Lecture 9

Simple Linear Regression

Statistics for
Civil & Environmental Engineers

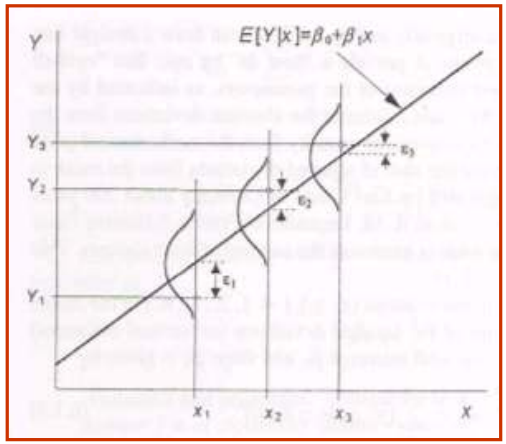
Relax a little and think about issue today!

- **Examples of Linear Models**
  - **Performance in college from test scores, weight from height, economic activity, population growth vs time, hydrologic statistics of a catchment from physiographic characteristics, deflections from load, accident rate from traffic volume and speed limit, runoff from snowpack**
- **Issues of Concern**
  - **How to predict Y best?**
  - **How accurately can parameters be estimated?**

Concept

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

where
Y: dependent, output,
or predictand variable
x: independent, input, predictor,
or explanatory variable
β_0 : intercept, β_1: slope,
ε: random error term



Simple Model I

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \varepsilon_i \sim \text{NID}[0,\ \sigma^2]$$

Engineer picks x_i and then observes Y_i

x_i: " fixed " independent variable

$$E[Y_i | x_i] =$$

$$Var[Y_i | x_i] =$$

Simple Model II

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \varepsilon_i \sim \text{NID}[0, \sigma^2]$$

X_i: " random " independent variable **$X_i \sim \text{NID}[\mu_x, \sigma_x^2]$**

$$E[Y_i | X_i = x_i] =$$

$$Var[Y_i | X_i = x_i] =$$

Model Parameter Estimators

❖ **Least Squares**

$$\min \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

$$-2\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad \text{and} \quad -2\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

$$\hat{\beta}_0 = \qquad \hat{\beta}_1 =$$

The Sum of Squares and Cross-Products

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

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2 / n$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i) / n$$

$$\Rightarrow \quad \hat{\beta}_1 =$$

Simple Linear Regression Example(1)

Example 6.1. Simple linear regression model for concrete strengths. From Table E.1.2 with observations x and y for concrete density and strength, respectively, the following summaries are obtained:

$n = 40$; $\bar{x} = 2445$ kg/m^3; $\bar{y} = 60.14$ N/mm^2;

$S_{xx} = 9977$; $S_{yy} = 980.8$; $S_{xy} = 1365$;

Properties of β_0

$$E[\hat{\beta}_0] = \beta_0$$

$$Var[\hat{\beta}_0] \sim \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

$$E[\hat{\beta}_1] = \beta_1$$

$$Var[\hat{\beta}_1] =$$

Properties of β_0

$$E[\hat{\beta}_1] = \frac{1}{S_{xx}} E[\sum_{i=1}^{n} Y_i (x_i - \bar{x})] = \frac{1}{S_{xx}} E[\sum_{i=1}^{n} (\beta_0 + \beta_1 x_i + \varepsilon_i)(x_i - \bar{x})]$$

$$= \frac{1}{S_{xx}} \{E[\beta_0 \sum_{i=1}^{n} (x_i - \bar{x})] + E[\beta_1 \sum_{i=1}^{n} x_i (x_i - \bar{x}) + E[\sum_{i=1}^{n} \varepsilon_i (x_i - \bar{x})]\}$$

$$Var[\hat{\beta}_1] = Var[\sum_{i=1}^{n} Y_i (x_i - \bar{x}) / S_{xx}] = \frac{1}{S_{xx}^2} \sigma^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{\sigma^2}{S_{xx}}$$

