Probability and Statistics
Lecture 11: Regression and Correlation

to accompany

Probability and Statistics for Engineers and Scientists

Fatih Cavdur

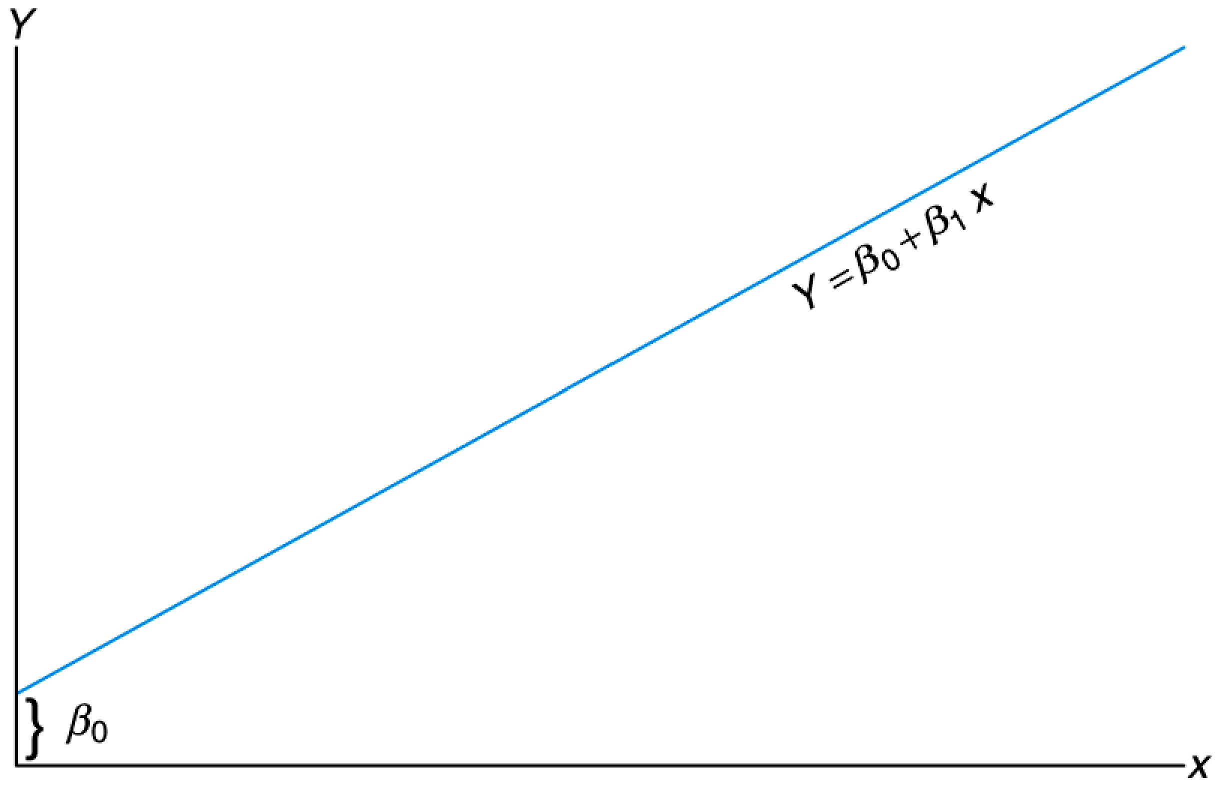
Introduction

A simple relationship between the dependent variable (response) Y and the independent variable (regressor) x can be written as

$$Y = \beta_0 + \beta_1 x$$

where β_0 is the intercept and β_1 is the slope.

Introduction



Introduction

In many applications, there might be more than one independent variable, such as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

which is a multiple linear regression equation with 2 independent variables.

