11.4.1 Estimation of Multiple Regression Coefficients

In multiple linear regression, we essentially solve *n* equations for the p unknown parameters. Thus *n* must be equal to or greater than *p* and in practice *n* should be at least 3 or 4 times as large as *p*. The difference between the observed and predicted value of *y* (using regression) or the error is = $y_i - \hat{y}_i$. The regression coefficients are obtained by minimizing the sum of squares of errors.

In matrix form, the *n* equations can be written as

$$Y = Xb + e \qquad (11.36)$$

where **Y** = (n x 1) column vector of the dependent variable, **X** = (n x p) matrix of independent variables, b = (p x 1) column vector of the regression coefficients, and e = (n x 1) column vector of residuals. The residuals are conditioned by:

$$E[\mathbf{e}] = 0 \qquad (11.37)$$

$$Cov(e) = \sigma_\varepsilon^2 I \qquad (11.38)$$

where **I** = (n x n) diagonal identity matrix with diagonal elements = 1 and off-diagonal elements = 0; and σ_e^2 = variance of (Y|X).

According to the least squares principle the estimates of regression parameters are those which minimize the residual sum of squares $\mathbf{e^T e}$. Hence

$$e^T e = (Y - Xb)^T (Y - Xb) \qquad (11.39)$$

is differentiated with respect to **b**, and the resulting expression is set equal to zero. This gives:

$$X^T Xb = X^T Y \qquad (11.40)$$

which are called the normal equations. Multiplying both sides with $(X^T X)^{-1}$ leads to an explicit expression for b:

$$\frac{b}{(p*1)} = \frac{(X^T X)^{-1}}{(p*n)(n*p)} \frac{X^T}{(p*n)} \frac{Y}{(n*1)}$$

(11.41)

Note that the independent variables should be chosen such that none of these is a linear combination of other independent variables. The properties of the estimator **b**:

$$Cov(b) = \sigma_{\varepsilon}^{2}\left(X^{T}X\right)^{-1} \tag{11.42}$$

By (21) and (22) the total adjusted sum of squares $\mathbf{Y^TY}$ can be partitioned into an explained part due to regression and an unexplained part about regression, as follows:

$$Y^{T}Y = b^{T}X^{T}Y + e^{T}e. \tag{11.43}$$

where $\mathbf{(Xb)^TY}$ = sum of squares due to regression; $\mathbf{e^Te}$ = sum of squares about regression**.**

This equation states:

Total sum of squares about mean = regression sum of squares + residual sum of Squares

The mean squares values of the right hand side terms in (11.43) are obtained by dividing the sum of squares by their corresponding degrees of freedom. If **b** is a (p x 1)-column vector, i.e. there are p-independent variables in regression, then the regression sum of squares has p-degrees of freedom. Since the total sum of squares has (n-1)-degrees of freedom (note: 1 degree of freedom is lost due to the estimation of $\bar{y}$), it follows by subtraction that the residual sum of squares has (n-1-p)-degrees of freedom. It can be shown that the residual mean square S_e^2:

$$S_e^2 = \frac{e^{T}e}{n-1-p} \tag{11.44}$$

Is an unbiased estimate of σ_ε^2 .The estimate se of σ_ε is the standard error of estimate.

The analysis of variance (ANOVA) table (see Table 11.2) summarizes the sum of squares quantities.

Table 11.2: Analysis of variance table (ANOVA)

Source	Sum of squares	Degrees of freedom
Total	$S_Y = Y^TY$	n

Mean	$n\bar{Y}^2$	1
Regression	$b^T X^T Y - n\bar{Y}^2$	p-1
Residual	$Y^T Y - b^T X^T Y$	n-p

As for the simple linear regression a measure for the quality of the regression equation is the coefficient of determination, defined as the ratio of the explained or regression sum of squares and the total adjusted sum of squares.

$$R_m^2 = \frac{b^T X^T Y}{Y^T Y} = 1 - \frac{e^T e}{Y^T Y} \qquad (11.45)$$

11.4.2 Confidence Intervals on the Regression Line

To place confidence limits on Y_0 where $Y_0 = \mathbf{X}_0\mathbf{b}$ it is necessary to have an estimate for the variance of $\hat{Y}_0$. Considering Cov(**b**) as given in (25) the variance Var($\hat{Y}_0$) is given by:

$$Var\left(\hat{Y}_0\right) = S_e^2 X_0 \left(X^T X\right)^{-1} X_0^T \qquad (11.46)$$

