743- Regression and Time Series

Mamikon S. Ginovyan

© 2012 Mamikon Ginovyan

The Prediction Problem
(Statistical Solution)

© 2012 Mamikon Ginovyan

2. The Least Squares Method (LSM)

The **Least Squares Method** (**LSM**) is a general method of estimation of **unknown parameters,** which is of great importance in many areas of statistics such as

- **the analysis of variance** (**ANOVA**) and
- **regression theory.**

© 2012 Mamikon Ginovyan

General Regression Model

Suppose that we have observations $Y_1, Y_2, ..., Y_n$, and suppose that we can write these observations in the form

$$(1) \qquad Y_i = f_i(q_1, \cdots, q_p) + e_i$$
$$= \text{"}\mathbf{Deterministic}\text{"} + \text{"}\mathbf{Random}\text{"}, \quad i = \overline{1, n},$$

where $f_i, i = \overline{1, n}$, are <u>**known functions,**</u> and the real numbers $q_1, \cdots, q_p$ are <u>**unknown parameters**</u> of interest that we want **to estimate.**

We assume that $q = (q_1, \cdots, q_p) \in \Theta \subset R^p$.

© 2012 Mamikon Ginovyan

General Regression Model

Suppose also that the RV's e_i satisfy the following conditions:

(2) $E(e_i) = 0, \quad i = \overline{1, n}.$

(3) $Var(e_i) = s^2, \quad 0 < s^2 < \infty, \quad i = \overline{1, n} \qquad (s^2$ is known).

(4) $Cov(e_i, e_j) = 0, \quad i, j = \overline{1, n}, \quad i \neq j.$

$$Cov(e_i, e_j) = d_{ij} s_j^2 = \begin{cases} s_j^2, & \text{if } i = j \\ 0, & \text{if } i \neq j, \end{cases} \qquad i, j = \overline{1, n}.$$

© 2012 Mamikon Ginovyan

General Regression Model

- **Remark 1.**

  Conditions (2) - (4) hold when $e_i, i = \overline{1, n}$ are ***i.i.d.*** RV's with mean $m = 0$ and variance s^2: $0 < s^2 < \infty$.

Most models in which the LSM is applied are of this type.

An important special case occurs when the RV's $e_i, i = \overline{1, n}$, form a sample (***i.i.d.***) from a $N(0, s^2)$ population.

© 2012 Mamikon Ginovyan

Least Squares Estimate

- **<u>The idea</u>** behind LSM is the following.

Consider the random vector $\underline{Y} = (Y_1, \cdots, Y_n)$ as a random point in R^n.

Since $E(e_l) = 0, \quad i = \overline{1, n},$

the "**expected value**" of $\boldsymbol{Y}$ is the vector

$$(5) \qquad f(q) = (f_1(q), \cdots, f_n(q)),$$

that is, $E[\underline{Y}] = f(q) \Leftrightarrow E[Y_i] = f_i(q), \quad i = \overline{1, n},$

where $q = (q_1, \cdots, q_p).$

© 2012 Mamikon Ginovyan

Least Squares Estimate

The LSM merely says that as an estimate of q we should to take that point $\hat{q} = (\hat{q}_1, \cdots, \hat{q}_p)$
which makes the **expected value vector** $f(q)$ **as close as possible** to the **observed point** $\boldsymbol{Y}$.
That is, if we observe

$$Y_1 = y_1, \cdots, Y_n = y_n,$$

$\hat{q} = (\hat{q}_1, \cdots, \hat{q}_p)$ should **minimize** the sum of the squares of the distances from the data points (y_i) to the expected value points $(f_i(q))$,
that is, the **function** $\boldsymbol{L}$ defined by

$$(6) \quad L = \sum_{i=1}^{n} (y_i - f_i(q))^2, \quad q = (q_1, \cdots, q_p) \in \Theta.$$

© 2012 Mamikon Ginovyan

Least Squares Estimate

The estimate $\hat{q} = (\hat{q}_1, \cdots, \hat{q}_p)$ is then called a **Least Squares Estimate (LSE)**.

Thus, the **LSE** $\hat{q} = (\hat{q}_1, \cdots, \hat{q}_p)$ is defined by

$$(7) \quad \sum_{i=1}^{n} (y - f_i(\hat{q}))^2 = \min_{q \in \Theta} \sum_{i=1}^{n} (y - f_i(q))^2 .$$

• **Remark 2 (Invariance Principle of LSE).**

If $g(q)$ is some function of q, then the **LSE** of $h = g(q)$ is

$\hat{h} = g(\hat{q})$.

© 2012 Mamikon Ginovyan

Least Squares Estimate

How to find the **LSE** $\hat{q}$?

That is, how to solve the **minimization problem (7)**?

We use **calculus** methods.

The procedure of finding LSE's

• **First,** if the functions $f_i, i = \overline{1, n},$ are **differentiable** and the range of $(f_1, \cdots, f_n)$ is **closed** in R^n, then the **LSE** $\hat{q}$ is **always defined**.

© 2012 Mamikon Ginovyan

Least Squares Estimate

- **Second,** if the parameter set Θ is **open** in R^p,

it follows from vector calculus that the LSE $\hat{q}$ must satisfy the equations:

$$(8)\qquad \begin{cases} \dfrac{\partial}{\partial q_j} \sum_{i=1}^{n} [y - f_i(q)]^2 = 0 \\ j = \overline{1.p.} \end{cases} \qquad q = (q_1, \cdots, q_p) \in \Theta \subset R^p.$$

The equations (8) are called **normal equations**.

It is clear that (8) is equivalent to

$$(9)\qquad \begin{cases} \sum_{i=1}^{n} [y - f_i(q)] \dfrac{\partial}{\partial q_i} f_i(q) = 0 \\ j = \overline{1, p} \end{cases} \qquad q \doteq (q_1, \cdots, q_p) \in \Theta \subset R^p.$$

© 2012 Mamikon Ginovyan

Least Squares Estimate

Remark 3.

The system (9) is a system of **non-linear** equations, which is **difficult to solve**.

In the special case where the functions $f_i(q_1, \cdots, q_p)$ are **linear** in the parameters $q_1, \cdots, q_p$ the normal equations become a system of **linear** equations and may be solved **explicitly**.

This model with $e_i \sim IIDN(0, s^2)$, which is called **linear model**, we will considers in detail later.

© 2012 Mamikon Ginovyan

Examples

1. The Measurement model.

Consider the ease where ***p* =1**, and $f_1(q_1) = q_1$.

That is, the model is given by

$$Y_i = q_1 + e_i, i = \overline{1, n},$$

which is called **measurement model**.

Since $\dfrac{\partial f_1(q_1)}{\partial q_1} = 1,$

the system (9) becomes

$$\sum_{i=1}^{n}(Y_i - q_1) = 0 \Leftrightarrow \sum_{i=1}^{n} Y_i - nq_1 = 0 \Leftrightarrow \hat{q}_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}.$$

Thus, in this case, the **LSE** $\hat{q}_1$ is the sample mean $\overline{Y}$.