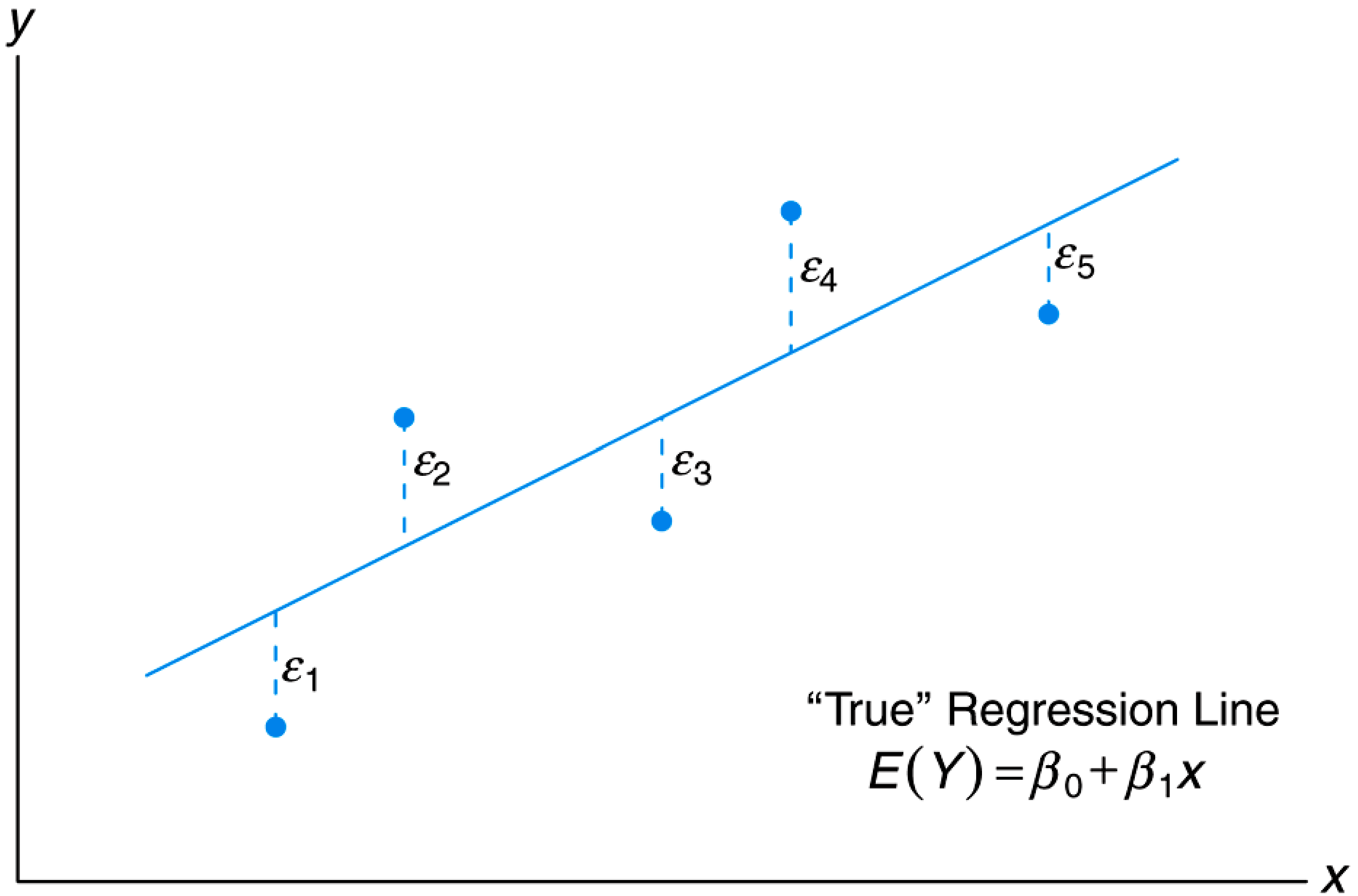
The Simple Linear Regression (SLR) Model

The SLR model can be written as

$$Y = \beta_0 + \beta_1 x + \epsilon$$

where $E(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$, and hence, Y is an RV.

The Simple Linear Regression (SLR) Model



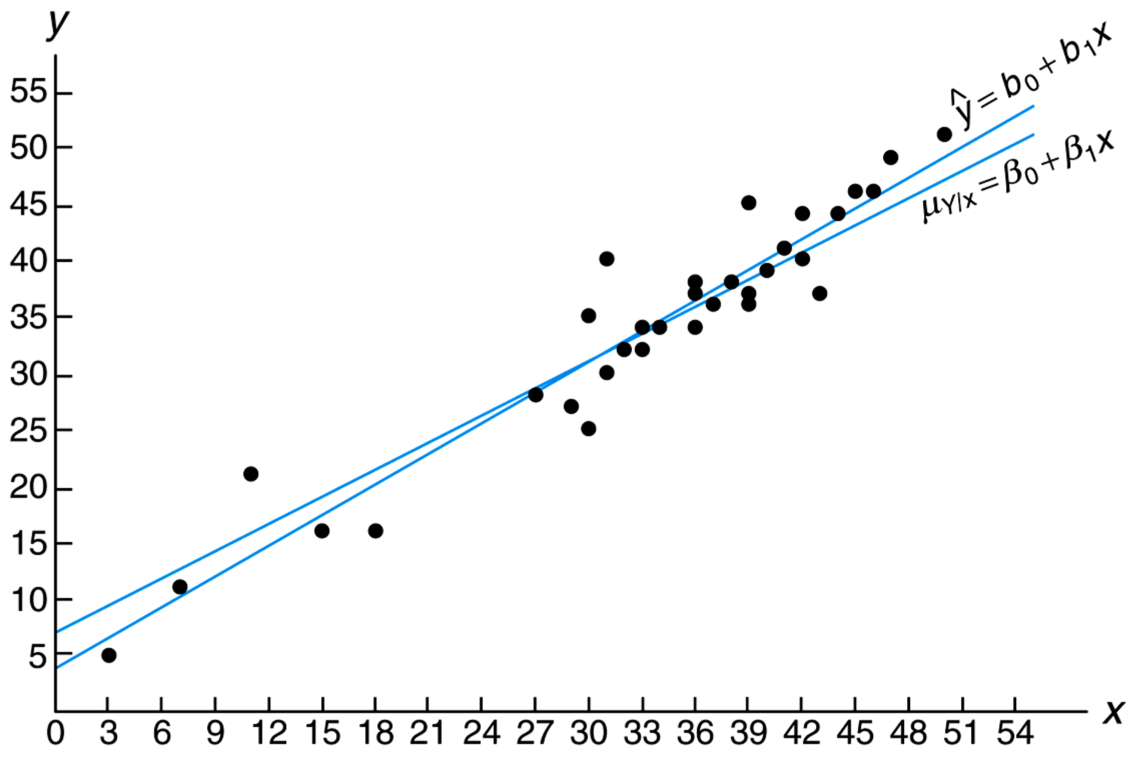
The Simple Linear Regression (SLR) Model

Example 11.1: One of the more challenging problems confronting the water pollution control field is presented by the tanning industry. Tannery wastes are characterized by high values of chemical oxygen demand, volatile solids and other pollution measures. Consider the following data where 33 samples of chemically treated waste in a study conducted at Virginia Tech in which x is the percent reduction in total solids and y is the percent reduction in chemical oxygen demand.

