Simple Linear Regression:

- Many engineering problems involve relationships between variables which are not deterministic.
- In stochastic situations the value of the response (dependent variable) cannot be predicted perfectly from the independent variables (input variables).

Regression Analysis: collection of statistical tools that are used to model and explore relationships between variables tjhat are not related in a deterministic manner.

Objective of Regression:

Build a model based on a set of observations which can be used for:

- Prediction
- Interpolation or extrapolation
- Optimization
- Control

The parameters in the model are called regression coefficients. e.g intercept and slope in a linear model.

Regressor(s) or predictor(s): is (are) the set of independent variable(s). Input variables for the model.
Response: is the dependent variable (output of the model)

In a distillation process,

y is the purity of oxygen produced in a chemical distillation process, and *x* is the percentage of hydrocarbons that are present in the main condenser of the distillation unit

Observation Number	Hydrocarbon Level x (%)	Purity y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

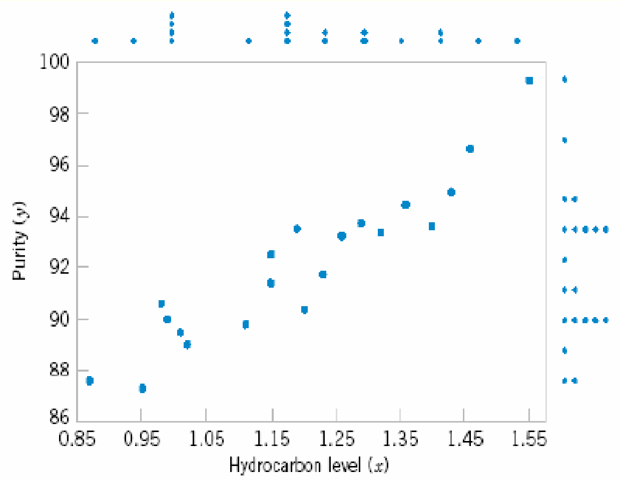

Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

Each observation, Y can be described by the model (as an estimation):

$$\hat{y} = \beta_o + \beta_1 x + \varepsilon \qquad \varepsilon \text{ is random error}$$

Where the intercept β_o and the slope β_1 are unknown regression coefficients

Assumptions for ε :

- Zero mean value
- Variance σ^2 is constant
- Normally distributed

The method of least squares is used to estimate the parameters β_o and $\beta_1$by minimizing the sum of the squares of the vertical deviations [(observed – calculated)] of the dependent variable.

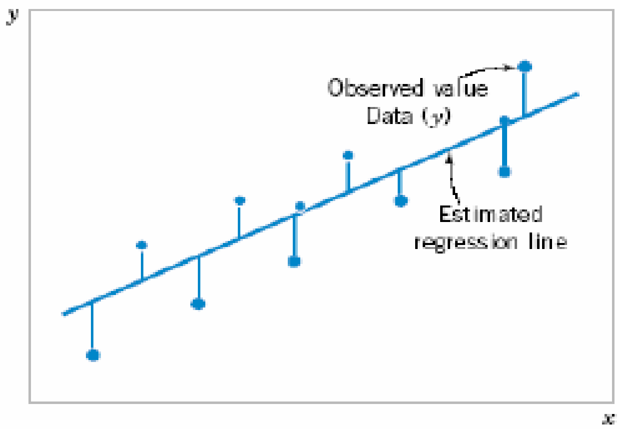

Figure 11-3 Deviations of the data from the estimated regression model.

$$y_i = \beta_o + \beta_1 x_i + \varepsilon_i, \qquad\qquad i = 1,2, \dots \dots, n$$

- The sum of the squares of the deviations of the observations from the true regression line is:

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_o - \beta_1 x_i)^2$$

For the function L to be minimum; its first derivatives with respect to all parameters must equal zero. This yields a number of first order differential equations that equals the number of parameters of concern. Solving this system of equations, we obtain estimates of the parameters.

$$\left.\frac{\partial L}{\partial \beta_o}\right|_{\beta_o \beta_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_o - \hat{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{\beta_o \beta_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_o - \hat{\beta}_1 x_i)\, x_i = 0$$

After rearrangement:

$$\hat{\beta}_o n + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_o \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i \; x_i$$

From these equations we obtain the least square estimates of the intercept and the slope in the simple linear regression model as:

$$\hat{\beta}_o = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i \; x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

$$where\ \bar{y} = \left({}^{1}\!/_{n}\right) \sum_{i=1}^{n} y_i \ \ and\ \bar{x} = \left({}^{1}\!/_{n}\right) \sum_{i=1}^{n} x_i$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

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

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

$\mathrm{SSE} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ Error sum of squares

$\mathrm{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ Total sum of squares

