameters (separate t-tests)
At the significance level α we are testing for $j = 0,1,\dots,p$

$H_0: \quad \beta_j = 0$ versus $H_1: \quad \beta_j \neq 0.$

The null hypothesis asserts that the vector $\mathbf{y}$ is not influenced by the $j-$th column of the matrix $\mathbf{X}$. Rejecting the null it is concluded that the parameter β_j is relevant in our model.

The test statistic $T_j = \frac{b_j}{s_{b_j}}$ follows a distribution $t(n-p-1)$ if H_0 is true.
The critical region follows: $W = (-\infty;\ -t_{1-\alpha/2}(n-p-1)\rangle \cup \langle t_{1-\alpha/2}(n-p-1);\ \infty)$

10.10 Test of significance of regression (the overall F-test)
At the significance level α we are testing:

$H_0: \quad (\beta_1,\beta_2,\dots,\beta_p) = (0,0,\dots,0)$ versus $H_1: \quad (\beta_1,\beta_2,\dots,\beta_p) \neq (0,0,\dots,0).$

H_0 is a more extensive hypothesis that none of the explanatory variables has an influence on Y. If H_0 is true then the variation of Y from observation to observation is not affected by changes in any one of the explanatory variables, but is purely random, $Y = \beta_0 + \varepsilon$.

The test statistic: $F = \frac{S_R/p}{S_E/(n-p-1)} \sim F(p, n-p-1)$, if H_0 is true.
The critical region follows: $W = \langle F_{1-\alpha}(p, n-p-1); \infty)$

The $F-$test results are usually performed in ANOVA table:

Sources of variability	sum of squares	degrees of freedom	mean squares	test statistic
regression model	S_R	p	S_R/p	$\frac{S_R/p}{S_E/(n-p-1)}$
error	S_E	$n-p-1$	$S_E/(n-p-1)$	
total	S_T	$n-1$		