Error Variance and Residuals

- **Error Variance**

$$\hat{\sigma}^2 = \text{SSE} = \frac{1}{n-2} \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- **Assumptions of Residuals**
  - **Constant variance and normality in distribution of the residuals**
  - **Independence among the residuals**
  - **Independence between the residuals and the values of the explanatory variable *x***

Simple Linear Regression Example (2)

Example 6.2. Properties of the residuals of a linear regression model applied to concrete data. The 40 residuals $\hat{\epsilon}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ are determined from the data of Table 1.2.1 and the parameters estimated in Example 6.1. Here we examine their independence and distributional properties.

To determine whether the residuals are related to each other (in a case like this where there may be time-dependency between the observations), we firstly make an index plot of the residuals (that is, residual against observation number). Secondly, we plot the relationship $\hat{\epsilon}_i$ vs. $\hat{\epsilon}_{i-1}$. This type of plot is relevant when the observations are made at regular intervals of time. These two plots are shown in Fig. 6.1.3 and 6.1.4, respectively. Although the data are by no means perfect (there is evidence of some runs in Fig. 6.1.3) the plots do not show any significant autocorrelation in the residuals.[6] Such a relationship would result in a trend in Fig. 6.1.4.

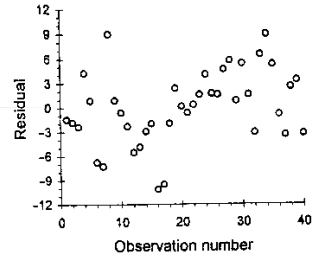


FIGURE 6.1.3
Index plot of residuals for regression of concrete strengths and densities.

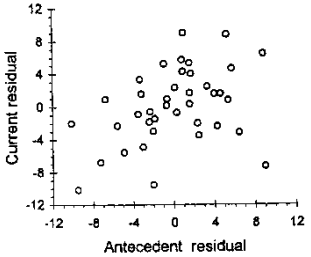


FIGURE 6.1.4
Plot of residual versus antecedent residual.

Simple Linear Regression Example (3)

As given earlier, $S_{xx} = 9977; S_{xy} = 1365; S_{yy} = 980.8$. From Eq. (6.1.14*d*) the variance of the errors is estimated as

$$\hat{\sigma}^2 = \frac{1}{38}\left(980.8 - \frac{1365^2}{9977}\right) = 20.89.$$

Thus the estimated standard deviation is $\hat{\sigma} = \sqrt{20.89} = 4.57$. Only 2 of the 40 residuals are outside the range $-2\hat{\sigma}$ to $+2\hat{\sigma}$, that is, from -9.14 to $+9.14$. The coefficient of skewness of the residuals is -0.1633, and the coefficient of kurtosis is 2.79. A normal probability plot is drawn as shown in Fig. 6.1.5. It does not indicate any outliers or untoward behavior. From the graph and the foregoing statistics we see that the distribution of the residuals is close to normality. Figure 6.1.6 is a plot of residuals $\hat{\varepsilon}_i$ against the densities of concrete x_i. The dispersion of the points indicates that the errors, as represented by the residuals, are independent of the explanatory variable, which is one of the assumptions made. Likewise, we can infer the independence by producing a plot of the residuals $\hat{\varepsilon}_i$ against the fitted values $\hat{y}_i$. It is reasonable to assume that the variance is constant; that would not be the case if there were a much larger spread above and below one part of the horizontal axis than another.

Simple Linear Regression Example (4)

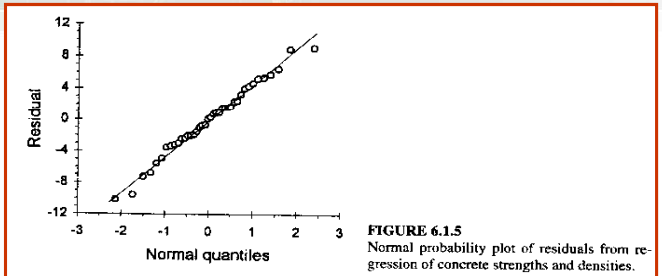


FIGURE 6.1.5
Normal probability plot of residuals from regression of concrete strengths and densities.