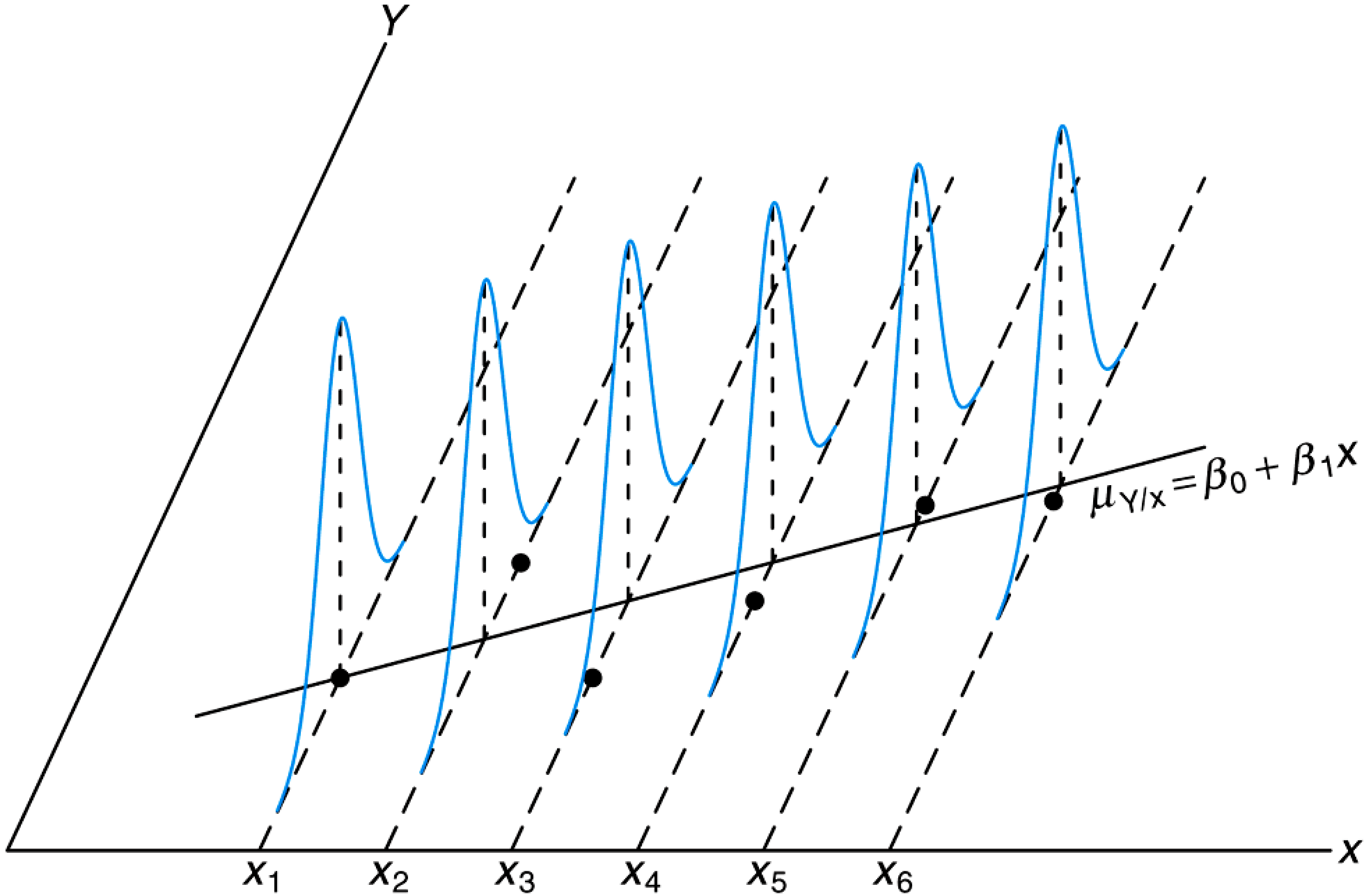
The Simple Linear Regression (SLR) Model

Solids Reduction, x (%)	Oxygen Demand Reduction, y (%)	Solids Reduction, x (%)	Oxygen Demand Reduction, y (%)
3	5	36	34
7	11	37	36
11	21	38	38
15	16	39	37
18	16	39	36
27	28	39	45
29	27	40	39
30	25	41	41
30	35	42	40
31	30	42	44
31	40	43	37
32	32	44	44
33	34	45	46
33	32	46	46
34	34	47	49
36	37	50	51
36	38		