© 2012 Mamikon Ginovyan

The Linear Regression Model

- **Statistical solution of MSE Linear prediction problem.**

Consider now the case where $\boldsymbol{p} = \mathbf{2}$,

(10) $f_i(q_1, q_2) = q_1 + q_2 x_i, i = \overline{1, n},$

so, our model is the **linear regression model**

(11) $Y_i = q_1 + q_2 x_i + e_i, i = \overline{1, n},$

where $e_i, i = \overline{1, n},$ are independent,

$$E(e_i) = 0, \; Var(e_i) = s^2, \; 0 < s^2 < \infty.$$

© 2012 Mamikon Ginovyan

The Linear Regression Model

The problem is:

Find **LSE**'s of unknown $q_1 \quad and \quad q_2$.

The following notations are in common used.

$$S_{xx} = S_x^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^{\;2} - \frac{1}{n} (\sum_{n=1}^{n} x_i)^2,$$

(12) $$S_{yy} = S_y^2 = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^{\;2} - \frac{1}{n} (\sum_{n=1}^{n} y_i)^2,$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n} (\sum_{n=1}^{n} x_i)(\sum_{n=1}^{n} y_i).$$

© 2012 Mamikon Ginovyan

The Linear Regression Model

Solution: By (10) we have

$$\frac{\partial}{\partial q_1} f_i(q_1, q_2) = \frac{\partial}{\partial q_1}[q_1 + q_2 x_i] = 1$$

$$\frac{\partial}{\partial q_2} f_i(q_1, q_2) = \frac{\partial}{\partial q_2}[q_1 + q_2 x_i] = x_i.$$

So, in this case the **normal equations** are

$$\begin{cases} \sum_{i=1}^{n} [y_i - f_i(q)] \frac{\partial}{\partial q_j} f_i(q) = 0 \\ j = 1, 2 \end{cases} \Leftrightarrow \begin{cases} \sum_{i=1}^{n} [y_i - q_1 - q_2 x_i] = 0 \\ \sum_{i=1}^{n} x_i [y_i - q_1 - q_2 x_i] = 0 \end{cases}$$

© 2012 Mamikon Ginovyan

The Linear Regression Model

$$\Leftrightarrow \begin{cases} \sum_{i=1}^{n} y_i = n \cdot q_1 + q_2 \sum_{i=1}^{n} x_i & (13) \\ \sum_{i=1}^{n} x_i y_i = q_1 \sum_{i=1}^{n} x_i + q_2 \sum_{i=1}^{n} x_i^2 & (14) \end{cases}.$$

To solve this system of equations for q_1 and q_2, we multiply the first equation by $\sum x_i$ and the second by $\boldsymbol{n}$ to obtain

$$\Leftrightarrow \begin{cases} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i = nq_1 \sum_{i=1}^{n} x_i + q_2 (\sum_{i=1}^{n} x_i)^2 & (15) \\ n \cdot \sum_{i=1}^{n} x_i y_i = nq_1 \sum_{i=1}^{n} x_i + q_2 n \sum_{i=1}^{n} x_i^2 & (16) \end{cases}.$$

© 2012 Mamikon Ginovyan

The Linear Regression Model

Now, subtracting (15) from (16) we get

$$n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i = q_2 \left[n\sum_{i=1}^{n} x_i^{\,2} - (\sum_{i=1}^{n} x_i)^2 \right] \qquad (17)$$

Using notation (12) we can write (17) as follows

$$n \cdot S_{xy} = q_2 \cdot nS_{xx} \Leftrightarrow q_2 = \frac{S_{xy}}{S_{xx}}.$$

Thus, the **LSE** $\hat{q}_2$ is given by

$$\hat{q}_2 = \frac{S_{xy}}{S_{xx}}. \qquad (18)$$

© 2012 Mamikon Ginovyan

The Linear Regression Model

Now, substituting (18) into (13) we find $\hat{q}_1$:

$$\hat{q}_1 = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{q}_2 \cdot \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{y} - \frac{S_{xy}}{S_{xx}} \cdot \overline{x}.$$

Thus, the **LSE** $\hat{q}_1$ is given by

$$\hat{q}_1 = \overline{y} - \frac{S_{xy}}{S_{xx}} \cdot \overline{x} \qquad (19)$$

Definition. The line

$$\hat{y} = \hat{q}_1 + \hat{q}_2 x \qquad (20)$$

with $\hat{q}_1$ and $\hat{q}_2$ as in (19) and (18) is called **sample (or estimated) regression line** or **line of best fit** of $\underline{y} = (y_1, \mathbf{K}, y_n)$ on $\underline{x} = (x_1, \mathbf{K}, x_n)$.

© 2012 Mamikon Ginovyan

Sample Regression Line

Geometric Interpretation

Given ***n*** points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
and a line $y = a + b\,x$.

If we measure the distance between a point (x_i, y_i) and a line $y = a + b\,x$:

$$d_i = \left| y_i - (a + b\,x_i) \right|,$$

then **what is the line that is "closest" to the given point?**

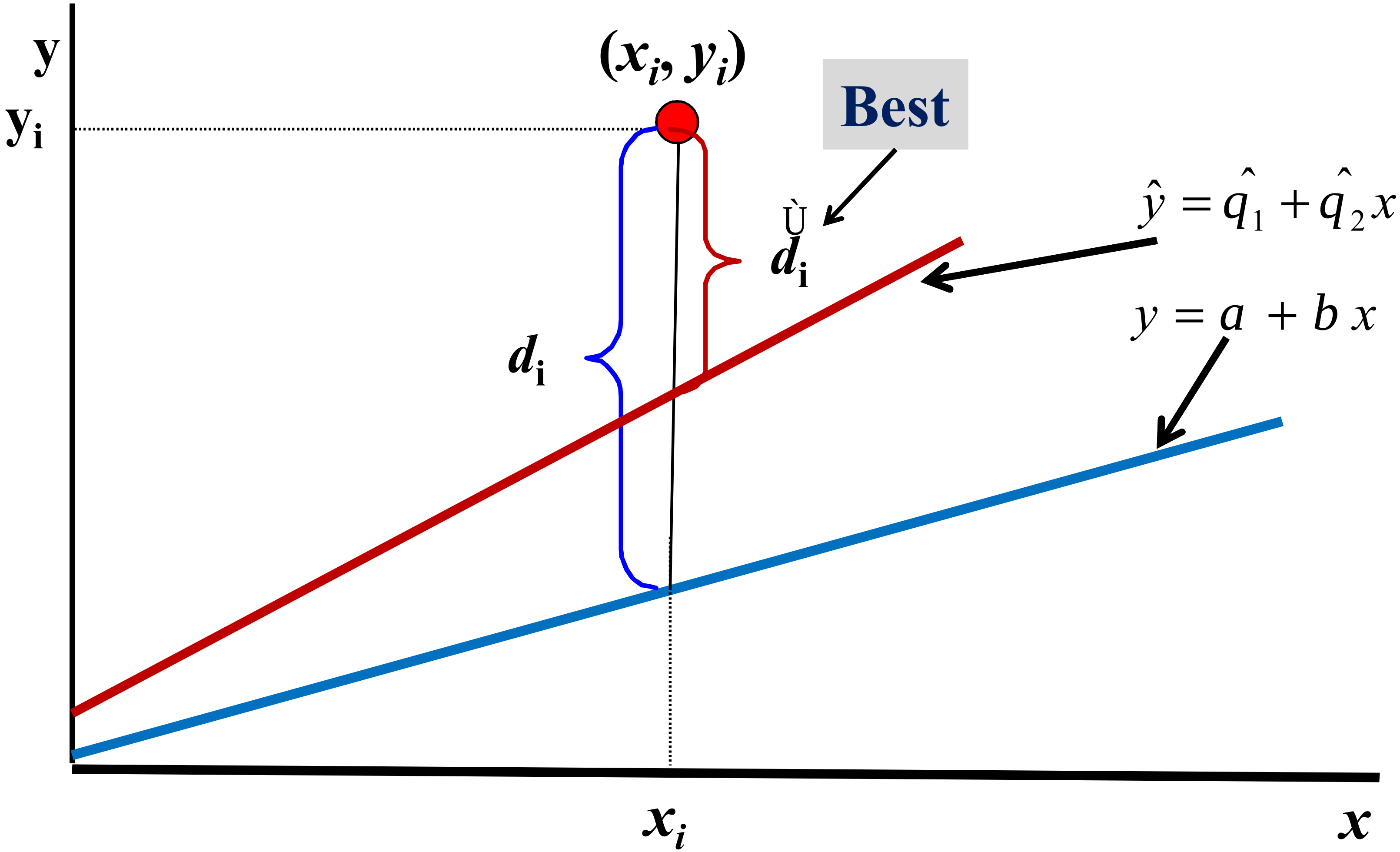
or, **what is the line that best fit given points?**

