11.3.3 Inferences on Regression Coefficients

We first compute the variances of coefficients *a* and *b* and the determine the confidence bands for these. Eq. (11.17a) gives

$b = S_{xy}/S_{xx}$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) / S_{xx} = \sum y_i (x_i - \bar{x}) / S_{xx} \quad =$$

Treating x_i as constent,

$$\text{var}(b) = \sigma_b^2 = \sum (x_i - \bar{x})^2 \, \text{var}(y_i) / S_{xx}^2 = S_{xx} s^2 / S_{xx}^2$$
$$= s^2 / S_{xx}$$

Thus $\sigma_b = s/\sqrt{S_{xx}}$

So, the standard error of *b*, $S_b = s/\sqrt{S_{xx}}$.

The variance of coefficient *a* can be computed by

$$\text{var}(a) = \text{var}(\bar{y} - b\bar{x}) = \text{var}(\bar{y}) - \bar{x}^2 \, \text{var}(b) \qquad \textit{because} \quad \text{cov}(\bar{y}, b) = 0$$
$$= s^2 / n + s^2 \bar{x} / S_{xx} = s^2 (1/n + \bar{x}^2 / S_{xx})$$

So, the standard error of *a*

$$S_a = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \qquad (11.26)$$

11.3.4 Confidence Intervals

In statestical analysis, the confidence interval at the α% significance level is determined suh that in repeated applications, the probability with which the confidence interval would contain the parameter value is (100 - α)%. Typically the value of α used analysis is 0.05 which corresponds to (1-0.05)*100% = 95% confidence band. In the computation of confidence interval, it is assumed that the variables are linear;ly related and the residuals e_i are independent, normally distributed random variables with constant variance. If the model is correct, then a/S_a and b/S_b should follow the *t* distribution with *(n-2)* degrees of freedom. Hence, for coefficient *a*, the lower and upper limits of the confidence intervel are:

$$(l_a, u_a) = \{a - t_{(1-\alpha/2),(n-2)} S_a, \; a + t_{(1-\alpha/2),(n-2)} S_a\} \qquad (11.27)$$

For coefficient *b*, the lower and upper limits are:

$$(l_b, u_b) = \{b - t_{(1-\alpha/2),(n-2)}\, S_b\ ,\ b + t_{(1-\alpha/2),(n-2)}\, S_b\} \qquad (11.28)$$

where $t_{(1-\alpha/2),\,(n-2)}$ represents Student's *t* values corresponding to the probability of exceedance $\alpha/2$ and *(n-2)* degrees of freedom.

Confidence intervals on the regression line

The depend on the variance of $\hat{\bar{y}}_k$ which is the predicted mean value of $\hat{y}_k$ for given x_k:

$$\hat{\bar{y}}_k = a + bx_k \qquad (11.29)$$

Then,

$$\begin{aligned} \text{var}(\hat{\bar{y}}_k) &= \text{var}(a) + x_k^2\, \text{var}(b) + 2x_k\, \text{cov}(a,b) \\ &= s^2\left[\frac{1}{n} + \frac{(x_k - \bar{x})^2}{S_{xx}}\right] \end{aligned} \qquad (11.30)$$

Hence, the standard error of $\hat{\bar{y}}_k$ would be

$$S_{\hat{\bar{y}}_k} = s\left[\frac{1}{n} + \frac{(x_k - \bar{x})^2}{S_{xx}}\right]^{1/2} \qquad (11.31)$$

So, the lower and upper confidence limits on the regression line are:

$$(L,U) = [\hat{\bar{y}}_k - S_{\hat{\bar{y}}_k} t_{(1-\alpha/2),(n-2)},\ \hat{\bar{y}}_k + S_{\hat{\bar{y}}_k} t_{(1-\alpha/2),(n-2)}] \qquad (11.32)$$

Example 11.3: The precipitation and runoff for a catchment for the month of August are given below in Table 11.1. (a) Develop the rainfall-runoff relationship in the form: $y = a + bx$; where *y* represents runoff and *x* represents precipitation. (b) What percent of the variation in runoff is accounted for by the developed regression equation?