The Simple Linear Regression (SLR) Model



The Simple Linear Regression (SLR) Model

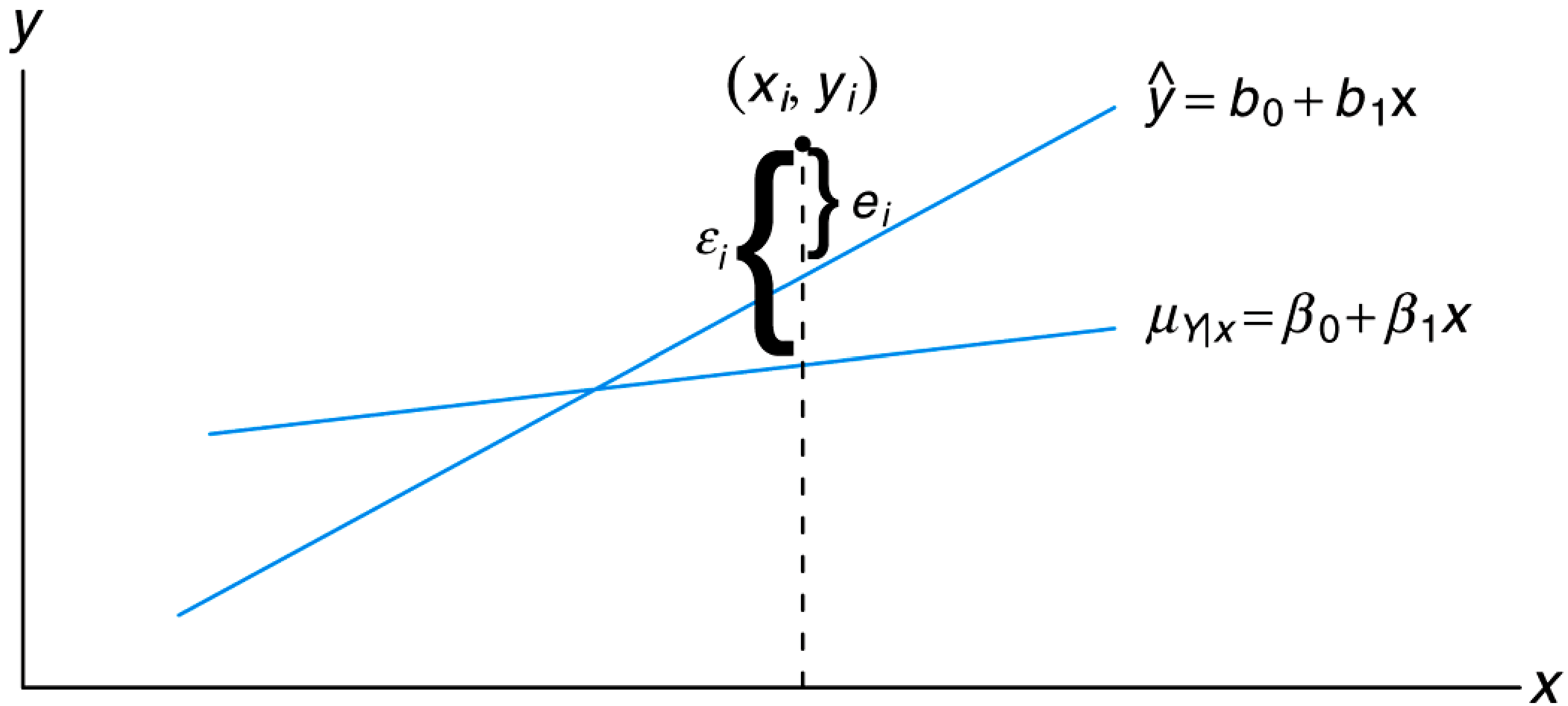


Least Squares and the Fitted Model

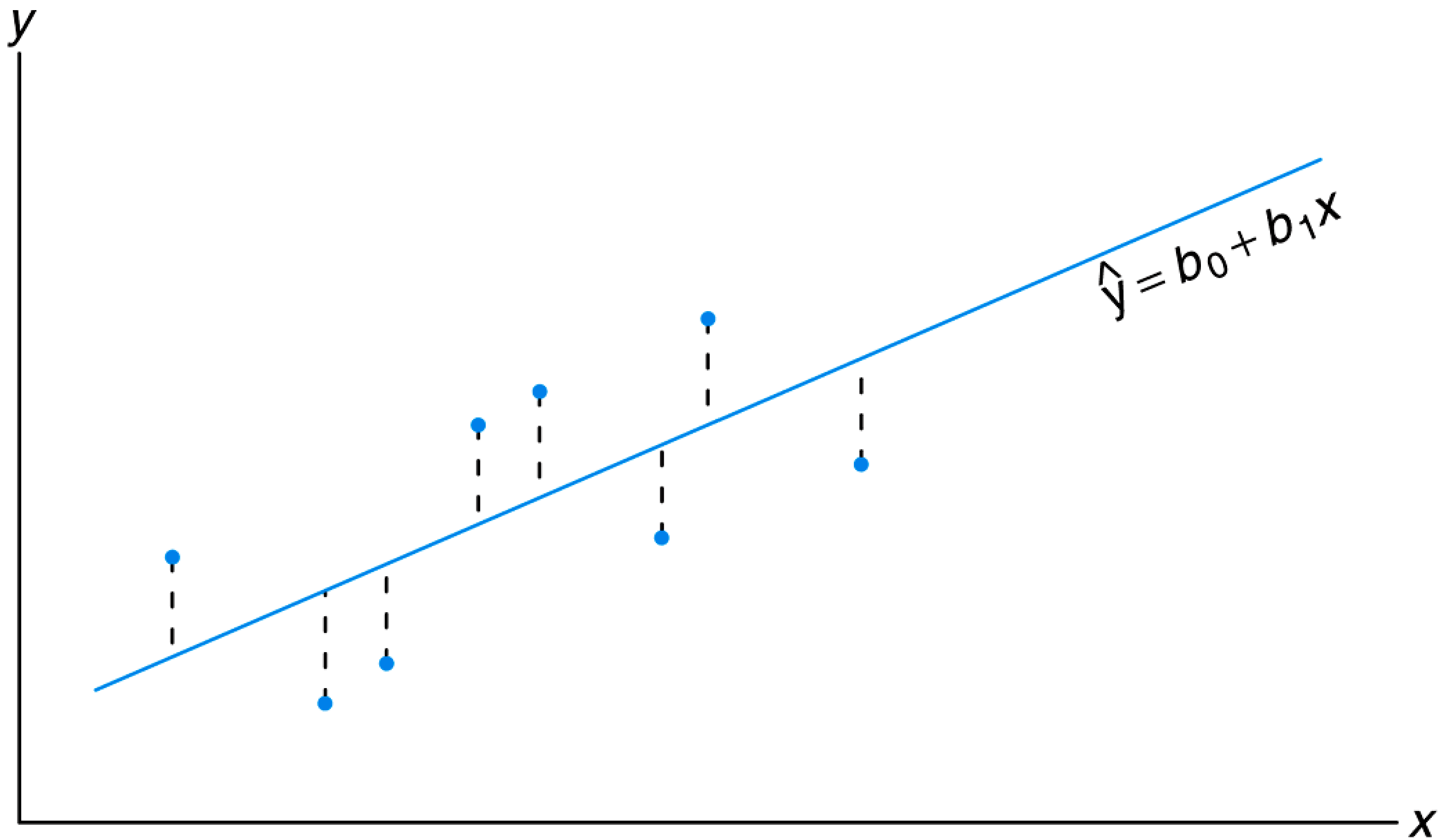
Given a set of regression data $\{(x_i, y_i), i = 1,2, \dots, n\}$ and a fitted model, $\hat{y}_i = b_0 + b_1 x$, the ith residual e_i is given by