The confidence limits for the mean regression equation are given by

$$CL_{\pm} = X_0 b \pm t_{1-a/2,n-p}\sqrt{Var\left(\hat{Y}_0\right)} \qquad (11.47)$$

Coefficient of Determination (R^2)

Let $Z_{i,j} = (X_{i,j} - \bar{x}_j)/S_j$ (11.48)

where $\bar{x}_j$ and S_j are the mean and standard deviation of the j^{th} independent variable. The correlation matrix is:

$R = Z^T Z/(n-1) = [R_{i,j}]$ (11.49)

where $R_{i,j}$ is the correlation between the i^{th} and j^{th} independent variables. R is a symmetric matrix since $R_{i,j} = R_{j,i}$. The coefficient of determination is defined as

R^2 = Sum of squares due to regression / Sum of squares about mean

or $R^2 = (b^T X^T Y - n\bar{Y}^2)/(Y^T Y - n\bar{Y}^2)$ (11.50)

$(L_{\hat{b}_i}, U_{\hat{b}_i}) = (\hat{b}_i - t_{(1-\alpha/2)(n-p)} S_{\hat{b}_i}, \hat{b}_i + t_{(1-\alpha/2)(n-p)} S_{\hat{b}_i})$ Here b^T is the transpose of vector b of size $(1xp)$, and Y^T is the transpose of vector Y of size $(1xn)$. Let the residual error be $\varepsilon = Y - X\,b$. R^2 is the part of the total sum of squares conceted for mean that is explain by the regression equation. It ranges between 0 and 1 and closer it is to1, the better is the regression.

11.4.3 Inferences on Regression Coefficients

(i) Confidence intervals on b_i

Assuming that the model is correct, the quantity $\hat{b}_i / S_{\hat{b}_i}$ follows a t-distribution with $(n-p)$ degrees of freedom. The confidence intervals on b_i are given as

(11.51)

(ii) Test of hypothesis concerning b_i

The hypothesis that the i^{th} variable is not contributing significantly to explaining the variation in the dependent variable is equivalent to testing the hypothesis $H_o : b_i = b_o$ versus $H_a : b_i \neq b_o$. The test is conducted by computing:

$$t = (\hat{b}_i - b_o) / S_{\hat{b}_i} \qquad (11.52)$$

Null hypothesis H_o is rejected if $|t| > t_{(1-\alpha/2),\,(n-p)}$. If this hypothesis is accepted, it is advisable to delete the concerned variable from the regression model.

Significance of the overall regression

The null hypothesis $H_o : b_1 = b_2 = \ldots b_p = 0$ versus H_a : at least one of these b's is not zero is used to test whether the regression equation is able to explain a significant amount of variation of Y or not. The ratio of the mean square error due to regression to the residual mean square has an F distribution with $p-1$ and $n-p$ degrees of freedom. Hence, the hypothesis is tested by computing the test statistic:

$$F = \frac{(b^T X^T Y - n\bar{Y}^2)/(p-1)}{(Y'Y - \hat{b}'X'Y)/(n-p)}$$

(11.53)

H_o is rejected if F exceeds the critical value $F_{(1-\alpha),\,(p-1),\,(n-p)}$.

Confidence Intervals on Regression Line:

To put the confidence limits on $Y_k = X_k b$, it is necessary to estimate the variance of $\hat{y}_k$. This is given by

$$S^2_{\hat{Y}_k} = S^2 X_k (X'X)^{-1} X'_k \qquad (11.54)$$

where $(L,U) = \{\hat{Y}_k - t_{(1-\alpha/2)(n-p)} S_{\hat{Y}_k}, \hat{Y}_k + t_{(1-\alpha/2)(n-p)} S_{\hat{Y}_k}\}$

Confidence Intervals on Individual Predicted Value of Y

$$\hat{Y}_K = X_K \hat{b} \qquad (11.55)$$

$$(L',U') = \{\hat{Y}_k - t_{(1-\alpha/2)(n-p)} S'_{\hat{Y}_k}, \hat{Y}_k + t_{(1-\alpha/2)(n-p)} S'_{\hat{Y}_k}\} \qquad (11.56)$$

$$S'^2_{\hat{Y}_k} = S^2[1 + X_k (X'X)^{-1} X'_k]$$

Example 11.5: Table contains rainfall for the months of July and August and discharge for the August month for a catchment. Estimate the parameters of linear regression and multiple linear regression and find out if there is an advantage in using multiple linear regression in this case.

