University of Jordan
Chemical Engineering Department
Process Modeling by Statistical Methods– 905331

Lecture 09: Linear Regression

Dr. Ali Khalaf Al-Matar

Introduction

- Objective of regression is to build a model of a set of data
  - Can be used for prediction.
  - Interpolation and/or extrapolation.
  - Optimization.
- The parameters in the model are called **regression coefficients.**
  - Intercept and slope in a linear model.
  - Parameters (A,B and C) in Antoine equation.
- **Regressor** or (predictor) is (are) the set of **independent** variable(s).
- **Response** is the **dependent** variable.

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

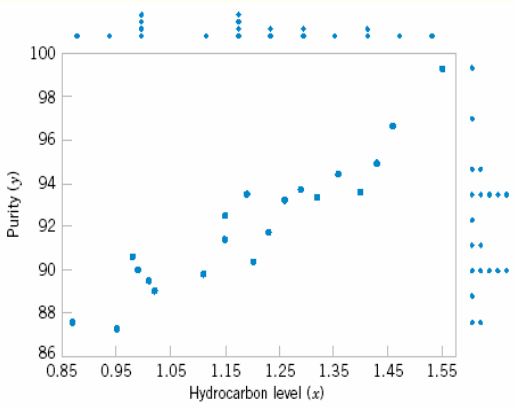


Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

The Simple Linear Regression Model

- each observation, Y, can be described by the model

$$\hat{y} = \beta_0 + \beta_1 x + \varepsilon$$

random error term

- Assumptions for ε
  - Zero mean value
  - Variance is σ^2.

Method of Least Squares

- The **method of least squares** is used to estimate the parameters, β_0 and β_1 by minimizing the sum of the squares of the vertical deviations
- The sum of the squares of the errors SSE (**residuals**) of the observations from the true regression line is

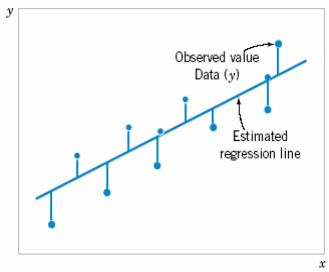


Figure 11-3 Deviations of the data from the estimated regression model.

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

Minimization of Squares

- For the function L to be minimum; its derivatives with respect to all parameters must be zero.
  - Generates a system of exactly the size of the number of parameters we have at hand (**normal equations**).
  - Solution to this system of equation provides estimates for the parameters

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\beta_0,\beta_1} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{\beta_0,\beta_1} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)x_i = 0$$

$$\rightarrow$$

$$\hat{\beta}_0 n + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

Simple Linear Regression Estimates

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

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

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}, \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

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

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

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

Error sum of squares

$$SST = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2$$

Total sum of squares

Estimating σ^2

- **Unbiased estimator**

$$\hat{\sigma}^2 = \frac{SSE}{n-2}$$

Confidence Limits

- Under the assumption that the observations are normally and independently distributed, a $100(1-\alpha)\%$ **confidence interval on the slope** β_1 in simple linear regression

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

- Similarly, a $100(1-\alpha)\%$ confidence interval on the intercept β_0

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]} \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]}$$

