Statistics for Management and Economics

OPRE 6301 (Statistics and Data Analysis)
by
Ayşegül Toptal Bilhan

Chapter 17
(Multiple Regression)

Multiple Linear Regression

- The simple linear regression model was used to analyze how the dependent variable y is related to one independent variable x.
- Multiple regression allows for any number of independent variables.
- Our model is now:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Multiple Linear Regression
(Coefficients)

Let us consider a regression model with two independent variables. That is,

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i}$$

- The coefficient estimators are:

$$b_1 = \frac{s_y\left(r_{x_1y} - r_{x_1x_2} r_{x_2y}\right)}{s_{x_1}\left(1 - r_{x_1x_2}^2\right)},\ b_2 = \frac{s_y\left(r_{x_2y} - r_{x_1x_2} r_{x_1y}\right)}{s_{x_2}\left(1 - r_{x_1x_2}^2\right)},\ b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2$$

where

r_{x_1y} is the sample correlation between X_1 and Y
r_{x_2y} is the sample correlation between X_2 and Y
$r_{x_1x_2}$ is the sample correlation between X_1 and X_2
s_{x_1} is the sample standard deviation of X_1
s_{x_2} is the sample standard deviation of X_2
s_y is the sample standard deviation of Y

Multiple Linear Regression (Cont'd)

Example: Programmer Salary Survey

A software firm collected data for a sample of 20 computer programmers. A suggestion was made that regression analysis could be used to determine if salary was related to the years of experience and the score on the firm's programmer aptitude test.

The years of experience, score on the aptitude test, and corresponding annual salary (\$1000s) for a sample of 20 programmers is shown on the next slide.

Multiple Linear Regression (Cont'd)

Example (Cont'd): Programmer Salary Survey

Exper. (Yrs.)	Test Score	Salary ($000s)	Exper. (Yrs.)	Test Score	Salary ($000s)
4	78	24.0	9	88	38.0
7	100	43.0	2	73	26.6
1	86	23.7	10	75	36.2
5	82	34.3	5	81	31.6
8	86	35.8	6	74	29.0
10	84	38.0	8	87	34.0
0	75	22.2	4	79	30.1
1	80	23.1	6	94	33.9
6	83	30.0	3	70	28.2
6	91	33.0	3	89	30.0

Multiple Linear Regression (Cont'd)

Example (Cont'd): Programmer Salary Survey

Suppose we believe that salary (y) is related to the years of experience (x_1) and the score on the programmer aptitude test (x_2) by the following regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where

y = annual salary ($000)
x_1 = years of experience
x_2 = score on programmer aptitude test

Multiple Linear Regression (Cont'd)

Example (Cont'd): Programmer Salary Survey

Excel's Regression Equation Output

	A	B	C	D	E
38					
39		Coeffic.	Std. Err.	t Stat	P-value
40	Intercept	3.17394	6.15607	0.5156	0.61279
41	Experience	1.4039	0.19857	7.0702	1.9E-06
42	Test Score	0.25089	0.07735	3.2433	0.00478
43					

Note: Columns F-I are not shown.

E[SALARY]= 3.174 + 1.404(EXPER) + 0.251(SCORE)

Multiple Linear Regression (Cont'd)

Example (Cont'd): Programmer Salary Survey

Excel's Regression Equation Output

	A	B	C	D	E	F
32						
33	ANOVA					
34		df	SS	MS	F	Significance F
35	Regression	2	500.3285	250.1643	42.76013	2.32774E-07
36	Residual	17	99.45697	5.85041		
37	Total	19	599.7855			
38						

SSR

SST

R^2 = SSR/SST, therefore, R^2 = 500.3285/599.7855 = .83418

Multiple Linear Regression
(Estimation of Error Variance)

Let s_ε^2 be an unbiased estimate of error variance σ_ε^2.

$$s_\varepsilon^2 = \frac{SSE}{n-k-1}$$

If number of independent variables is large relative to sample size n, we use Adjusted R^2.

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

Multiple Linear Regression
(Coefficient Standard Errors)

Square roots of the below variance estimators (s_{b_1}, s_{b_2}, …) are called coefficient standard errors.

$$s_{b_1}^2 = \frac{s_\varepsilon^2}{(n-1)s_{x_1}^2\left(1-r_{x_1x_2}^2\right)},\quad s_{b_2}^2 = \frac{s_\varepsilon^2}{(n-1)s_{x_2}^2\left(1-r_{x_1x_2}^2\right)}$$

Multiple Linear Regression
(F-test for Overall Significance)

To test the validity of the regression model, we have

$$H_0:\ \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$$

$$H_a: At\ least\ one\ \beta_i\ is\ not\ equal\ to\ 0.$$

Test Statistic: $F = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)}$

Rejection Region: $F > F_{\alpha,k,n-k-1}$

Multiple Linear Regression
(Testing the Validity of the Model)

SSE	s_ε	R^2	F	Assessment of Model
0	0	1	∞	Perfect
small	small	close to 1	large	Good
large	large	close to 0	small	Poor
$\sum(y_i - \bar{y})^2$	$\sqrt{\frac{\sum(y_i-\bar{y})^2}{n-k-1}}$	0	0	Invalid

Once we're satisfied that the model fits the data as well as possible, and that the required conditions are satisfied, we can interpret and test the individual coefficients and use the model to predict and estimate…

Multiple Linear Regression
(Tests of Hypotheses for Regression Coefficients)

To test the null hypothesis

$$H_0\text{: } \beta_j = \beta^*$$

against the two-sided alternative

$$H_1\text{: } \beta_j \neq \beta^*$$

the decision rule is as follows:

Reject H_0 if $\frac{b_j - \beta^*}{s_{b_j}} > t_{n-k-1,\alpha/2}$ or $\frac{b_j - \beta^*}{s_{b_j}} < -t_{n-k-1,\alpha/2}$