Lecture & Examples

Topic 1: Parameter Estimations in Multiple Regression Models

Although most practical applications of the regression analysis utilize models that are more complex than the simplest straight-line model discussed from Chapter 11, the analyzing steps for a multiple regression model is similar to the analyzing steps for the simple straight-line model.

Suppose that we have one response variable, y, and k independent variables, $x_1, x_2, \ldots, x_k$. Then we can use similar steps discussed in Lecture 5 to analyze the data.

Step 1: Decide the deterministic portion of the multiple regression model. In multiple regression models, the deterministic portion of the multiple regression models is $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$.

Step 2: Make assumptions about the distribution of the random error term of the model. We usually assume the random error term of the model has the following properties:

- $E(\varepsilon) = 0$: the mean of the random error term is zero,
- $\text{Var}(\varepsilon) = \sigma^2$: the variance of the probability distribution of, ε, is σ^2,
- ε's are independent,
- have a normal distribution

Step 3: Estimate the unknown parameters, $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$, and σ^2, with least-squares method.

Step 4: Check the usefulness of the model.

Step 5: When the model fits data adequately, we can use the model for prediction and estimation.

In this lecture, we will discuss the least-squares method to estimate the unknown parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ and the variance of the error term, σ^2. The least-squares estimators, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k$, for the regression coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ can be obtained with computer packages such as SAS.

Let $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$. The least-squares estimator for σ^2 based on the assumptions stated in Step 3 is $s^2 = \text{MSE} = \dfrac{\text{SSE}}{n-(k+1)} = \dfrac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n-(k+1)}$. We can then test the individual regression parameter and construct the $(1-\alpha)*100\%$ confidence interval of these individual regression parameters with the following procedures:

- **Test of an Individual Regression Coefficient in the Multiple Regression Model:**

Two-Tailed Test:

Hypothesis:

H_0: $\beta_i = 0$

H_a: $\beta_i \neq 0$

Test Statistic:

$$t_c = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}}$$

Rejection Region:

$$t_c < -t_{\alpha/2,\, n-(k+1)} \text{ or } t_c > t_{\alpha/2,\, n-(k+1)}$$

Right-Tailed Test:

Hypothesis:

H_0: $\beta_i = 0$

H_a: $\beta_i > 0$

Test Statistic: $t_c = \dfrac{\hat{\beta}_i}{s_{\hat{\beta}_i}}$

Rejection Region: $t_c > t_{\alpha,\, n-(k+1)}$

Left-Tailed Test:

Hypothesis:

H_0: $\beta_i = 0$

H_a: $\beta_i < 0$

Test Statistic: $t_c = \dfrac{\hat{\beta}_i}{s_{\hat{\beta}_i}}$

Rejection Region: $t_c < -t_{\alpha,\, n-(k+1)}$

- **(1−α)*100% Confidence Interval for an Individual Regression Coefficient in the Multiple Regression Model:**

(1-α)*100% confidence interval for β_i

$$\hat{\beta}_i \pm t_{\alpha/2,\, n-(k+1)} s_{\hat{\beta}_i}.$$

- **Example 12.1:**

Regression analysis was employed to investigate the determinants of survival size of nonprofit hospitals (*Applied Economics,* Vol. 18, 1986). For a given sample of hospitals, survival size, y, is defined as the largest size hospital (in terms of number of beds) exhibiting growth in market share over a specific time interval. Suppose 10 states are randomly selected and the survival size for all non-profit hospitals in each state is determined for two time periods five years apart, yielding two observations per state. The 20 survival sizes are listed in the following table, along with the data for each state, for the second year in each time interval:

x_1 = Percentage of beds that are in for-profit hospitals
x_2 = Ratio of the number of persons enrolled in health maintenance organizations (HMOs) to the number of persons covered by hospital insurance
x_3 = State population (in thousands)
x_4 = Percent of state that is urban

Fit a multiple regression model for data in Table 12.1 with the following SAS Printout.

Table 12.1

STATE	TIME	Y	X1	X2	X3	X4
1	1	370	0.13	0.090	5800	89
1	2	390	0.15	0.090	5955	87
2	1	455	0.08	0.110	17648	87
2	2	450	0.10	0.160	17895	85
3	1	500	0.03	0.040	7332	79
3	2	480	0.07	0.050	7610	78
4	1	550	0.06	0.005	11731	80
4	2	600	0.10	0.005	11790	81
5	1	205	0.30	0.120	2932	44
5	2	230	0.25	0.130	3100	45
6	1	425	0.04	0.010	4148	36
6	2	445	0.07	0.020	4205	38
7	1	245	0.20	0.010	1574	25
7	2	200	0.30	0.010	1560	28
8	1	250	0.07	0.080	2471	38
8	2	275	0.08	0.100	2511	38
9	1	300	0.09	0.120	4060	52
9	2	290	0.12	0.200	4175	54
10	1	280	0.10	0.020	2902	37
10	2	270	0.11	0.050	2925	38

SAS Printout for Example 12.1

Model: MODEL1
Dependent Variable: Y
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	4	246538.27382	61634.56845	28.181	0.0001
Error	15	32806.72618	2187.11508		
C Total	19	279345.00000			

Root MSE	46.76660	R-square	0.8826
Dep Mean	360.50000	Adj R-sq	0.8512
C.V.	12.97271		

Parameter Estimates

Variable	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	295.326893	40.17805033	7.350	0.0001
X1	-480.782607	150.39171920	-3.197	0.0060
X2	-829.425309	196.46881427	-4.222	0.0007
X3	0.007936	0.00355388	2.233	0.0412
X4	2.360346	0.76161020	3.099	0.0073

(a) What are the sample estimates for β_0, β_1, β_2, …, β_4 ?