Table 11.1 Precipitation runoff data and calculations.

SN	Year	Precip (x)	Runoff (y)	$x - \bar{x}$	$(y - \bar{y})$	$(x - \bar{x})*(y - \bar{y})$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$\hat{y}$	$S_{se} = (y - \hat{y})^2$
1	1973	33.48	8.31	-0.036	-1.893	0.068	0.001	3.583	10.917	6.799
2	1974	47.67	15.17	-8.946	-6.843	61.216	80.033	46.824	17.635	6.078
3	1975	50.24	13.55	5.244	0.017	0.090	27.498	0.000	18.852	28.112

4	1976	43.28	14.22	7.814	-1.603	-12.524	61.057	2.569	15.557	1.788
5	1977	42.39	13.26	0.854	-0.933	-0.796	0.729	0.870	15.136	3.518
6	1978	52.57	21.21	10.144	6.057	61.444	102.898	36.690	19.955	1.575
7	1979	31.06	10.72	-11.366	-4.433	50.383	129.188	19.650	9.772	0.899
8	1980	50.02	17.64	7.594	2.487	18.888	57.667	6.186	18.748	1.227
9	1981	47.08	23.91	4.654	8.757	40.755	21.659	76.689	17.356	42.955
10	1982	43.06	18.89	0.634	3.737	2.369	0.402	13.967	15.453	11.814
11	1983	40.89	12.82	-1.536	-2.333	3.583	2.360	5.442	14.426	2.578
12	1984	37.31	11.58	-5.116	-3.573	18.279	26.175	12.765	12.731	1.324
13	1985	33.15	15.17	-9.276	0.017	-0.160	86.046	0.000	10.761	19.437
14	1986	40.38	10.12	-2.046	-5.033	10.298	4.187	25.329	14.184	16.517
15	1987	45.39	18.02	2.964	2.867	8.498	8.785	8.221	16.556	2.143
16	1988	41.03	16.25	-1.396	1.097	-1.532	1.949	1.204	14.492	3.091
17	1989	36.49	10.76	-5.936	-4.393	26.076	35.237	19.296	12.342	2.504
18	1990	48.18	21.15	5.754	5.997	34.507	33.107	35.967	17.877	10.714
	Sum	763.67	272.75	0.000	0.000	321.443	678.979	315.251	272.750	163.073
	Average	42.43	15.15	0.000	0.000	17.858	37.721	17.514	15.153	9.060

Solution: In regression anlysis, it is always helpful to first plot the data and note the variation in the dependent and independent variables. Fig. 11.5 gives a plot of the precipitation and runoff data which shows that there is not much scatter around the line of best fit.

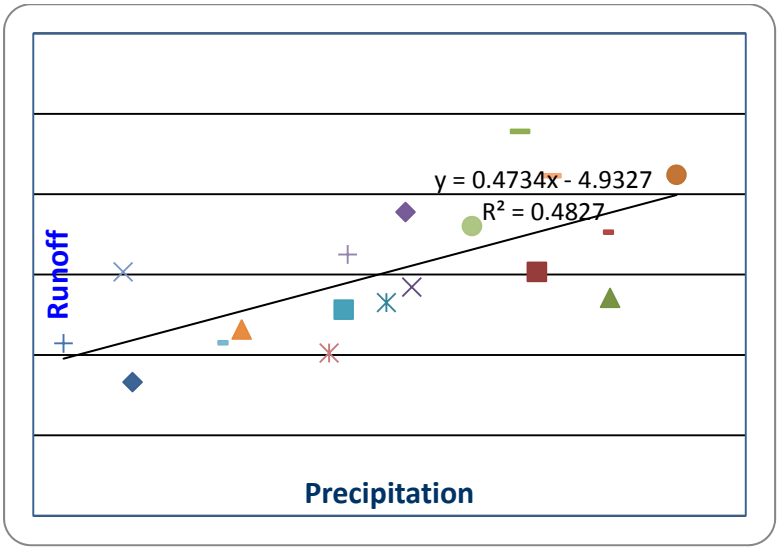

Fig. 11.5 Plot of the precipitation and runoff data.