$$e_i = y_i - \hat{y}_i, \quad i = 1,2, \dots, n$$

Least Squares and the Fitted Model



Least Squares and the Fitted Model



Least Squares and the Fitted Model

We want to minimize the Sum of Squared Errors (SSE) defined as

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

and

$$\frac{\partial(SSE)}{\partial b_0} = -2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)$$

$$\frac{\partial(SSE)}{\partial b_1} = -2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) x_i$$

Least Squares and the Fitted Model

We then have

$$\frac{\partial(SSE)}{\partial b_0} = 0 \Rightarrow nb_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\frac{\partial(SSE)}{\partial b_1} = 0 \Rightarrow b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Least Squares and the Fitted Model

By solving the above equations (normal equations), we obtain

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

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - b_1\bar{x}$$

Least Squares and the Fitted Model

In our example, we have

$$\sum_{i=1}^{33} x_i = 1{,}104; \quad \sum_{i=1}^{33} y_i = 1{,}124; \quad \sum_{i=1}^{33} x_i y_i = 41{,}355; \quad \sum_{i=1}^{33} x_i^2 = 41{,}086$$

$$\begin{aligned} b_1 &= \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\ &= \frac{(33)(41{,}355) - (1{,}104)(1{,}124)}{(33)(41{,}086) - (1{,}104)^2} \\ &= 0.904 \end{aligned}$$

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n} = \frac{1{,}124 - (0{,}904)(1{,}104)}{33} = 3.830$$

Least Squares and the Fitted Model

We can use the following notation in the following sections:

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \, ; \quad S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \, ; \quad S_{xy} = S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Least Squares and the Fitted Model

We can then write the sum of squared error as

$$SSE = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$$
$$= \sum_{i=1}^{33}[(y_i - \bar{y}) - b_1(x_i - \bar{x})]^2$$
$$= \sum_{i=1}^{33}(y_i - \bar{y})^2 - 2b_1\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) + b_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2$$
$$= S_{yy} - 2b_1 S_{xy} + b_1^2 S_{xx}$$
$$= S_{yy} - b_1 S_{xy}$$

Least Squares and the Fitted Model

An unbiased estimate of σ^2 is

$$s^2 = \frac{SSE}{n-2} = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n-2} = \frac{S_{yy} - b_1 S_{xy}}{n-2}$$

Least Squares and the Fitted Model

```
Regression Analysis: COD versus Per_Red
The regression equation is COD = 3.83 + 0.904 Per_Red

Predictor     Coef   SE Coef      T      P
Constant     3.830    1.768    2.17  0.038
Per_Red    0.90364  0.05012   18.03  0.000

S = 3.22954   R-Sq = 91.3%   R-Sq(adj) = 91.0%
Analysis of Variance
Source          DF      SS       MS       F      P
Regression       1  3390.6   3390.6  325.08  0.000
Residual Error  31   323.3     10.4
Total           32  3713.9
```

Least Squares and the Fitted Model

A $100(1-\alpha)\%$ CI for β_1 is given by

$$b_1 - t_{\alpha/2}\frac{s}{\sqrt{S_{xx}}} < \beta_1 < b_1 + t_{\alpha/2}\frac{s}{\sqrt{S_{xx}}}$$

A $100(1-\alpha)\%$ CI for β_0 is given by

$$b_0 - t_{\alpha/2}\frac{s}{\sqrt{nS_{xx}}}\sqrt{\sum_{i=1}^{33} x_i^2} < \beta_0 < b_0 + t_{\alpha/2}\frac{s}{\sqrt{nS_{xx}}}\sqrt{\sum_{i=1}^{33} x_i^2}$$

Least Squares and the Fitted Model

We can perform a test about the slope as follows:

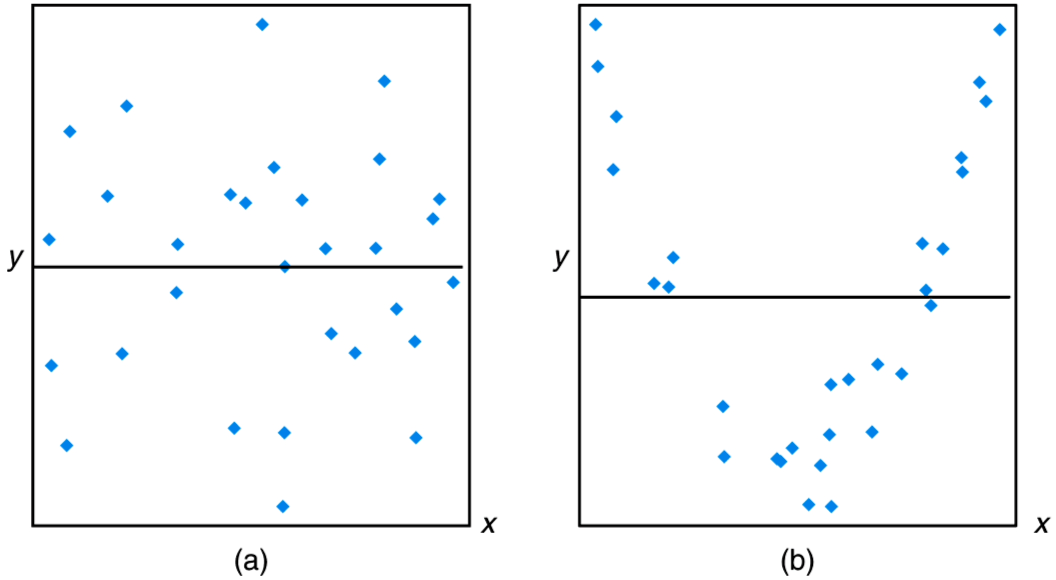
$$H_0\colon \beta_1 = \beta_{10}$$

$$H_1\colon \beta_1 \neq \beta_{10}$$

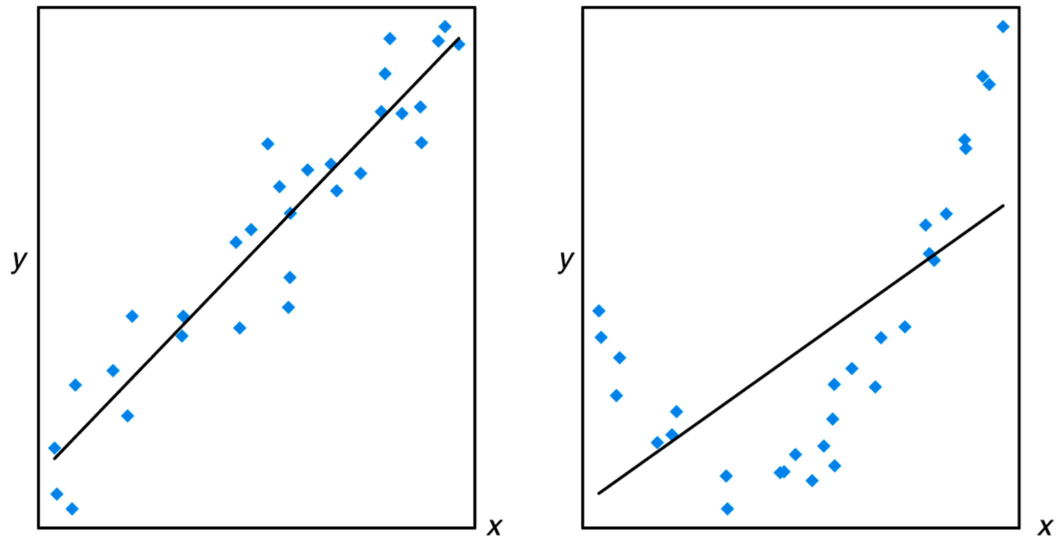
where we use the t-statistic as

$$t = \frac{b_1 - \beta_{10}}{s/\sqrt{S_{xx}}}$$

Least Squares and the Fitted Model



Least Squares and the Fitted Model



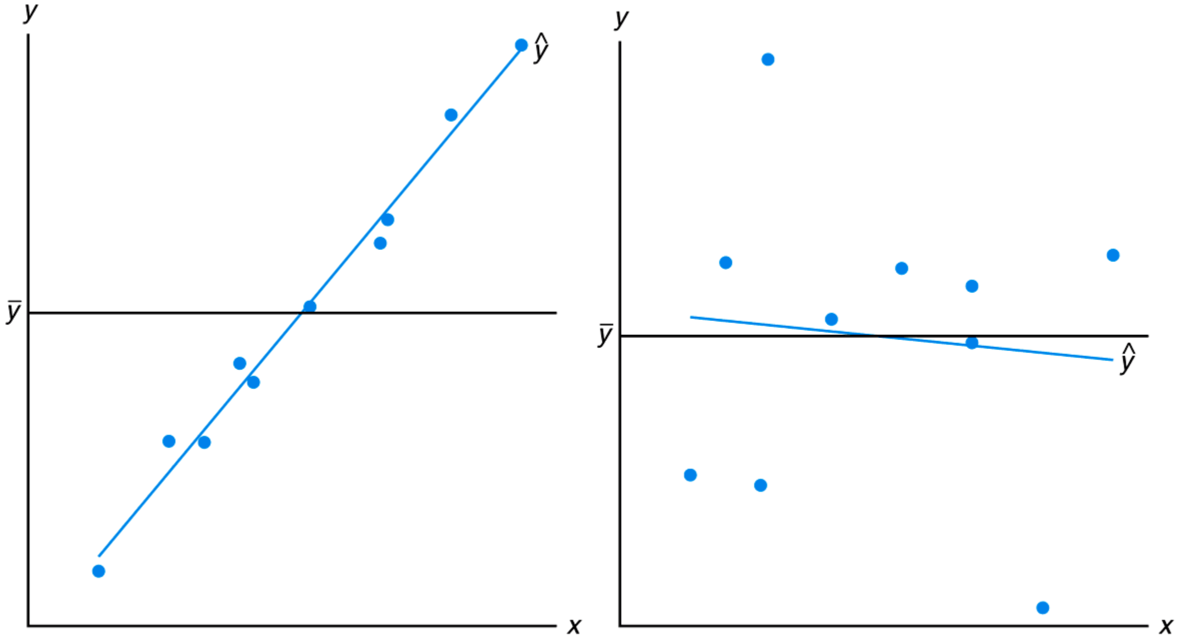
(a) (b)