The **answer is:**

© 2012 Mamikon Ginovyan

Sample Regression Line. Geometric Interpretation

The answer is the **sample regression line defined by (20) with slope $\hat{q}_2$ and intercept $\hat{q}_1$ given by (18) and (19).**



Sample Regression Line

- **Statistical Interpretation: Statistical solution of prediction problem.**

Theorem (SLS).

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be $\boldsymbol{n}$ independent observations from the distribution of a random vector $(\boldsymbol{X}, \boldsymbol{Y})$. Then the **best linear predictor** $\hat{Y}$ of $\boldsymbol{Y}$ based on $\boldsymbol{X}$ is given by

$$\hat{Y} = \hat{a}_0 + \hat{b}_0 X, \qquad (21)$$

where

$$\hat{a}_0 = \overline{Y} - \frac{S_{XY}}{S_{XX}} \cdot \overline{X}, \qquad (22)$$

$$\hat{b}_0 = \frac{S_{XY}}{S_{XX}}. \qquad (23)$$

© 2012 Mamikon Ginovyan

Probabilistic Solution vs. Statistical Solution

Comparison of Probabilistic and Statistical Solutions of the Best Linear MSE Prediction Problem

- **Probabilistic Solution.**

The **unique** **best MSE-linear predictor** of $\boldsymbol{Y}$ given $\boldsymbol{X}$ is given by

$$\hat{Y} = a_0 + b_0 X$$

with

$$a_0 = E(Y) - b_0 E(X), \quad b_0 = \frac{Cov(X,Y)}{s^2(X)}.$$

© 2012 Mamikon Ginovyan

Probabilistic Solution vs. Statistical Solution

- **Statistical Solution.**

Given ***n*** independent observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, the **unique** **best MSE-linear predictor** of ***Y*** given ***X*** is given by

$$\hat{Y}_n = \hat{a}_0 + \hat{b}_0 X$$

with

$$\hat{a}_0 = \overline{Y} - \frac{S_{XY}}{S_{XX}} \cdot \overline{X}, \quad \hat{b}_0 = \frac{S_{XY}}{S_{XX}}.$$

© 2012 Mamikon Ginovyan

Probabilistic Solution vs. Statistical Solution

Since $\overline{X}, \overline{Y}, S_{XX}$ and S_{XY} are **statistical estimators** for $E(X), E(Y), s^2(X)$ and $Cov(X, Y)$, the **sample regression coefficients** $\hat{a}_0$ and $\hat{b}_0$ are statistical estimators for "**theoretical**" regression coefficients a_0 and b_0, respectively.

© 2012 Mamikon Ginovyan

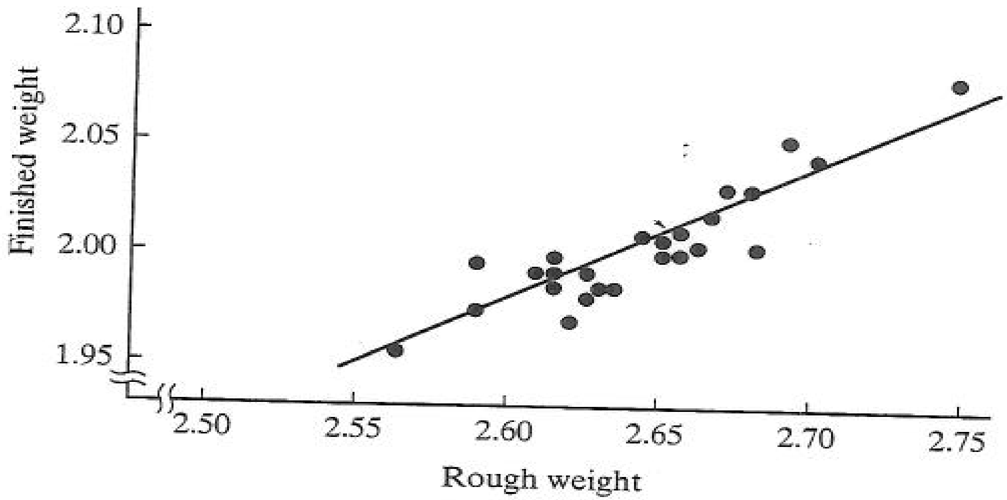
Example 1 (Manufacturer).

- A manufacturer of air conditioning units is having assembly problems due to the failure of a connecting rod to meet finished-weight specifications. Too many rods are being completely tooled, then rejected as overweight.
- To reduce that cost, the company's quality control department wants to quantify the relationship between the weight of the finished rod (y), and that of the rough casting (x).
- Casting likely to produce rods that are too heavy can then be discarded before undergoing the final (and costly) tooling process.

© 2012 Mamikon Ginovyan

Example 1 (Manufacturer).

- As the first step in examining ***xy*** -relationship, ***n*=25** (x_i, y_i) -pairs are measured. The data are given in the following table.
- Use the **LSM** to find the **best straight line** approximating the ***xy*** -relationship and state your conclusion.
- The scatter plot suggests that ***y*** is linearly related to the ***x***.



© 2012 Mamikon Ginovyan

Example 1 (Manufacturer).

Data Table 1.