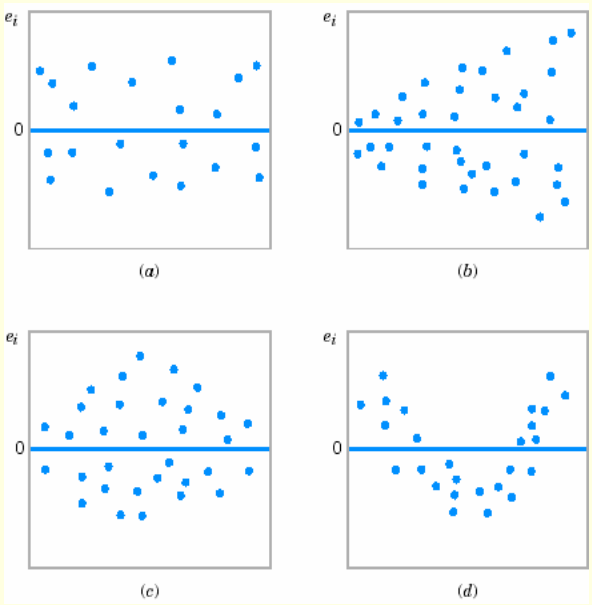
Adequacy of a Regression Model

- Fitting a regression model requires several **assumptions.**
  - Errors are uncorrelated random variables with mean zero;
  - Errors have constant variance; and,
  - Errors be normally distributed.
- The analyst should always consider the validity of these assumptions to be doubtful and conduct analyses to examine the adequacy of the model

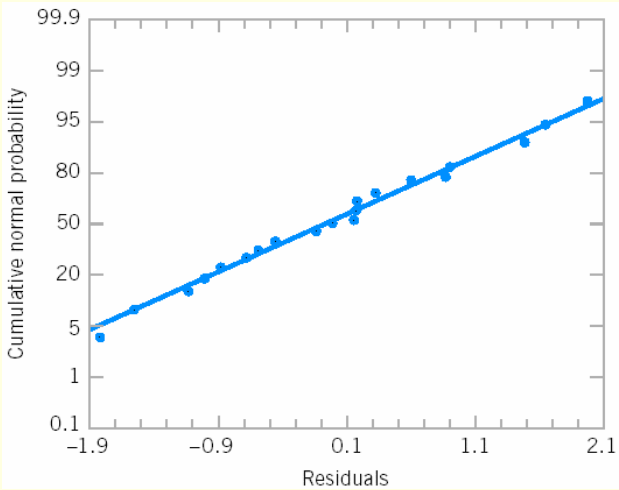
Residual Analysis

- The **residuals** from a regression model are $e_i = y_i - \hat{y}_i$, where y_i is an actual observation and $\hat{y}_i$ is the corresponding fitted value from the regression model.
- Analysis of the residuals is frequently helpful in checking the assumption that the errors are approximately normally distributed with constant variance, and in determining whether additional terms in the model would be useful.

Figure 11-9 Patterns for residual plots. (a) satisfactory, (b) funnel, (c) double bow, (d) nonlinear.
[Adapted from Montgomery, Peck, and Vining (2001).]



Normal Probability Plots



Correlation Coefficient and Coefficient of Determination

- The **coefficient of determination** is often used to judge the adequacy of a regression model

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

- The range is $0 \le R^2 \le 1$.
- Correlation coefficient (R) is the square root of R^2
  - The range is $-1 \le R \le 1$.
- We often refer (loosely) to R^2 as the amount of variability in the data explained or accounted for by the regression model.

Transformation to a Linear Model

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

Power law

$y = ax^b$

$\xrightarrow{\ln} \ln y = \ln a + b \ln x$

$\longrightarrow y' = \beta_0 + \beta_1 x'$

Transform using logarithms then the new variables will be ln *y* and ln *x*. Also, the parameters will be ln *a* and original *b*.

Exponential

$y = ae^{bx}$

$\xrightarrow{\ln} \ln y = \ln a + bx$

$\longrightarrow y' = \beta_0 + \beta_1 x$

Transform using logarithms then the new variables will be ln *y* and original *x*. Also, the parameters will be ln *a* and original *b*.

Saturation

$y = \dfrac{ax}{1 + bx}$

$\xrightarrow{\text{Reciprocal}} \dfrac{1}{y} = \dfrac{b}{a} + \dfrac{1}{a}\dfrac{1}{x}$

$\longrightarrow y' = \beta_0 + \beta_1 x'$

Transform using reciprocals then the new variables will be 1/*y* and 1/*x*. Also, the parameters will be b/*a* and reciprocal *a*.