(a) The values of various variables required to calculate *a* and *b* are computed in the Table 11.1. Here, $\bar{x}$ = 763.67/18 = 42.43, $\bar{y}$ = 272.75/18= 15.15. The regression coefficients are:

$b = S_{xy}/S_{xx}$ = 321.443/678.979=0.473

and $a = \bar{y} - b\,\bar{x}$ = 15.15-0.473*42.43=-4.933

Hence, the regression equation is: y = -4.933 + 0.473 x.

(b) The percent of variation in y that is accounted for by the regression is computed as the coefficient of determination (r^2) multiplied by 100. The value of S_{se} has been computed in Table 11.1.

Coefficient of determination $R^2 = 1 - S_{se} / S_{yy}$ = 1 – 163.073/315.251=0.483.

Thus, nearly 66 percent of variation in *y* is explained by the regression equation. The remaining 34 percent variation is due to unexplained causes.

The coefficient of correlation (r) = square root of coefficient of determination

= √ 0.483 = 0.695.

Example 11.4: Using the data of Example 11.3, (a) Compute the 95% confidence intervals on *a* and *b*, (b) test the hypothesis that *a* = 0 and the hypothesis that *b* = 0.50; (c) Calculate the 95% confidence limits for the regression line, (d) Calculate the 95% confidence interval for an individual predicted value of *y*.

Solution: (a) Computing 95% confidence intervals on *a* and *b*.

We first compute mean square error (Table 11.1): mse = S_{se} /(n-2) = 163.073/16=10.192.
Standard error of regression: $s_r = mse^{0.5} = 10.192^{0.5}$ = 3.192. This is a very useful indicator of the quality of regression relationship.

Standard error of *b* (S_b) = s_r /√ (S_{xx}) = 3.192 /√ 678.979 = 0.123.

Standard error of *a* (S_a) = $s_r\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}} = 3.192\sqrt{\frac{1}{18}+\frac{42.43^2}{678.979}} = 5.253$

From the t-table, for α = 0.05, n-2 = 16, $t_{(1-0.025),\,(18-2)} = t_{0.975,\,16}$ = 2.12.

So, 95% confidence intervals on *a*:

$(l_a, u_a) = \{ a - t_{(1-\alpha/2),(n-2)} . S_a, a + t_{(1-\alpha/2),(n-2)} . S_a\}$

$= (-4.933-2.12*5.253,-4.933+2.12*5.253)$

$= (-16.068,6.203).$

Similarly, 95% confidence intervals on *b*:

$(l_b , u_b) = \{b - t_{(1-\alpha/2),(n-2)} . S_b, b + t_{(1-a/2),(n-2)} . S_b\}$

$= (00.473-2.12*0.123,0.473+2.14*0.123)$

$= (0.214,0.473).$

(b) Testing the hypothesis H_o : $a = 0$ versus $a \neq 0$.

Here, $t = (a - 0.00)/S_a = -4.933/5.253 = -0.939$. Since $t_{(1-\alpha/2),(n-2)} = t_{0.975,16} = 2.12$, $|t| > t_{0.995,16}$. Hence, the null hypothesis H_o : $a = 0$ is rejected.

Hypothesis H_o : $b = 0.5$ versus H_a : $b \neq 0.5$.

In this case, $t = (b-0.5)/S_b$, $= (0.473-0.50)/0.123 = -0.217$. Since $|t| < 0.217$, the hypothesis H_o cannot be rejected.

From the above tests, it is concluded that the intercept is significantly different from zero. However, the slope is not significantly different from 0.5.

(c) The significance of the overall regression can be evaluated by testing H_o: $b = 0$. To test this hypothesis, the test statisties is computed as

$$t = \frac{b - 0.00}{S_b} = \frac{(0.473 - 0)}{0.123} = 3.86$$

Since $|t| > t_{0.975,16}$, we reject H_o and conclude that the regression equation is able to explain a significant amount of the variation in *Y*.