Rod Number	Rough Weight, x	Finished Weight, y	Rod Number	Rough Weight, x	Finished Weight, y
1	*2.745*	2.080	14	*2.635*	1.990
2	*2.700*	2.045	15	*2.630*	1.990
3	*2.690*	2.050	16	*2.625*	1.995
4	*2.680*	2.005	17	*2.625*	1.985
5	*2.675*	2.035	18	*2.620*	1.970
6	*2.670*	2.035	19	*2.615*	1.985
7	*2.665*	2.020	20	*2.615*	1.990
8	*2.660*	2.005	21	*2.615*	1.995
9	*2.655*	2.010	22	*2.610*	1.990
10	*2.655*	2.000	23	*2.590*	1.975
11	*2.650*	2.000	24	*2.590*	1.995
12	*2.650*	2.005	25	*2.565*	1.995
13	*2.645*	2.015			

© 2012 Mamikon Ginovyan

Example 1 (Manufacturer).

Solution.

From Data Table 1 we find $\sum_{i=1}^{25} x_i = 66.1, \ \sum_{i=1}^{25} x_i^2 = 174.7,$

$$\sum_{i=1}^{25} y_i = 50.1, \ \sum_{i=1}^{25} y_i^2 = 100.5, \ \sum_{i=1}^{25} x_i y_i = 132.5.$$

Therefore

$$\hat{b}_0 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{n=1}^{n} x_i\right)\left(\sum_{n=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{n=1}^{n} x_i\right)^2}$$

$$= \frac{25(132.5) - (66.1)(50.1)}{25(174.7) - (66.1)^2} = .64$$

© 2012 Mamikon Ginovyan

Example 1 (Manufacturer).

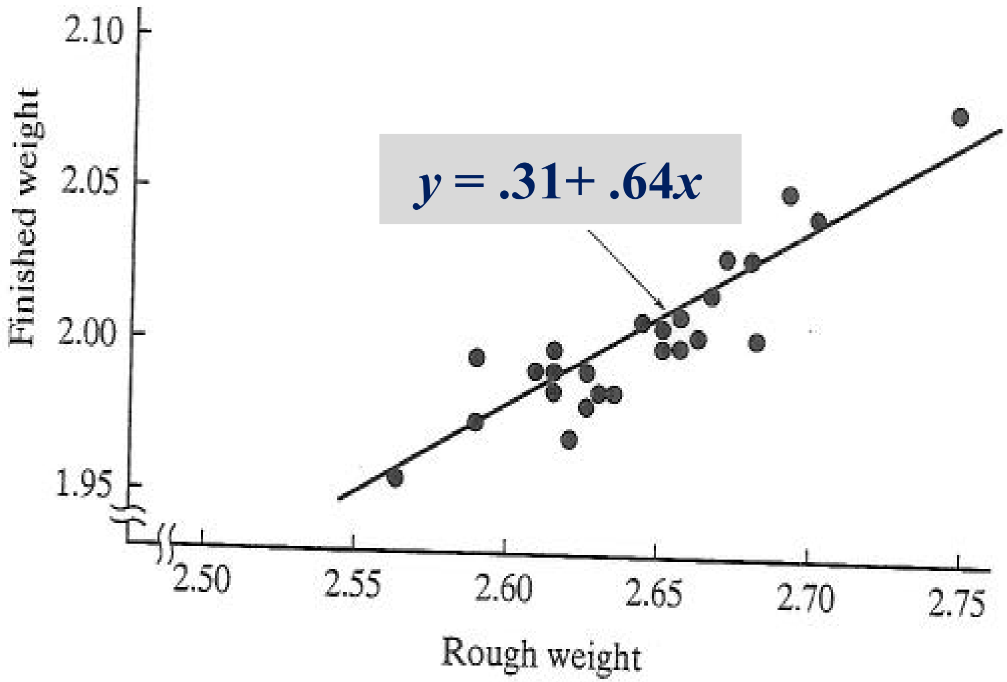
and

$$\hat{a}_0 = \bar{Y} - \frac{S_{XY}}{S_{XX}} \cdot \bar{X} = \bar{Y} - \hat{b}_0 \cdot \bar{X} = \frac{50.1 - .642(66.1)}{25} = .31.$$

<u>**Thus,**</u> the LSM - **best straight line approximating** ***xy***-relationship is :

$$\boldsymbol{y = .31 + .64x.}$$

The manufacturer is now in a position to make some informed policy decisions.



© 2012 Mamikon Ginovyan

Example 1 (Manufacturer).

If the weight of a rough casting is, say, $\boldsymbol{x}$ **= 2.71 oz.**, the **least-squares line predicts** that its finished weight $\hat{y}$ will be $\hat{y}$ = **2.05 oz.**:

Estimated weight $= \hat{y} = \hat{a}_0 + \hat{b}_0(2.71) = .308 + .642(2.71) = 2.05.$

•In the event that finished weights of **2.05 oz.** are considered to be too heavy, rough casting weighing **2.71 oz**. (or more) should be discarded.

© 2012 Mamikon Ginovyan

Residuals

Let $\hat{a}_0$ and $\hat{b}_0$ be the **least-squares coefficients** associated with the sample $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

We know that for any value of $\boldsymbol{x}$, the quantity

$$\hat{y} = \hat{a}_0 + \hat{b}_0 x$$

is the **predicted (linear) value** of $\boldsymbol{y}$.

© 2012 Mamikon Ginovyan

Residuals

Definition 1.

For each $i (i = \overline{1, n})$, the difference between an **observed** y_i and **predicted** $\hat{y}_i$ values, that is,

$$e_i = y_i - \hat{y}_i = y_i - (\hat{a}_0 + \hat{b}_0 x_i)$$

is called the ***i*-th residual.**

The **magnitude** of the ***i*-th residual** e_i reflects the **failure of the least-squares line to "model" that particular point.**

Definition 2. A **residual plot** is graph of the ***i*-th residual** e_i versus x_i, for all $i = \overline{1, n}$.

© 2012 Mamikon Ginovyan

Interpreting Residual Plots

Applied statisticians find **residual plot** to be very helpful in assessing the appropriateness of **fitting a straight line** through a set of n points.

- If the relationship between x and y is **linear**, the corresponding **residual plot** typically shows **no patterns, cycles, trends, or outliers**.

- For **nonlinear relationships**, though, **residual plots** often take on dramatically **nonrandom** appearances that can very effectively, highlight and illuminate the underlying association between x and y.

© 2012 Mamikon Ginovyan

Example 2.

- Construct the residual plot for the date in the Manufacturer-Example.
- What does its appearance imply about the suitability of fitting those points with a straight line?

© 2012 Mamikon Ginovyan

Example 2.

- **Solution.**

We begin by calculating the residuals for each of the twenty-five date points. The first observation recorded, for example, was $(x_1, y_1) = (2.745,\ 2.080).$

The corresponding predicted value is $\hat{y}_1 = 2.070$:

$$\hat{y}_1 = .308 - .642(2.745) = 2.070.$$

The first residual, then, is

$$e_1 = y_1 - \hat{y}_1 = 2.080 - 2.070 = .01.$$

The complete set of residuals appears in the last column of the following table.

© 2012 Mamikon Ginovyan

Example 2.