Table 11.3 Data and computations for multiple linear regression example

YEAR	RF-JUL (MCM)	RF-AUG (MCM)	Obs Q Aug (MCM)	Comp. Q by Lin Reg (Q_L)	(Qob-Q_L)^2	Comp. Q by Mult. Lin Reg (Q_M)	(Qob-Q_M)^2
1982	5020.04	15664.05	5996.939	6830.0	694015.6	6873.0	767532
1983	7980.13	6546.24	2557.916	3263.6	497987.7	3572.3	1028983
1984	3002.36	13086.63	4395.515	5821.9	2034467.0	4736.5	116242
1985	8572.75	7532.13	5725.02	3649.2	4308915.7	4314.0	1990914
1986	5242.03	5799.34	2532.373	2971.4	192787.2	2045.2	237329
1987	6311.05	9522.80	2774.517	4427.9	2733589.2	4353.5	2493329
1988	6040.00	7285.46	4163.013	3552.7	372427.7	3123.1	1081472

1989	1597.33	6922.49	2046.694	3410.8	1860702.7	1068.8	956235
1990	8561.71	6889.43	4190.084	3397.8	627652.7	3988.7	40541
1991	7153.31	12566.82	6107.452	5618.5	239036.6	6227.3	14365
1992	5623.67	10263.08	5145.44	4717.4	183188.8	4433.0	507510
1993	4233.30	7108.91	2300.774	3483.7	1399281.8	2273.2	759
1994	13076.88	10472.23	8994.085	4799.2	17596705.8	7679.9	1727088
1995	6843.64	8068.47	3695.11	3859.0	26865.5	3852.6	24788
1996	7819.49	9330.16	4870.4	4352.5	268196.6	4893.4	531
1997	9403.82	7424.92	3943.455	3607.3	113005.8	4610.9	445541
1998	7040.85	8306.55	3801.727	3952.1	22624.6	4054.5	63883
1999	7380.56	9987.30	5895.899	4609.6	1654653.4	5036.2	739043
2000	8620.28	4283.79	1501.445	2378.6	769480.6	2713.5	1469074
2001	9113.46	5071.52	2670.739	2686.8	256.8	3314.4	414338
2002	1296.93	11168.68	3192.95	5071.7	3529547.4	3060.5	17531
2003	7493.84	7784.62	3708.33	3748.0	1572.8	3985.1	76593
Sum	147427.41	191085.61	90209.88	90209.88	3.91E+07	9.02E+04	1.42E+07
Average	6701.25	8685.71	4100.45	4100.449			

Solution: Using the data given in the table, linear regression equation of the following form was established between the rainfall and observed discharge for the August month.

$Q_A = a + b\ R_A$ where Q_A = discharge for August and R_A = rainfall for August.

The parameters a and b were estimated to yield the following equation:

$Q_A = 703.05 + 0.391\ R_A$

Coefficient of determination $R^2 = 1 - 3.91*10^7/6.33*10^7 = 0.382$.

Next, discharge for the August month was computed by using the above equation and the sum of square of residuals turned out to be $3.91*10^7$.

In case of multiple linear regression, the independent variables were the rainfall for the month July and August and dependent variable as the discharge for August. Regression equation of the following form was envisaged

$Q_A = a + b_1 R_J + b_2 R_A$

where R_J = rainfall for July month.

After computations, the following regression equation was obtained.

$$Q_A = -3058.24 + 0.42\ R_J + 0.50\ R_A$$

Coefficient of determination $R^2 = 1 - 1.42*10^7/6.33*10^7 = 0.78$.

The discharge for August was computed by LR and MLR equations and the sum of squares of errors were computed. The values were $3.91*10^7$ for LR and $1.42*10^7$ for MLR. When these values are compared along with R^2 for the two cases, it can be concluded that MLR gives much improved estimates of the discharge compared to LR.