Multiple Linear Regression

Osama Bin Ajaz
Lecturer at FAST – NUCES

Multiple Regression Model

- In **multiple regression,** there are several independent variables and one dependent variable, and the equation is:

$$y' = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

Normal equations for Regression Coefficient for two independent variables

$$\sum Y = na + b_1 \sum X_1 + b_2 \sum X_2$$

$$\sum X_1 Y = a \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2$$

$$\sum X_2 Y = a \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2$$

Simultaneous solution to above normal equations give values of b_1 & b_2 as:

$$b_1 = \frac{(\sum X_1 Y)\left(\sum X_2^2\right) - (\sum X_2 Y)(\sum X_1 X_2)}{\left(\sum X_1^2\right)\left(\sum X_2^2\right) - (\sum X_1 X_2)^2}$$

$$b_2 = \frac{(\sum X_2 Y)\left(\sum X_1^2\right) - (\sum X_1 Y)(\sum X_1 X_2)}{\left(\sum X_1^2\right)\left(\sum X_2^2\right) - (\sum X_1 X_2)^2}$$

Normal equations for K independent variables

$$
\begin{aligned}
nb_0 + b_1 \sum_{i=1}^{n} x_{1i} \quad &+ b_2 \sum_{i=1}^{n} x_{2i} \quad + \cdots + b_k \sum_{i=1}^{n} x_{ki} \quad = \sum_{i=1}^{n} y_i \\
b_0 \sum_{i=1}^{n} x_{1i} + b_1 \sum_{i=1}^{n} x_{1i}^2 \quad &+ b_2 \sum_{i=1}^{n} x_{1i}x_{2i} + \cdots + b_k \sum_{i=1}^{n} x_{1i}x_{ki} = \sum_{i=1}^{n} x_{1i}y_i \\
&\vdots \qquad\qquad \vdots \qquad\qquad\quad \vdots \qquad\qquad\quad \vdots \qquad\qquad \vdots \\
b_0 \sum_{i=1}^{n} x_{ki} + b_1 \sum_{i=1}^{n} x_{ki}x_{1i} &+ b_2 \sum_{i=1}^{n} x_{ki}x_{2i} + \cdots + b_k \sum_{i=1}^{n} x_{ki}^2 \quad = \sum_{i=1}^{n} x_{ki}y_i
\end{aligned}
$$

Here a = b_o

Example # 01

- For example, suppose a teacher wishes to see whether there is a relationship between a student's grade point average, age, and score on the state board examination. The two independent variables are GPA (denoted by $x1$) and age (denoted by x_2).
-

Student	GPA, x_1	Age, x_2	State board score, y
A	3.2	22	550
B	2.7	27	570
C	2.5	24	525
D	3.4	28	670
E	2.2	23	490

Example # 01

- The multiple regression obtained from the data is:

$$y' = -44.572 + 87.679x_1 + 14.519x_2$$

- if a student has a GPA of 3.0 and is 25 years old, the student's predicted state board score is:

$$\begin{aligned} y' &= -44.572 + 87.679(3.0) + 14.519(25) \\ &= 581.44 \text{ or } 581 \end{aligned}$$

Assumptions for Multiple linear Regression

- $E(\epsilon) = 0$
- The regression equation is linear in the parameters.
- $V(\epsilon) = \sigma^2$.
- The population distribution of ϵ is normal.
- The values for the *y variables are independent.*
- The independent variables are not correlated.

Multiple Correlation (R)

- The value of ***R*** *takes into account* all the independent variables.
- The value of ***R*** can range from 0 to 1; *R can never be negative.*
- The closer to 1, the stronger the relationship; the closer to 0, the weaker the relationship.

The formula for R is

$$R = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1} \cdot r_{yx_2} \cdot r_{x_1x_2}}{1 - r_{x_1x_2}^2}}$$

Example # 02

- For the data regarding state board scores, find the value of *R*.

$$R = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1} \cdot r_{yx_2} \cdot r_{x_1x_2}}{1 - r_{x_1x_2}^2}}$$

$$R = \sqrt{\frac{(0.845)^2 + (0.791)^2 - 2(0.845)(0.791)(0.371)}{1 - 0.371^2}}$$

$$R = \sqrt{\frac{0.8437569}{0.862359}} = \sqrt{0.9784288} = 0.989$$

Coefficient of Multiple Determination (R^2)

- R^2 is the *coefficient of multiple determination, and it is* the amount of variation explained by the regression model.