Table 2.

x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$
2.745	2.080	2.070	0.010
2.700	2.045	2.041	0.004
2.690	2.050	2.035	0.015
2.680	2.005	2.029	-0.024
2.675	2.035	2.025	0.010
2.670	2.035	2.022	0.013
2.665	2.020	2.019	0.001
2.660	2.005	2.016	-0.011
2.655	2.010	2.013	-0.003
2.655	2.000	2.013	-0.013
2.650	2.000	2.009	-0.009
2.650	2.005	2.009	-0.004
2.645	2.015	2.006	0.009
2.635	1.990	2.000	-0.010

© 2012 Mamikon Ginovyan

Example 2.

Table 2

x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$
2.630	1.990	1.996	-0.006
2.625	1.995	1.993	0.002
2.625	1.985	1.993	-0.008
2.620	1.970	1.990	-0.020
2.615	1.985	1.987	-0.002
2.615	1.990	1.987	0.003
2.615	1.995	1.987	0.008
2.610	1.990	1.984	0.006
2.590	1.975	11.971	0.004
2.590	1.995	1.971	0.024
2.565	1.955	1.955	0.000

© 2012 Mamikon Ginovyan

Example 2.

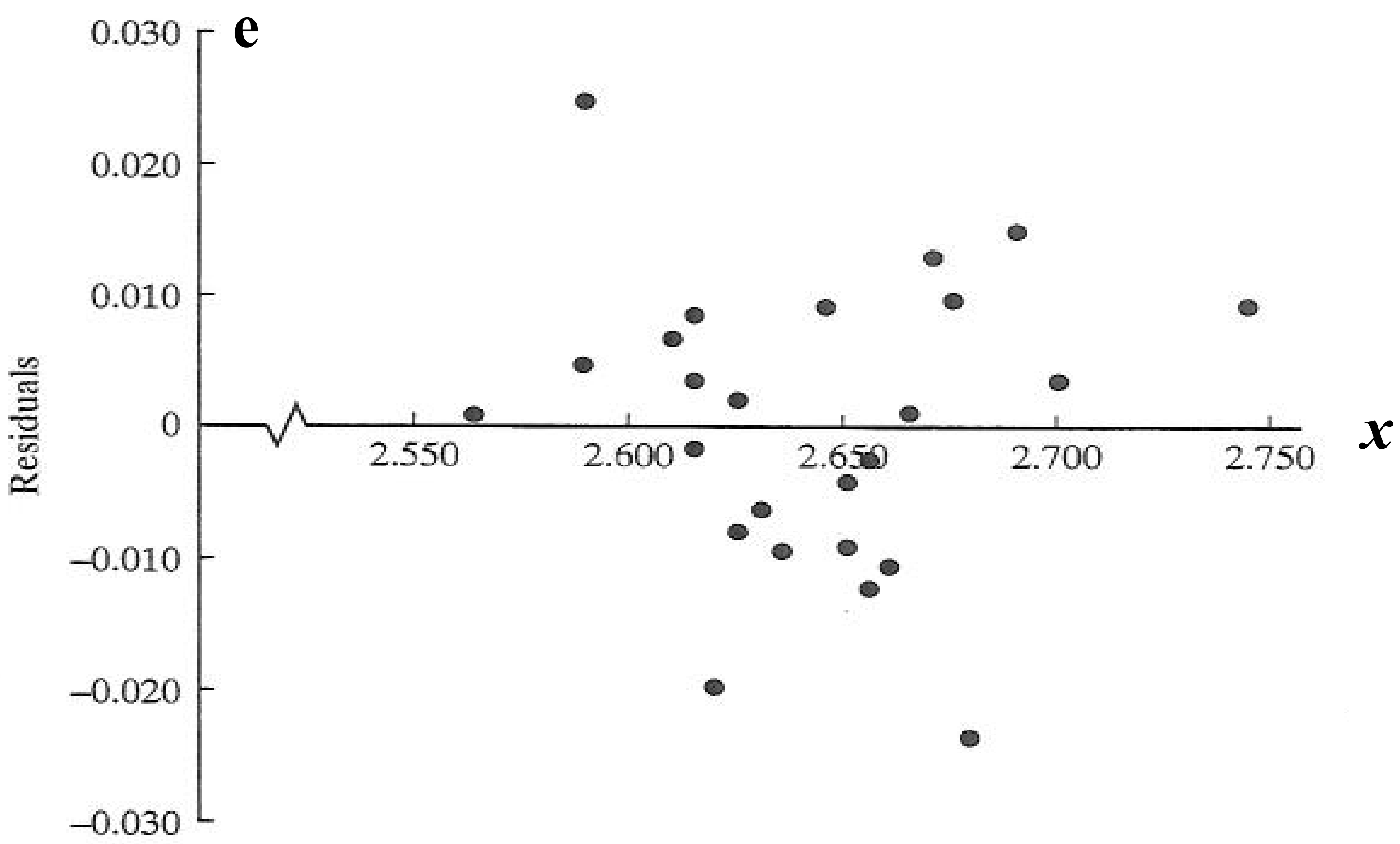


Fig.2

© 2012 Mamikon Ginovyan

Example 2.

Figure 2 shows the residual plot generated by fitting the least squares straight line $y = .308 + .642x,$
to the twenty-five points (x_i, y_i).

To an applied statistician, there is nothing here that would raise serious doubts about using a straight line to describe the
xy –relationship:
the points (in Fig.2) appear to be **randomly scattered** and exhibit **no obvious anomalies or patterns**.

© 2012 Mamikon Ginovyan

Example 3.

Table 3 below lists Social Security costs for selected years from **1965** through **1992**.
During that period, payouts rose from **$17.1** billion to **$285.1** billion.

•Is it reasonable to predict the Social Security costs in the year **2010** by using the **linear predictor**?

• Why or why not?

© 2012 Mamikon Ginovyan

Example 3.

Table 3.

Year	Year after 1960 x	Social Security Cost ($ Billion) y
1965	5	$17.1
1970	10	29.6
1975	15	63.6
1980	20	117.1
1985	25	186.4
1990	30	346.5
1992	32	285.1

© 2012 Mamikon Ginovyan

Example 3.

Solution. As in the Manufacturer-Example, we find

$$\hat{b}_0 = \frac{S_{XY}}{S_{XX}} = 10.3 \quad \text{and} \quad \hat{a}_0 = \bar{Y} - \hat{b}_0 \bar{X} = -66.2.$$

So, the **LSM best straight line** approximating ***xy*** -relationship, that is, the **best linear predictor** of ***y*** using ***x***, is

$$\hat{y} = \hat{a}_0 + \hat{b}_0 x = -66.2 + 10.3x.$$

Thus, if we will use linear prediction, we should to predict that Social Security costs in the year **2010** (that is, when ***x* = 50**) will be **$448.8** billion:

$$\hat{y}(50) = -66.2 + 10.3(50) = 448.8.$$

© 2012 Mamikon Ginovyan

Example 3.

- <u>**Is it reasonable?**</u>
- At the first glance, the least-squares line <u>**does appear to fit**</u> the data quite well (see Fig. 3).

