

Chapter 9

Multiple Linear Regression Analysis

L1 – Multiple Linear Regression



Lecture 3

Learning Outcomes:

At the end of the lesson, the student should be able to

- Use the least squares method to estimate a multiple linear model
- Carry out tests to determine if the model obtained is an adequate fit to the data
- Carry out test for inferences on regression parameters
- Find the CI for the slope

MULTIPLE LINEAR REGRESSION

- an extension of a simple linear regression model
- allows the dependent variable y to be modeled as a linear function of **more than one** input variable x_i
- Consider the following data consisting of n sets of values

$$(y_1, x_{11}, x_{21}, \dots, x_{k1})$$

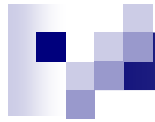
$$(y_2, x_{12}, x_{22}, \dots, x_{k2})$$

.

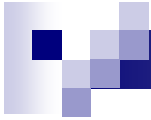
$$(y_n, x_{1n}, x_{2n}, \dots, x_{kn})$$

- the value of the dependent variable y_i is modeled as

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon$$




- the dependent variable is related to k independent or regressor variables
- the multiple linear regression model can provide a rich variety of functional relationships by allowing some of the input variables x_i to be functions of other input variables.
- As in simple linear regression, the parameters $\beta_0, \beta_1, \dots, \beta_k$ are estimated using the **method of least squares**.
- However, it would be tedious to find these values by hand, thus we use the computer to handle the computations.
- the ANOVA is used to test for significance of regression
- the t - test is used to test for inference on individual regression coefficient



Data for Multiple Linear Regression

y	x_1	x_2	\dots	x_k
y_1	x_{11}	x_{12}	\dots	x_{1k}
y_2	x_{21}	x_{22}	\dots	x_{2k}
\vdots	\vdots	\vdots	\vdots	\vdots
y_n	x_{n1}	x_{n2}	\dots	x_{nk}

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \epsilon_i \\ &= \beta_0 + \sum_{j=1}^k \beta_j x_{ij} + \epsilon_i \quad i = 1, 2, \dots, n \end{aligned}$$



$$L = \sum_{i=1}^n \epsilon_1^2 = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^k \beta_j x_{ij} \right)^2$$

The **least squares estimates** of $\beta_0, \beta_1, \dots, \beta_k$ must satisfy

$$\left. \frac{\partial L}{\partial \beta_0} \right|_{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \sum_{j=1}^k \hat{\beta}_j x_{ij}) = 0$$

$$\left. \frac{\partial L}{\partial \beta_j} \right|_{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \sum_{j=1}^k \hat{\beta}_j x_{ij}) x_{ij} = 0 \quad j = 1, 2, \dots, k$$

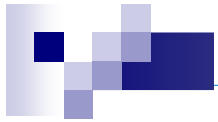


Least squares normal equations :

[illegible]

For two independent variables, X1 and X2,
the normal equations :

$$\begin{aligned} n\beta_0 + \beta_1 \sum X_1 + \beta_2 \sum X_2 &= \sum Y \\ \beta_0 \sum X_1 + \beta_1 \sum X_1^2 + \beta_2 \sum X_1 X_2 &= \sum X_1 Y \\ \beta_0 \sum X_2 + \beta_1 \sum X_1 X_2 + \beta_2 \sum X_2^2 &= \sum X_2 Y \end{aligned}$$



Example 1:
pg. 310

Observation Number	Pull Strength y	Wire Length x_1	Die Height x_2
1	9.95	2	50
2	24.45	8	110
3	31.75	11	120
4	35.00	10	550
5	25.02	8	295
6	16.86	4	200
7	14.38	2	375
8	9.60	2	52
9	24.35	9	100
10	27.50	8	300
11	17.08	4	412
12	37.00	11	400
13	41.95	12	500
14	11.66	2	360
15	21.65	4	205
16	17.89	4	400
17	69.00	20	600
18	10.30	1	585
19	34.93	10	540
20	46.59	15	250
21	44.88	15	290
22	54.12	16	510
23	56.63	17	590
24	22.13	6	100
25	21.15	5	400