Significance of R

- An *F test is used to test the significance of R. The*

$H_0: \rho = 0$ and $H_1: \rho \neq 0$

The formula for the *F* test is

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)}$$

where n is the number of data groups ($x_1, x_2, \ldots, y$) and k is the number of independent variables.

The degrees of freedom are d.f.N $= n - k$ and d.f.D $= n - k - 1$.

Example # 03

- Test the significance of the *R obtained in Example 02 at* $\alpha = 0.05$

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)}$$

$$= \frac{0.978/2}{(1 - 0.978)/(5 - 2 - 1)} = \frac{0.489}{0.011} = 44.45$$

- The critical value obtained, d.f.N (υ_1)= 3, and d.f.D (υ_2)=5 - 2 -1 = 2 is 19.16. Hence, the decision is to reject the null hypothesis.

Polynomial Regression

Now suppose that we wish to fit the polynomial equation

$$\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_r x^r$$

to the n pairs of observations $\{(x_i, y_i);\ i = 1, 2, \ldots, n\}$. Each observation, y_i, satisfies the equation

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_r x_i^r + \epsilon_i$$

or

$$y_i = \hat{y}_i + e_i = b_0 + b_1 x_i + b_2 x_i^2 + \cdots + b_r x_i^r + e_i,$$

where r is the degree of the polynomial and ϵ_i and e_i are again the random error and residual associated with the response y_i and fitted value $\hat{y}_i$, respectively. Here, the number of pairs, n, must be at least as large as $r+1$, the number of parameters to be estimated.

Polynomial Regression (Contd.)

- Notice that the polynomial model can be considered a special case of the more general multiple linear regression model, where v$x_1 = x, x_2 = x^2, \ldots, x_r = x^r$

- The normal equations assume the same form as those given in previous slides (slides 3 & 4).

Example # 04

- Given the data

x	0	1	2	3	4	5	6	7	8	9
y	9.1	7.3	3.2	4.6	4.8	2.9	5.7	7.1	8.8	10.2

- fit a regression curve of the forn$\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2$ hen estimate $\mu_{Y|2}$.

Example # 04 (Contd.)

$$\begin{aligned} 10b_0 + \quad 45\,b_1 + \quad 285\,b_2 &= 63.7, \\ 45b_0 + \quad 285b_1 + \quad 2025\,b_2 &= 307.3, \\ 285b_0 + 2025\,b_1 + 15{,}333b_2 &= 2153.3. \end{aligned}$$

$$b_0 = 8.698, \quad b_1 = -2.341, \quad b_2 = 0.288.$$

$$\hat{y} = 8.698 - 2.341x + 0.288x^2.$$

When $x = 2$, our estimate of $\mu_{Y|2}$ is

$$\hat{y} = 8.698 - (2.341)(2) + (0.288)(2^2) = 5.168.$$

Example # 05

- The data given below represent the percent of impurities that resulted for various temperatures and sterilizing times during a reaction associated with the manufacturing of a certain beverage. Estimate the regression coefficients in the pc

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{11} x_{1i}^2 + \beta_{22} x_{2i}^2 + \beta_{12} x_{1i} x_{2i} + \epsilon_i,$$

for $i = 1, 2, \ldots, 18$.

Sterilizing Time, x_2 (min)	Temperature, x_1 (°C)		
	75	100	125
15	14.05	10.55	7.55
	14.93	9.48	6.59
20	16.56	13.63	9.23
	15.85	11.75	8.78
25	22.41	18.55	15.93
	21.66	17.98	16.44

Example # 05 (Contd.)