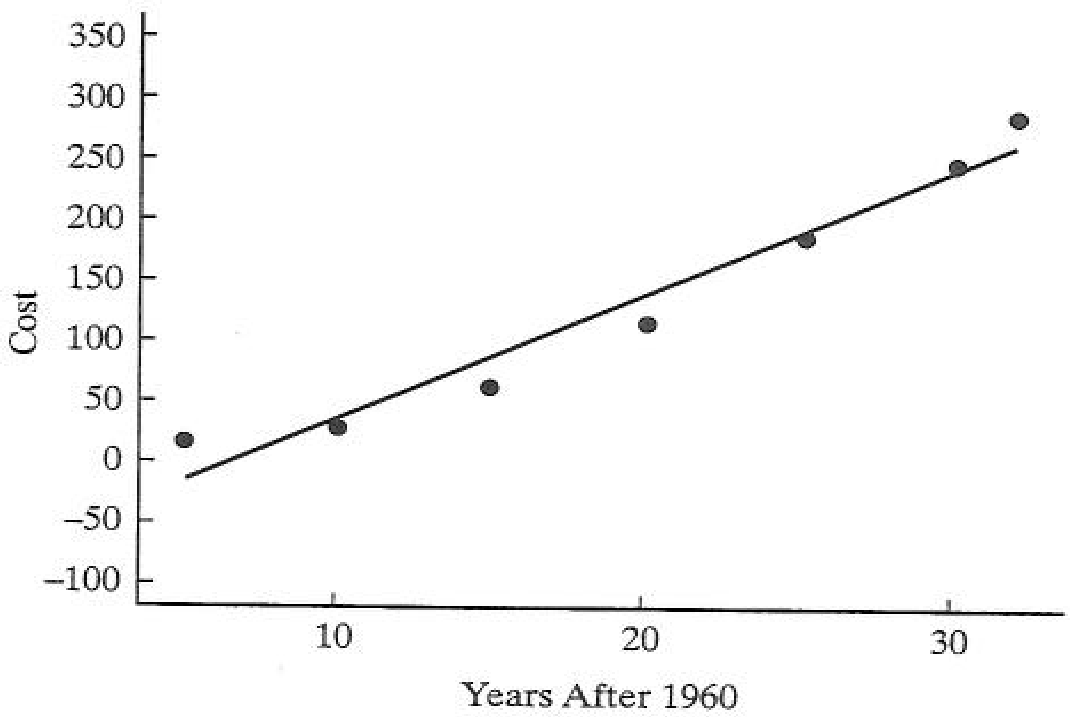


Fig. 3

© 2012 Mamikon Ginovyan

Example 3.

- A closer look, however, suggests that the underlying

xy -relationship may be **curvilinear** rather than **linear**.
The **residual plot** (see Fig.4) confirms that suspicion – there we see a distinctly **nonrandom pattern**.

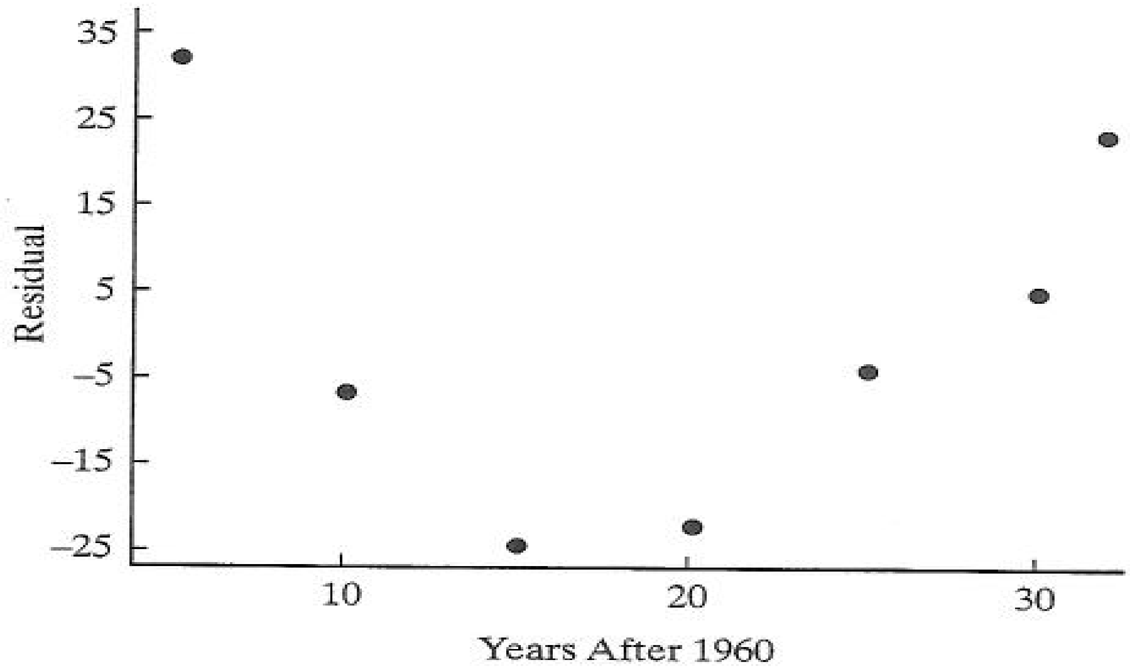


Fig. 4

© 2012 Mamikon Ginovyan

Example 3.

Conclusion.

- Using linear prediction procedure to predict the Social Security costs in the year **2010** to be **$448.8** billion **is not reasonable decision.**

- Based on the information in Table 3 the **$448.8** billion prediction is likely to **underestimate** substantially the cost of Social Security at the end of this decade.

© 2012 Mamikon Ginovyan

Some Nonlinear Models

- Obviously, not all ***xy*** **-**relationship can be adequately described by **linear models**, that is, by straight lines.

- **Curvilinear** relationships of all sorts can be found in every field of endeavor.
- Many of these nonlinear models, however, can still be **fit using SLS-Theorem**, provided that the data have been initially **"linearized"** by a suitable transformation.

© 2012 Mamikon Ginovyan

1. Exponential Regression

Suppose the relationship between x and y is best described by an **exponential function** of the form (See scatter plots in Fig.1).