We calculate

$$n = 25, \sum_{i=1}^{25} y_i = 725.82, \sum_{i=1}^{25} x_{i1} = 206, \sum_{i=1}^{25} x_{i2} = 8,294$$

$$\sum_{i=1}^{25} x_{i1}^2 = 2,396, \sum_{i=1}^{25} x_{i2}^2 = 3,531,848$$

$$\sum_{i=1}^{25} x_{i1}x_{i2} = 77,177, \sum_{i=1}^{25} x_{i1}y_i = 8,008.47, \sum_{i=1}^{25} x_{i2}y_i = 274,816.71$$



$$25\hat{\beta}_0 + 206\hat{\beta}_1 + 8,294\hat{\beta}_2 = 725.82$$

$$206\hat{\beta}_0 + 2,396\hat{\beta}_1 + 77,177\hat{\beta}_2 = 8,008.47$$

$$8,294\hat{\beta}_0 + 77,177\hat{\beta}_1 + 3,531,848\hat{\beta}_2 = 274,816.71$$

$$\hat{\beta}_0 = 2.26379, \hat{\beta}_1 = 2.74427, \hat{\beta}_2 = 0.01253$$

Using these estimated model parameters, the fitted regression equation is

$$\hat{y} = 2.26379 + 2.74427x_1 + 0.01253x_2 .$$

MULTIPLE LINEAR REGRESSION ANALYSIS

- testing for the significance of regression.

Hypotheses: $H_0 : \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$
 $H_1 : \text{at least one } \beta_j \neq 0$

Test statistic: $F_0 = \frac{MS_R}{MS_E}$

where: $MS_R = \frac{SSR}{k}$ $MS_E = \frac{SSE}{n - p}$

Rejection criteria: $F_0 = \frac{MS_R}{MS_E} > f_{\alpha, k, n-p}$

ANOVA Table for multiple linear regression

Source Of variation	Sum of Squares	Degrees of freedom (df)	Mean Square (Sum of squares / df)	Computed F
Regression	SSR	k	$MS_R = \frac{SSR}{k}$	$F = MS_R / MS_E$
Error	SSE	$n - (k+1)$	$MS_E = \frac{SSE}{n - (k + 1)}$	
Total	SST	$n - 1$		

Inferences on the model parameters in multiple regression.

- The hypotheses are $H_0 : \beta_j = \beta_{j,0}$

$$H_1 : \beta_j \neq \beta_{j,0}$$

- Test statistic

$$T_0 = \frac{\hat{\beta}_j - \beta_{j,0}}{se(\hat{\beta}_j)}$$

- Reject H_0 if $|T_0| > t_{\alpha/2, n-p}$

- A 100 (1 - α)% CI for an individual regression coefficient is

$$\hat{\beta}_j - t_{\alpha/2, n-(k+1)} (se)(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2, n-(k+1)} (se)(\hat{\beta}_j)$$

Regression Analysis: y versus x_1, x_2

The regression equation is Strength =

Predictor	Coef	SE Coef	T	P-value
Constant	2.264 $\leftarrow \beta_0$	1.060	?	
x_1	2.74427 $\leftarrow \beta_1$	0.09	?	
x_2	0.0125 $\leftarrow \beta_2$	0.002	?	

S = ?

R-Sq = ?

Reject H_0 if $|T| > t_{\alpha/2, n-p}$

Analysis of Variance

Source	DF	SS	MS	F	P-value
Regression	2	5990.4	2995.4	572.17	0.000
Residual Error	22	115.2	5.2		
Total	24	6105.9			

σ^2

Regression Analysis: y versus x_1, x_2

The regression equation is $\text{Strength} = 2.26 + 2.74 \text{ Wire Ln} + 0.0125 \text{ Die Ht}$

Predictor	Coef	SE Coef	T	P-value
Constant	2.264	1.060	2.14	0.044
x_1	2.74427	0.09	30.49	0.000
x_2	0.0125	0.002	6.25	0.000

$S = 2.288$ $R\text{-Sq} = 98.1\%$

Analysis of Variance

Source	DF	SS	MS	F	P-value
Regression	2	5990.4	2995.4	572.17	0.000
Residual Error	22	115.2	5.2		
Total	24	6105.6			

Reject H_0 if $F_0 > f_{\alpha, k, n-(p)}$



Example 2:

A set of experimental runs were made to determine a way of predicting cooking time y at various levels of oven width $x1$, and temperature $x2$. The data were recorded as follows:

y	$x1$	$x2$
6.4	1.32	1.15
15.05	2.69	3.4
18.75	3.56	4.1
30.25	4.41	8.75
44.86	5.35	14.82
48.94	6.3	15.15
51.55	7.12	15.32
61.5	8.87	18.18
100.44	9.8	35.19
111.42	10.65	40.4

a) Estimate the multiple linear regression equation for the data.