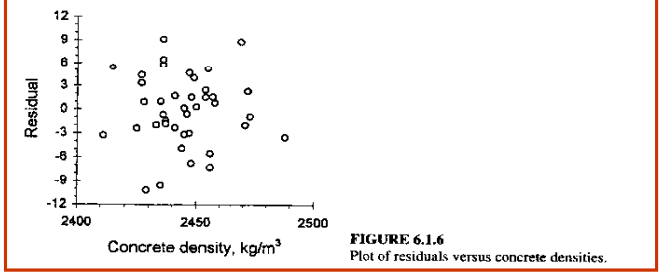


FIGURE 6.1.6
Plot of residuals versus concrete densities.

Test of Significance

- **Hypothesis**

$$H_0 : \beta_1 = \beta^* \qquad H_1 : \beta_1 \neq \beta^*$$

β_1: slope

β^*: constant

- **Test statistic**

$$T = \frac{\hat{\beta}_1 - \beta^*}{\sigma / \sqrt{S_{xx}}} \sim t_{n-2}$$

Confidence Intervals

- **100(1-α) percent confidence interval for β_1**

$$\Pr[\hat{\beta}_1 - t_{n-2,\alpha/2}\hat{\sigma}/\sqrt{S_{xx}} < \beta_1 < \hat{\beta}_1 + t_{n-2,\alpha/2}\hat{\sigma}/\sqrt{S_{xx}}] = 1-\alpha$$

- **100(1-α) percent confidence interval for the mean value of Y**

$$\Pr[\hat{\beta}_0 + \hat{\beta}_1 x_0 - t_{n-2,\alpha/2}\hat{\sigma}/\sqrt{(1/n)+(x_0-\bar{x})^2/S_{xx}} < \beta_0 + \beta_1 x_0$$
$$< \hat{\beta}_0 + \hat{\beta}_1 x_0 + t_{n-2,\alpha/2}\hat{\sigma}/\sqrt{(1/n)+(x_0-\bar{x})^2/S_{xx}}] = 1-\alpha$$

- **100(1-α) percent prediction interval for a future value of Y, Y_0**

$$\Pr[\hat{\beta}_0 + \hat{\beta}_1 x_0 - t_{n-2,\alpha/2}\hat{\sigma}/\sqrt{1+(1/n)+(x_0-\bar{x})^2/S_{xx}} < Y_0$$
$$< \hat{\beta}_0 + \hat{\beta}_1 x_0 + t_{n-2,\alpha/2}\hat{\sigma}/\sqrt{1+(1/n)+(x_0+\bar{x})^2/S_{xx}}] = 1-\alpha$$

Simple Linear Regression Example(2)

Example 6.3. Test of significance of simple linear regression model applied to concrete data. This is effectively a test on the slope parameter. The null and alternate hypotheses are

$H_0: \beta_1 = 0,$
$H_1: \beta_1 \neq 0,$
Level of significance $\alpha = .05.$

Calculations: $n = 40$, $S_{xx} = 9977$, and from Examples 6.1 and 6.2, $\hat{\beta}_1 = 0.1368$, $\hat{\sigma}^2 = 20.89$, that is, $\hat{\sigma} = 4.57$. Under the null hypothesis it follows from Eq. (6.1.15) that the t statistic for the test is

$$\hat{\beta}_1 \sqrt{S_{xx}}/\hat{\sigma} = 0.1368 \times \sqrt{9978}/4.57 = 2.99.$$

From Table C.2, this has a probability of exceedance of around .999.

Decision: We reject the null hypothesis $\beta_1 = 0$.