Using the normal equations, we obtain

$$b_0 = 56.4411, \quad b_1 = -0.36190, \quad b_2 = -2.75299,$$
$$b_{11} = 0.00081, \quad b_{22} = 0.08173, \quad b_{12} = 0.00314,$$

and our estimated regression equation is

$$\hat{y} = 56.4411 - 0.36190x_1 - 2.75299x_2 + 0.00081x_1^2 + 0.08173x_2^2 + 0.00314x_1x_2.$$

Practice Questions

12.1 A set of experimental runs was made to deter mine a way of predicting cooking time y at variou values of oven width x_1 and flue temperature x_2. Th coded data were recorded as follows:

y	x_1	x_2
6.40	1.32	1.15
15.05	2.69	3.40
18.75	3.56	4.10
30.25	4.41	8.75
44.85	5.35	14.82
48.94	6.20	15.15
51.55	7.12	15.32
61.50	8.87	18.18
100.44	9.80	35.19
111.42	10.65	40.40

Estimate the multiple linear regression equation

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

12.4 An experiment was conducted to determine if the weight of an animal can be predicted after a given period of time on the basis of the initial weight of the animal and the amount of feed that was eaten. The following data, measured in kilograms, were recorded:

Final Weight, y	Initial Weight, x_1	Feed Weight, x_2
95	42	272
77	33	226
80	33	259
100	45	292
97	39	311
70	36	183
50	32	173
80	41	236
92	40	230
84	38	235

(a) Fit a multiple regression equation of the form

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

(b) Predict the final weight of an animal having an initial weight of 35 kilograms that is given 250 kilograms of feed.

Practice Questions (Contd.)

12.5 The electric power consumed each month by a chemical plant is thought to be related to the average ambient temperature x_1, the number of days in the month x_2, the average product purity x_3, and the tons of product produced x_4. The past year's historical data are available and are presented in the following table.

y	x_1	x_2	x_3	x_4
240	25	24	91	100
236	31	21	90	95
290	45	24	88	110
274	60	25	87	88
301	65	25	91	94
316	72	26	94	99
300	80	25	87	97
296	84	25	86	96
267	75	24	88	110
276	60	25	91	105
288	50	25	90	100
261	38	23	89	98

(a) Fit a multiple linear regression model using the above data set.

(b) Predict power consumption for a month in which $x_1 = 75°\text{F}$, $x_2 = 24$ days, $x_3 = 90\%$, and $x_4 = 98$ tons.

Practice Questions (Contd.)

12.1 A set of experimental runs was made to determine a way of predicting cooking time y at various values of oven width x_1 and flue temperature x_2. The coded data were recorded as follows:

y	x_1	x_2
6.40	1.32	1.15
15.05	2.69	3.40
18.75	3.56	4.10
30.25	4.41	8.75
44.85	5.35	14.82
48.94	6.20	15.15
51.55	7.12	15.32
61.50	8.87	18.18
100.44	9.80	35.19
111.42	10.65	40.40

Estimate the multiple linear regression equation

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

Practice Questions (Contd.)

12.10 The following data are given:

x	0	1	2	3	4	5	6
y	1	4	5	3	2	3	4

(a) Fit the cubic model $\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$.

(b) Predict Y when $x = 2$.

12.6 An experiment was conducted on a new model of a particular make of automobile to determine the stopping distance at various speeds. The following data were recorded.

Speed, v (km/hr)	35	50	65	80	95	110
Stopping Distance, d (m)	16	26	41	62	88	119

(a) Fit a multiple regression curve of the form $\mu_{D|v} = \beta_0 + \beta_1 v + \beta_2 v^2$.

(b) Estimate the stopping distance when the car is traveling at 70 kilometers per hour.

Practice Questions

12.13 A study was performed on a type of bearing to find the relationship of amount of wear y to x_1 = oil viscosity and x_2 = load. The following data were obtained. (From *Response Surface Methodology*, Myers, Montgomery, and Anderson-Cook, 2009.)

y	x_1	x_2	y	x_1	x_2
193	1.6	851	230	15.5	816
172	22.0	1058	91	43.0	1201
113	33.0	1357	125	40.0	1115

(a) Estimate the unknown parameters of the multiple linear regression equation

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

(b) Predict wear when oil viscosity is 20 and load is 1200.