$$y = a \cdot e^{bx}, a > 0, b \in R. \qquad (1)$$

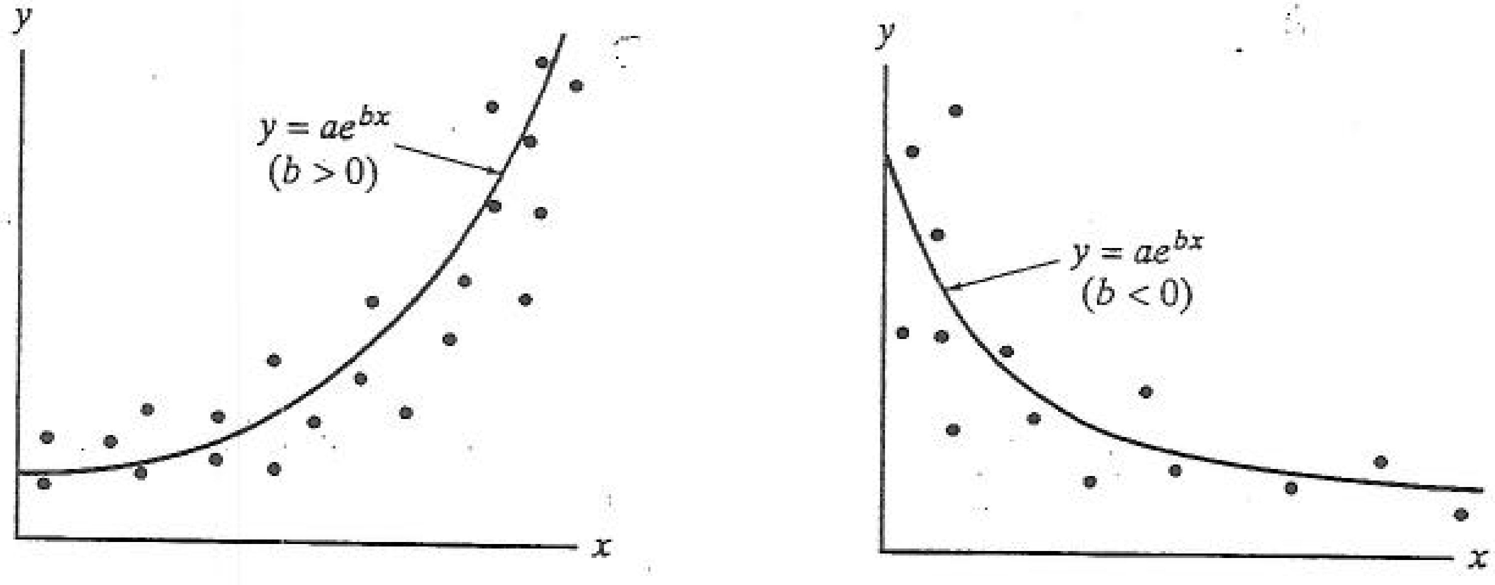


Fig. 1 (Scatter Plots)

© 2012 Mamikon Ginovyan

1. Exponential Regression

- To these **nonlinear** relationships between $\boldsymbol{x}$ and $\boldsymbol{y}$, however, we can associate a **linear model** if we observe that

$$y = a \cdot e^{bx} \Leftrightarrow \ln y = \ln a + bx.$$

Denoting by $y_1 = \ln y$ and $a_1 = \ln a$ we get the **linear model** (associated with (1))

$$y_1 = a_1 + bx \qquad (2)$$

Therefore, we can apply **SLS-Theorem** to the model (2) , and obtain the **best slope** and $\boldsymbol{y}$ -**intercept** of **nonlinear model** (1).

© 2012 Mamikon Ginovyan

1. Exponential Regression

Specifically,

$$\hat{b} = \frac{n\sum_{i=1}^{n} x_i \ln y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} \ln y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

and

$$\hat{a}_1 = \ln \hat{a} = \frac{1}{n}\left[\sum_{i=1}^{n} \ln y_i - \hat{b}\sum_{i=1}^{n} x_i\right].$$

•**Remark.**

The exponential regression models are particularly useful in Computer Science.

© 2012 Mamikon Ginovyan

2. Logarithmic Regression

Suppose that the relationship between **x** and **y** is best described by a **power function**

$$y = a \cdot x^b, \; x > 0, \;\; a > 0, \;\; b \in R. \qquad (3)$$

This model can be easily **linearized** if we take log of both side of equation (3).

Thus, $$y = a \cdot x^b \Leftrightarrow \log y = \log a + b \cdot \log x.$$

© 2012 Mamikon Ginovyan

2. Logarithmic Regression

Denoting by $y_1 = \log y$, $a_1 = \log a$, and $x_1 = \log x$ we get the **linear model**

$$y_1 = a_1 + bx_1 \qquad (4).$$

Therefore, we can apply **SLS-Theorem** to obtain

$$\hat{b} = \frac{n\sum_{i=1}^{n}(\log x_i)(\log y_i) - (\sum_{i=1}^{n}\log x_i)(\sum_{i=1}^{n}\log y_i)}{n\sum_{i=1}^{n}(\log x_i)^2 - (\sum_{i=1}^{n}\log x_i)^2}$$

and

$$\hat{a}_1 = \log \hat{a} = \frac{1}{n}\left[\sum_{i=1}^{n}\log y_i - \hat{b}\sum_{i=1}^{n}\log x_i\right].$$

© 2012 Mamikon Ginovyan

2. Logarithmic Regression

- **Remark.**

The model (3), called **logarithmic regression model**, has **slower growth rates** than **exponential models,** and are particular useful in describing **biological and engineering** phenomena.

© 2012 Mamikon Ginovyan

3. Logistics Regression (The Logit Model).

- **Growth** is a fundamental characteristic of living organisms, institution and ideas.
- Many growth models in **biology** (the change in size of a Drosophila population); in **economics** (proliferation of global market); in political science (the gradual acceptance of tax reform) can be described by the **logistic equation**

$$y = \frac{L}{1 + e^{a+bx}}, x \in R, \qquad (1)$$

where, a, b and L are constants.

For different values of a and b, equation (1) generates a variety of **S-shaped curves**.

© 2012 Mamikon Ginovyan

3. Logistics Regression (The Logit. Model).

To **linearize** the model (1) we make the following transformations:

$$y = \frac{L}{1 + e^{a+bx}}$$

$$\Leftrightarrow \frac{1}{y} = \frac{1 + e^{a+bx}}{L} \quad (the \quad reciprocal)$$

$$\Leftrightarrow \frac{L}{y} = 1 + e^{a+bx}$$

$$\Leftrightarrow \frac{L - y}{y} = e^{a+bx}$$

$$\Leftrightarrow \ln(\frac{L - y}{y}) = a + bx.$$

© 2012 Mamikon Ginovyan

3. Logistics Regression (The Logit. Model).

Denoting

$$y_1 = \ln\left(\frac{L-y}{y}\right),$$

we get the linear model

$$y_1 = a + bx. \qquad (2)$$

Remark 1.

The parameter $\boldsymbol{L}$ is interpreted as the limit to which $\boldsymbol{y}$ is converging as $\boldsymbol{x}$ increases $(x \to \infty)$.

In practice, $\boldsymbol{L}$ is often estimated simply by plotting the data and "**eye-balling**" the $\boldsymbol{y}$ -asymptote.

© 2012 Mamikon Ginovyan

3. Logistics Regression (The Logit. Model).

Now we can apply **SLS-Theorem** to the model (2) and obtain the **best slope** and y -**intercept** of model (1).

$$\hat{b} = \frac{n\sum_{i=1}^{n} x_i \ln\left(\frac{\hat{L}-y_i}{y_i}\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} \ln\left(\frac{\hat{L}-y_i}{y_i}\right)\right)}{n\sum_{i=1}^{n} x_i^{2} - \left(\sum_{i=1}^{n} x_i\right)^{2}} \qquad (3)$$

$$\hat{a} = \frac{1}{n}\left[\sum_{i=1}^{n} \ln\left(\frac{\hat{L}-y_i}{y_i}\right) - \hat{b}\sum_{i=1}^{n} x_i\right], \qquad (4)$$

where $\hat{L}$ is an **estimate** for L.