Solution:

$$\hat{\beta}_0 = 295.33,$$

$$\hat{\beta}_1 = -480.78,$$

$$\hat{\beta}_2 = -829.43,$$

$$\hat{\beta}_3 = 0.0079,$$

$$\hat{\beta}_4 = 2.36.$$

(b) What is the least squares prediction equation?

Solution:

$$\hat{y} = 295.33 - 480.78x_1 - 829.43x_2 + 0.0079x_3 + 2.36x_4$$

(c) Find SSE, MSE and s^2.

Solution:

SSE = 32806.73

MSE = 2187.11

s = Root MSE = 46.77

(d) Test H_0: $\beta_1 = 0$ against H_a: $\beta_1 \neq 0$. Use $\alpha = 0.10$.

Solution:

$$t_c = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{-480.78}{150.39} = -3.197$$

Reject the null hypothesis because
the p-value = 0.0060 < α .

(e) Test H_0: $\beta_3 = 0$ against H_a: $\beta_3 > 0$. Use $\alpha = 0.01$.

Solution:

$$t_c = \frac{\hat{\beta}_3}{s_{\hat{\beta}_3}} = \frac{0.0079}{0.0035} = 2.233$$

Rejection region is $t_c > t_{0.01,15} = 2.602$.
Thus, we fail to reject the null hypothesis at $\alpha = 0.01$

(f) Find a 95% confidence interval for β_2.

Solution:

95% confidence interval for β_2 is:

$$\hat{\beta}_2 \pm t_{0.05/2,\, 20-(4+1)} s_{\hat{\beta}_2} = \hat{\beta}_2 \pm t_{0.025,\, 15} s_{\hat{\beta}_2} = -829.42 \pm (2.131)(196.47)$$

$$= [-1248.10,\ -410.74].$$

Example 12.2:

Analyze the data in Table 12.2 with multiple regression model and the SAS Printout.

Table 12.2

Y	X1	X2	X3	X4	X5
90300	4	82	4635	0	4266
384000	20	13	17798	0	14391
157500	5	66	5913	0	6615
676200	26	64	7750	6	34144
165000	5	55	5150	0	6120
300000	10	65	12506	0	14552
108750	4	82	7160	0	3040
276538	11	23	5120	0	7881
420000	20	18	11745	20	12600
950000	62	71	21000	3	39448
560000	26	74	11221	0	30000
268000	13	56	7818	13	8088
290000	9	76	4900	0	11315
173200	6	21	5424	6	4461
323650	11	24	11834	8	9000
162500	5	19	5246	5	3828
353500	20	62	11223	2	13680
134400	4	70	5834	0	4680
187000	8	19	9075	0	7392
155700	4	57	5280	0	6030
93600	4	82	6864	0	3840
110000	4	50	4510	0	3092
573200	14	10	11192	0	23704
79300	4	82	7425	0	3876
272000	5	82	7500	0	9542

SAS Printout for Example 12.2

Model: MODEL1
Dependent Variable: Y
Analysis of Variance

Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Prob>F**
Model	5	1.0528947E12	210578940102	190.749	0.0001
Error	19	20975246806	1103960358.2		
C Total	24	1.0738699E12			

Root MSE	33225.89891	R-square	0.9805
Dep Mean	290573.52000	Adj R-sq	0.9753
C.V.	11.43459		

Parameter Estimates

Variable	**Parameter Estimate**	**Standard Error**	**T for H0: Parameter=0**	**Prob > |T|**
INTERCEP	93074	28720.896862	3.241	0.0043
X1	4152.207009	1491.6258701	2.784	0.0118
X2	-854.941615	298.44765134	-2.865	0.0099
X3	0.924244	2.87673442	0.321	0.7515
X4	2692.461752	1577.2862258	1.707	0.1041
X5	15.542769	1.46287006	10.625	0.0001

(a) Report the least squares prediction equation.

Solution:

$$\hat{y} = 93074 + 4152.21x_1 - 854.94x_2 + 0.92x_3 + 2692.46x_4 + 15.54x_5$$

(b) Find the standard deviation of the regression model.

Solution:

Root MSE = 33225.90

(c) Does the data provide sufficient evidence to conclude that value increases with the number of units in an apartment building? Use $\alpha = 0.05$.

Solution:

Test H_0: $\beta_1 = 0$ against H_a: $\beta_1 > 0$.

$$t_c = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{4152.21}{1491.63} = 2.784$$

Rejection region is $t_c > t_{0.05,\,19} = 1.729$. Thus, we reject the null hypothesis, i.e., the data provide sufficient evidence to conclude that value increases with the number of units in an apartment building

(d) Test H_0: $\beta_2 = 0$ against H_a: $\beta_2 < 0$ using $\alpha = 0.01$. Why is it reasonable to conduct a one-tailed test rather than a two-tailed test of this hypothesis? What is the observed significance level for this test.

Solution:

$$t_c = \frac{\hat{\beta}_2}{s_{\hat{\beta}_2}} = \frac{-854.94}{298.45} = -2.865.$$

Rejection region is $t_c < -t_{0.01,\,19} = -2.539$. Thus, we reject the null hypothesis. Since the sample estimate is -854.94, it is reasonable to conduct a one-tailed (left-tailed) test to test this hypothesis. The observed significance level is 0.0099/2=0.00495.