Least Squares and the Fitted Model

The quality of fit is measured with a parameter called coefficient of determination, R^2 and computed as

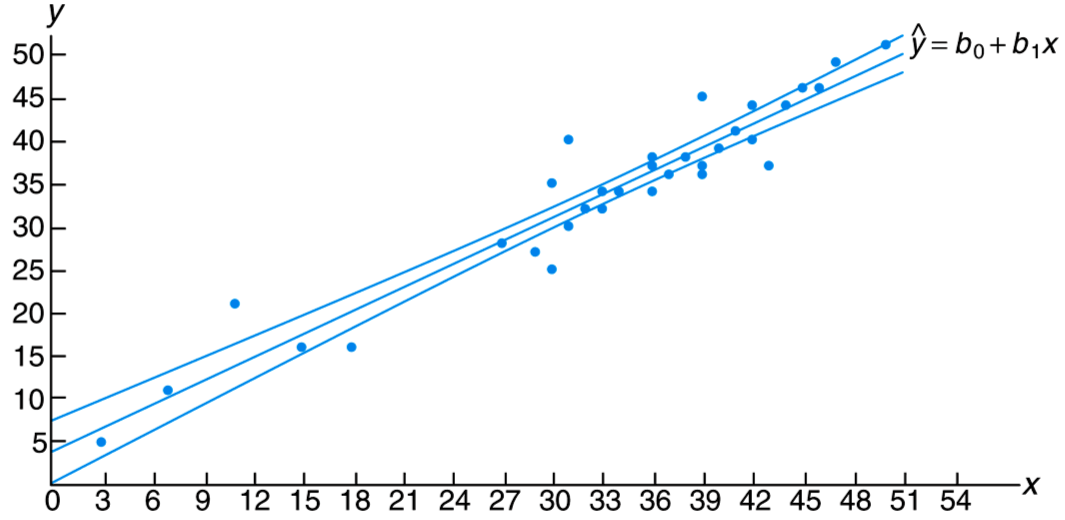
$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