Solution:

	Y	X1	X2	X1-square	X2-square	X1X2	X1Y	X2Y
	6.4	1.32	1.15	1.7424	1.3225	1.518	8.448	7.36
	15.05	2.69	3.4	7.2361	11.56	9.146	40.4845	51.17
	18.75	3.56	4.1	12.6736	16.81	14.596	66.75	76.875
	30.25	4.41	8.75	19.4481	76.5625	38.5875	133.4025	264.6875
	44.86	5.35	14.82	28.6225	219.6324	79.287	240.001	664.8252
	48.94	6.3	15.15	39.69	229.5225	95.445	308.322	741.441
	51.55	7.12	15.32	50.6944	234.7024	109.0784	367.036	789.746
	61.5	8.87	18.18	78.6769	330.5124	161.2566	545.505	1118.07
	100.44	9.8	35.19	96.04	1238.336	344.862	984.312	3534.484
	111.42	10.65	40.4	113.4225	1632.16	430.26	1186.623	4501.368
TOTAL	489.16	60.07	156.46	448.2465	3991.121	1284.037	3880.884	11750.03



Least squares normal equations :

$$10\beta_0 + 60.07\beta_1 + 156.46\beta_2 = 489.16$$

$$60.07\beta_0 + 488.2465\beta_1 + 1284.037\beta_2 = 3880.884$$

$$156.46\beta_0 + 1284.037\beta_1 + 3991.121\beta_2 = 11750.03$$

By solving this normal equations :

$$\beta_0 = 0.568, \beta_1 = 2.706, \beta_2 = 2.051$$

Then the estimated regression model is :

$$\Rightarrow \hat{Y} = 0.568 + 2.706X_1 + 2.051X_2$$

Solution:

Using the computer for computations, the following results were observed.

Regression Analysis: cooking time versus width, temperature

The regression equation is ?

Predictor	Coef	SE Coef	T	P
Constant	0.568	0.585	0.970	0.364
Width	2.706	0.194	?	0.000
Temp	2.051	0.046	?	0.000

S = ? R-Sq = ? R-Sq(adj) = 100%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	?	10953.334	5476.667	13647.872	0.000
Residual Error	7	2.809	0.401		
Total	?	10956.143			

Solution:

Using the computer for computations, the following results were observed.

Regression Analysis: cooking time versus width, temperature

The regression equation is

Cooking time = 0.568 + 2.706 width + 2.051 temperature

Predictor	Coef	SE Coef	T	P
Constant	0.568	0.585	0.970	0.364
Width	2.706	0.194	13.935	0.000
Temp	2.051	0.046	44.380	0.000

S = 0.6334 R-Sq = 99.9% R-Sq(adj) = 100%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	10953.334	5476.667	13647.872	0.000
Residual Error	7	2.809	0.401		
Total	9	10956.143			



Example 2 (continued):

b) Test whether the regression explained by the model obtained in part (a) is significant at the 0.01 level of significance.

Solution:

We use ANOVA to test for significance of regression

The hypotheses are

$$H_0 : \beta_1 = \beta_2 = 0$$

$$H_1 : \beta_1 \text{ and } \beta_2 \text{ are not both zero}$$

The test statistic is

$$F_0 = \frac{MSR}{MSE}$$



The following ANOVA table is obtained:

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	10953.334	5476.667	13647.872	0.000
Residual Error	7	2.809	0.401		
Total	9	10956.143			

Taking the significance $\alpha = 1\% = 0.01$

$$f_{\alpha, k, n-(k+1)} = f_{0.01, 2, 7} = 9.55$$

$$F = 13647.87 > 9.55$$

Decision : Reject H_0 , A linear model is fitted

QUIZ 5 (5/8/2011)

Using the computer for computations, the following results were observed.

Regression Analysis: Y versus X1, X2, X3

The regression equation is ?

Predictor	Coef	SE Coef	T	P
Constant	1470	5746	0.26	0.801
X1	0.8145	0.5122	?	0.131
X2	0.8204	0.2112	?	?
X3	13.529	6.586	2.05	0.057

S = ? **R-Sq = ?** **R-Sq(adj) = 87.8%**

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	?	8779676741	?	?	0.000
Residual Error	?	1003491259	?		
Total	19	?			

- Answer all questions.
- Is the regression model significant at alpha 0.05? Why?