© 2012 Mamikon Ginovyan

3. Logistics Regression (The Logit. Model).

Remark 2.

The distribution function *(cdf)* $F(x) = \dfrac{1}{1+e^{-x}}$

corresponding to $\boldsymbol{L} = \mathbf{1}$, $\boldsymbol{a} = \mathbf{0}$ and $\boldsymbol{b} = \mathbf{-1}$, is called the **standard logistic distribution**.

The corresponding *pdf* is

$$f(x) = \frac{e^{-x}}{(1+e^{-x})^2}. \qquad (5)$$

Problem. Show that the function *f(x)* given by (5) is indeed a *pdf*, that is,

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} \frac{e^{-x}}{(1+e^{-x})^2} dx = 1.$$

© 2012 Mamikon Ginovyan

Example

To determine how effective the **SAT scores** are in predicting academic success, the National Collegiate Athletic Association (NCAA) compiled the data in Table 1, showing the relationship between athletes' **SAT scores** **(*x*)** and their **graduation rates** **(*y*).**

- Quantify the **graduation rate/SAT** score relationship by choosing an **appropriate model of fitting the data**.

© 2012 Mamikon Ginovyan

Example

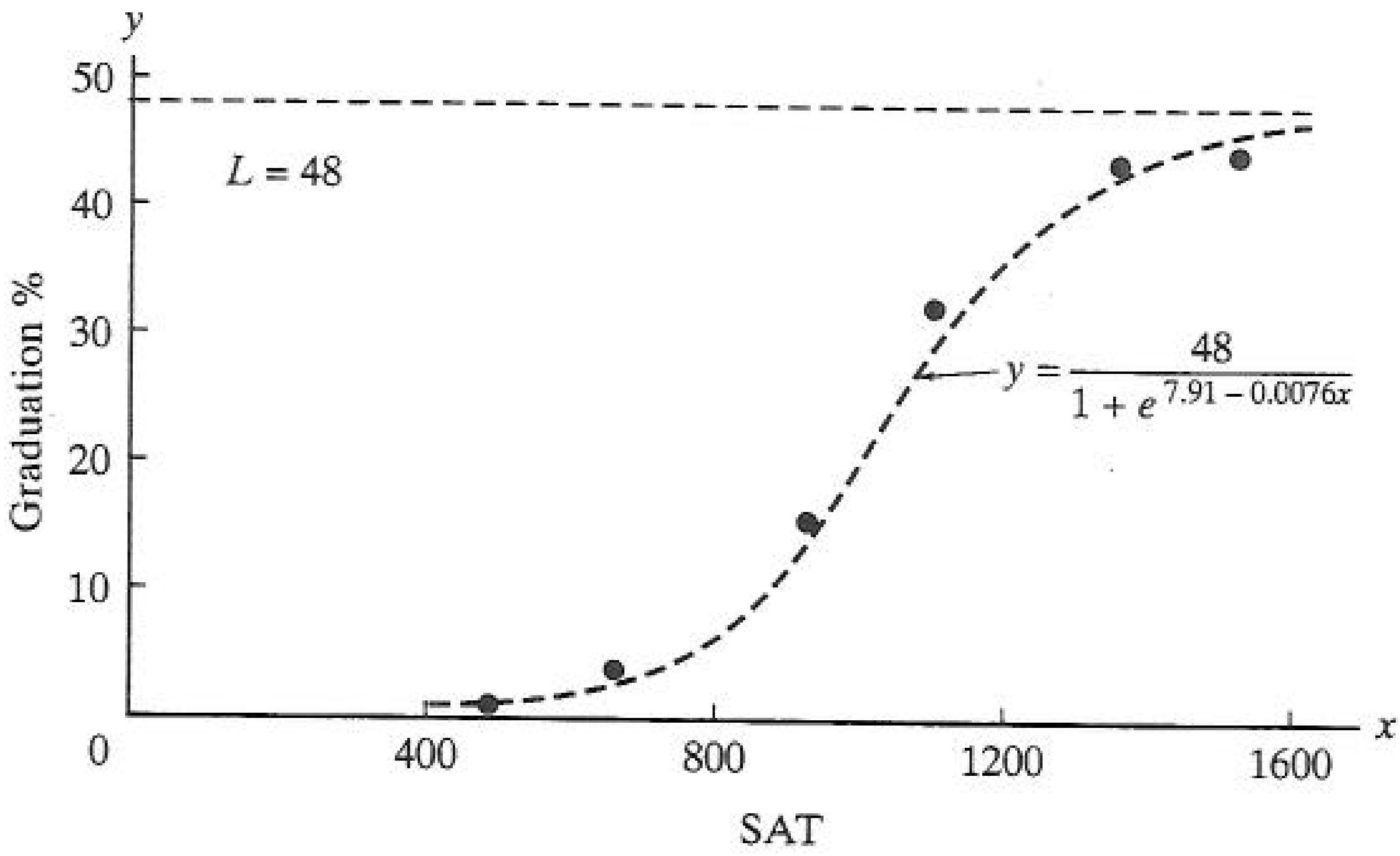
Table 1.

SAT Score, x	Graduation Rate (%), y
480	0.3
690	4.6
900	15.6
1100	33.4
1320	44.4
1530	45.7

© 2012 Mamikon Ginovyan

Example

- **Solution.** Start with **scatter plot**.



© 2012 Mamikon Ginovyan

Example

- The **scatter plot** for the six data points has a **definite S-shaped** appearance (see Fig. 1), which makes Equation (1) a good candidate for modeling the ***xy*** **-**relationship.

- The limit to which the graduation rates are converging (as SAT score increase) appears to be about 48.

 So we can fit the data points using **logistic model** with ***L*** **= 48**:

$$y = \frac{48}{1 + e^{a+bx}}.$$

- To find the best **LSE's** for a and b we use formulas (3), and the following table.

© 2012 Mamikon Ginovyan

Example

Table 2.

x_i	y_i	$\ln(\frac{48-y_i}{y_i})$	x_i^2	$x_i \cdot \ln(\frac{48-y_i}{y_i})$
480	0.3	5.06890	230,400	2433.072
690	4.6	2.24440	476,100	1548.636
900	15.6	0.73089	810,000	657.801
1100	33.4	-0.82753	1,210,000	-910.283
1320	44.4	-2.51231	1,742,400	-3316.249
1530	45.7	-2.98919	2,340,900	-4573.461
6020			**6,809800**	**-4160.461**

© 2012 Mamikon Ginovyan

Example

For $\hat{a}$ and $\hat{b}$ we have

$$\hat{b} = \frac{6(-4160.484) - (6020)(1.71516)}{6(6809800) - (6020)^2} = -.0076,$$

$$\hat{a} = \frac{1.71516 - (.0076)(6020)}{6} = 7.91.$$

Thus, the **best-fitting logistic curve** has the equation

$$y = \frac{48}{1 + e^{7.91 - .0076x}}.$$

© 2012 Mamikon Ginovyan