11.3.5 Extrapolation

Regression relation is frequently used for interpolation and extrapolation. However, the use of regression relation to extrapolate the dependent variable beyond the range of values of the independent variable used in estimating *a* and *b* may not be appeopriate under certain conditions. The confidence intervals on the regression line become wide as the point of interest moves away

from the mean of the independent variable. Second, the relation between the dependent and independent variables may be non-linear over the entire feasible range of the variable but it may be nearly linear for the range of data that was used in establishing the regression relation. An example of this behaviour is shown in Fig. 11.6.

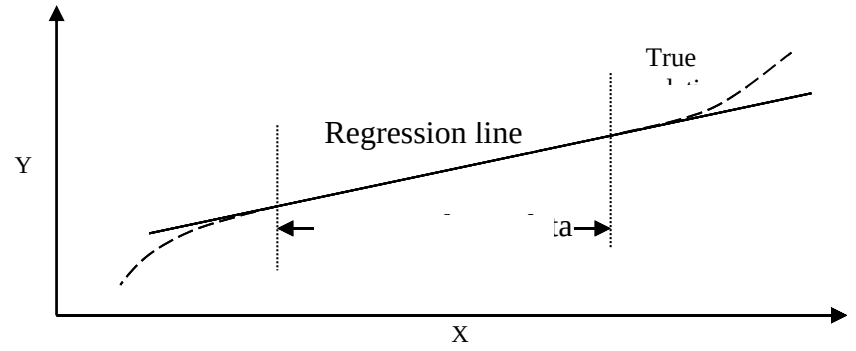

Fig. 11.6 Regression line and extrapolation.

Obviously, in this case, extrapolation will yield erroneous results and hence it should be attempted. Only after ascertaining that the assumption of linearity of relation helds good.

11.4 MULTIPLE LINEAR REGRESSION

Sometimes we may like to model the dependent variable as a function of two or more independent variables. For example, monthly runoff at the outlet of a catchment may be better modeled as a function of the rainfall in the current month and in the previous month(s). The association of more then two variables can be investigated by multiple linear regression. When the dependent variable are linear related with the independent variables, the regression is known as the multiple linear regression. A nonlinear association between the dependent and independent variable can be modeled by transforming the variables to linear form and applying the technique of multiple regression.

The general form of the multiple linear regression equation is:

$$y_i = b_1 x_{i,1} + b_2 x_{i,2} + \ldots + b_p x_{i,p} \qquad (11.33)$$

consider that a set of *n* observations is available on the dependent and each of the *p* independent variables. We will have *n* equations containing *p* unknown parameters.

$$y_1 = b_1 x_{1,1} + b_2 x_{1,2} + \ldots + b_p x_{1,p}$$
$$y_2 = b_1 x_{2,1} + b_2 x_{2,2} + \ldots + b_p x_{2,p} \qquad (11.34)$$

....

$y_n = b_1 x_{n,1} + b_2 x_{n,2} + ... + b_p x_{n,p}$

where y_i is the i^{th} observation of the dependent variable, $x_{i,1}$, $x_{i,2}$,…$x_{i,p}$ are the i^{th} observation on the independent variables and b_1, b_2, … b_p are the unknown parameters.

It is assumed here that n is much larger than p. If a regression equation with p parameters is fitted to a set of *n* observations of *p* variables, the degrees of freedom will be *n-p*. If the number of parameters p is equal to the sample size n, the regression equation will pass through all the points as there is no degree of freedom. Such an equation will not be suitable for prediction as the errors of parameters are inversely proportional to the degrees of freedom.

While establishing the multiple linear and nonlinear regression relations, the selection of dependent and independent variables is very important. The dependent variable is defined by the problem itself. The independent variables are selected based on the following considerations:

i. An analysis of physical phenomenon should indicate a cause-and-effect relation between dependent and independent variables.
ii. The variables should have been observed in the past concurrently with the dependent variable so that the regression equation may be established.
iii. There should be no plan to discontinue observing these in future, they may be used to predict the dependent variable.

The variables that are known to have little or no effect on the dependent variable are neglected. In matrix notation, eq. (11.34) can be written as:

$$\frac{Y}{n*1} = \frac{X}{n*p}\frac{B}{p*1} \qquad (11.35)$$