Least Squares and the Fitted Model

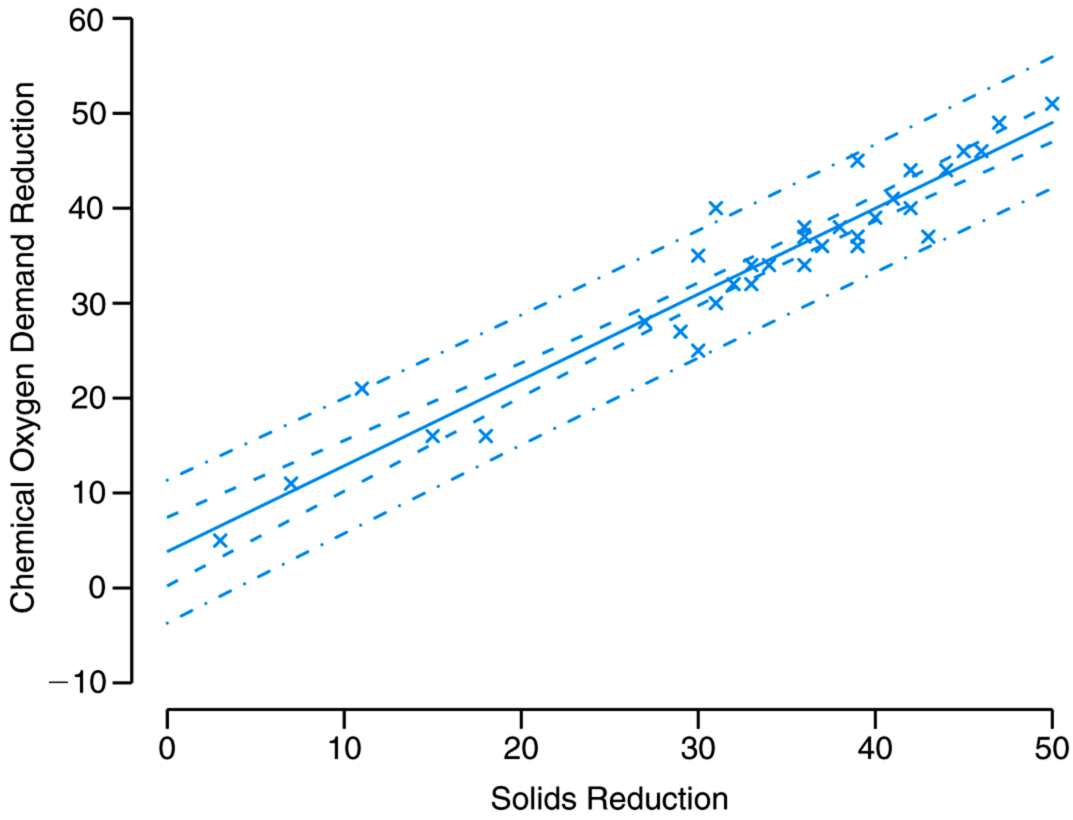


(a) $R^2 \approx 1.0$ (b) $R^2 \approx 0$

Least Squares and the Fitted Model



Least Squares and the Fitted Model



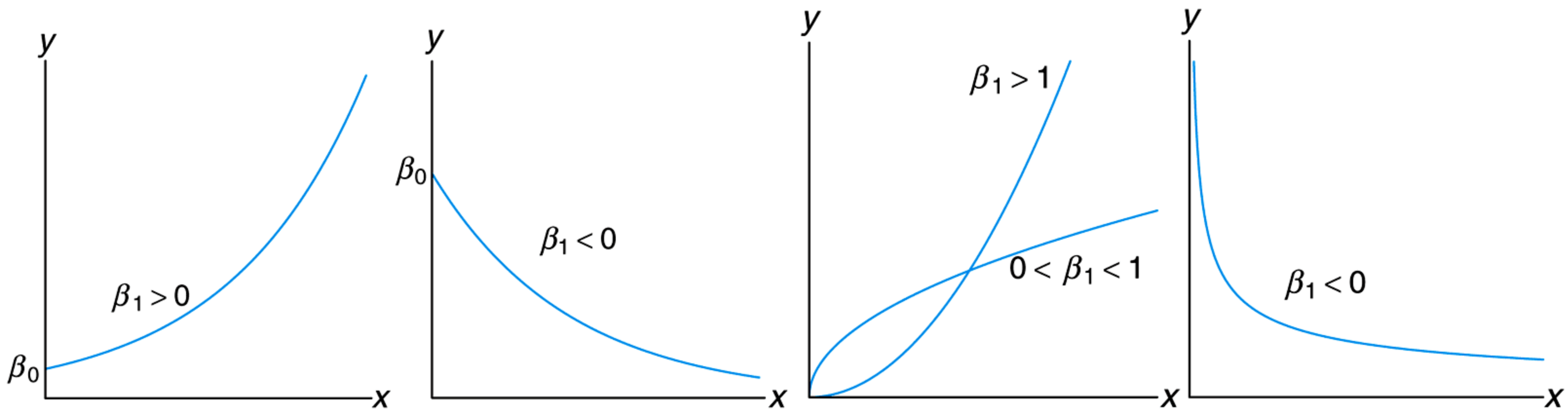
Least Squares and the Fitted Model

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Computed f
Regression	SSR	1	SSR	$\frac{SSR}{s^2}$
Error	SSE	$n-2$	$s^2 = \frac{SSE}{n-2}$	
Total	SST	$n-1$		

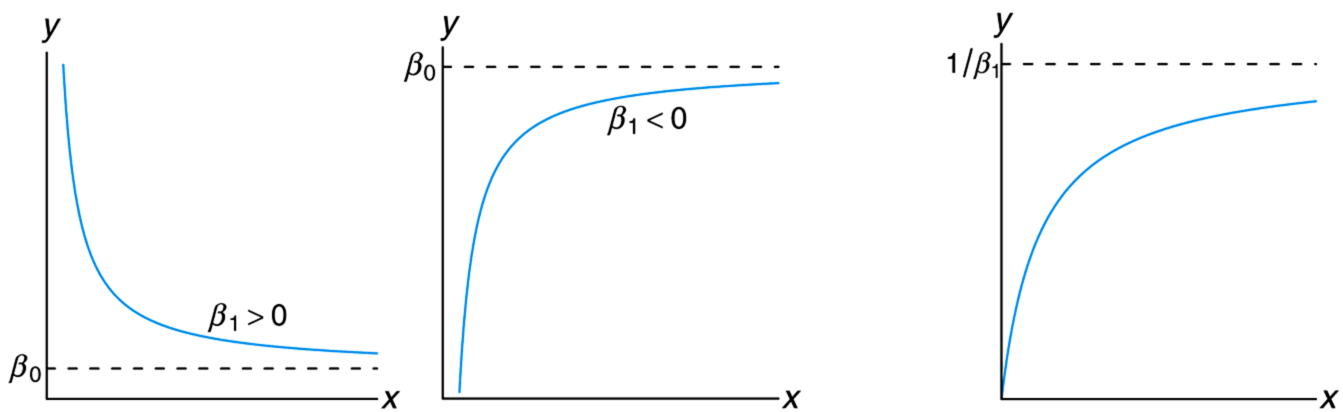
Some Useful Transformations

Functional Form Relating y to x	**Proper Transformation**	**Form of Simple Linear Regression**
Exponential: $y = \beta_0 e^{\beta_1 x}$	$y^* = \ln y$	Regress y^* against x
Power: $y = \beta_0 x^{\beta_1}$	$y^* = \log y; \quad x^* = \log x$	Regress y^* against x^*
Reciprocal: $y = \beta_0 + \beta_1 \left(\frac{1}{x}\right)$	$x^* = \frac{1}{x}$	Regress y against x^*
Hyperbolic: $y = \frac{x}{\beta_0 + \beta_1 x}$	$y^* = \frac{1}{y}; \quad x^* = \frac{1}{x}$	Regress y^* against x^*

Some Useful Transformations



(a) Exponential function (b) Power function



(c) Reciprocal function (d) Hyperbolic function

Fatih Cavdur – fatihcavdur@uludag.edu.tr

Correlation

The measure ρ of linear correlation between two variables X and Y is estimated by the sample correlation coefficient r as

$$r = b_1 \sqrt{\frac{S_{xx}}{S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

End of Lecture

Thank you! Questions?