We will fit a simple linear regression model to the oxygen purity data in Table 11-1. The following quantities may be computed:

$$n = 20 \quad \sum_{i=1}^{20} x_i = 23.92 \quad \sum_{i=1}^{20} y_i = 1{,}843.21 \quad \bar{x} = 1.1960 \quad \bar{y} = 92.1605$$

$$\sum_{i=1}^{20} y_i^2 = 170{,}044.5321 \quad \sum_{i=1}^{20} x_i^2 = 29.2892 \quad \sum_{i=1}^{20} x_i y_i = 2{,}214.6566$$

$$S_{xx} = \sum_{i=1}^{20} x_i^2 - \frac{\left(\sum_{i=1}^{20} x_i\right)^2}{20} = 29.2892 - \frac{(23.92)^2}{20} = 0.68088$$

and

$$S_{xy} = \sum_{i=1}^{20} x_i y_i - \frac{\left(\sum_{i=1}^{20} x_i\right)\left(\sum_{i=1}^{20} y_i\right)}{20} = 2{,}214.6566 - \frac{(23.92)(1{,}843.21)}{20} = 10.17744$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{10.17744}{0.68088} = 14.94748$$

$$\hat{\beta}_o = \bar{y} - \hat{\beta}_1 \bar{x} = 92.1605 - (14.94748)1.196 = 74.28331$$

The fitted simple linear regression model is:

$$\hat{y} = 74.283 + 14.947\,x$$

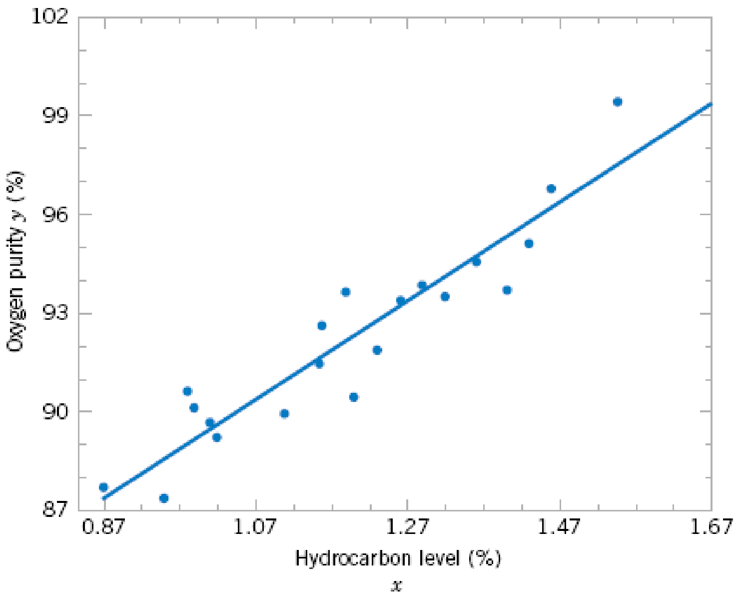

Figure 11-4 Scatter plot of oxygen purity y versus hydrocarbon level x and regression model $\hat{y} = 74.20 + 14.97x$.

The coefficient of determination R^2 is often used to judge the adequacy of a regression model:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Where SSR = Sum square of residuals

The range of R^2 is : $0 \le R^2 \le 1$

The correlation coefficient (R) is the positive square root of R^2

The coefficient of determination can sometimes be considered as the amount of variability in the data accounted for by the regression model.

For the oxygen purity regression model, we have

$$R^2 = SS_R/SS_T = 152.13/173.38 = 0.877,$$

that is, the model accounts for 87.7% of the variability in the data.

- Many models are intrinsically linear i.e., can be transformed to linear form by proper manipulations
  - Power law
  - Exponential
  - Saturation

Power law	Exponential	Saturation
$y = ax^b$ $\xrightarrow{\ln} \ln y = \ln a + b \ln x$ $\longrightarrow y' = \beta_0 + \beta_1 x'$	$y = ae^{bx}$ $\xrightarrow{\ln} \ln y = \ln a + bx$ $\longrightarrow y' = \beta_0 + \beta_1 x$	$y = \frac{ax}{1+bx}$ $\xrightarrow{\text{Reciprocal}} \frac{1}{y} = \frac{b}{a} + \frac{1}{a}\frac{1}{x}$ $\longrightarrow y' = \beta_0 + \beta_1 x'$
Transform using logarithms then the new variables will be ln y and ln x. Also, the parameters will be ln a and original b.	Transform using logarithms then the new variables will be ln y and original x. Also, the parameters will be ln a and original b.	Transform using reciprocals then the new variables will be $1/y$ and $1/x$. Also, the parameters will be b/a and reciprocal a.