MATH 3424 Regression Analysis

Chapter 3 : Multiple Linear Regression Model

3.5 An experiment was conducted in order to study the size of squid eaten by sharks and tuna. The regressor variables are characteristics of the beak or mouth of the squid. The regressor variables and response considered for the study are

x_1 : Rostral length in inches
x_2 : Wing length in inches
x_3 : Rostral to notch length
x_4 : Notch to wing length
x_5 : Width in inches
y : Weight in pounds

The study involved measurements and weight taken on 22 specimen. The data are shown in Table 3.2.

Table 3.2 Squid weight and beak measurements

x_1	x_2	x_3	x_4	x_5	y
1.31	1.07	0.44	0.75	0.35	1.95
1.55	1.49	0.53	0.90	0.47	2.90
0.99	0.84	0.34	0.57	0.32	0.72
0.99	0.83	0.34	0.54	0.27	0.81
1.05	0.90	0.36	0.64	0.30	1.09
1.09	0.93	0.42	0.61	0.31	1.22
1.08	0.90	0.40	0.51	0.31	1.02
1.27	1.08	0.44	0.77	0.34	1.93
0.99	0.85	0.36	0.56	0.29	0.64
1.34	1.13	0.45	0.77	0.37	2.08
1.30	1.10	0.45	0.76	0.38	1.98
1.33	1.10	0.48	0.77	0.38	1.90
1.86	1.47	0.60	1.01	0.65	8.56
1.58	1.34	0.52	0.95	0.50	4.49
1.97	1.59	0.67	1.20	0.59	8.49
1.80	1.56	0.66	1.02	0.59	6.17
1.75	1.58	0.63	1.09	0.59	7.54
1.72	1.43	0.64	1.02	0.63	6.36
1.68	1.57	0.72	0.96	0.68	7.63
1.75	1.59	0.68	1.08	0.62	7.78
2.19	1.86	0.75	1.24	0.72	10.15
1.73	1.67	0.64	1.14	0.55	6.88

(a) Fit the multiple regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i$$

(b) Write down 95% confidence intervals on the mean response and 95% prediction intervals on a new observation x_0 at the conditions of the 22 specimen.

(c) Compute a new multiple regression using regressors x_2, x_4, and x_5. Compute S^2, the standard errors of prediction, and 95% confidence intervals on the mean response at the regressor locations for the 22 specimen.

(d) Use the information from (a), (b), and (c) to make a choice between the full model and the reduced model in part (c).

(e) For the squid data, test

$$H_0 : \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = 0 \qquad H_1 : \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \neq 0$$

Draw the conclusions. Comment on the results.

3.6 The principal objective of many data collection exercises in forestry is developing models to use in predicting volume and general value of trees in a forested tract. The data in Table 3.6 give values of characteristics of a particular stand of trees, including: AGE, the age of a particular pine stand; HD, the average height of dominant trees in feet; N, the number of pine trees per acre at age, AGE; and MDBH, the average diameter at breast height (measured at 4.5 feet above ground) at age, AGE.

The data were used to build a model to predict MDBH. Theory suggests that a reasonable definition of regressor variables are x_1 = HD, x_2 = AGE · N, x_3 = HD/N with the response variable y = MDBH. Thus the following model was postulated.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i \qquad (i = 1, 2, \ldots, 20)$$

(a) Fit a regression line.

(b) Fit a linear regression with the term $\beta_3 x_3$ eliminated.

(c) Compute values of S^2; compute the standard error of prediction at the 20 data locations.

(d) Does the comparison between the results in part (a) and those in part (b) signify a superiority of the reduced model or not? Explain.

Table 3.6 Stand characteristics for pine trees

AGE	HD	N	MDBH
19	51.5	500	7.0
14	41.3	900	5.0
11	36.7	650	6.2
13	32.2	480	5.2
13	39.0	520	6.2
12	29.8	610	5.2
18	51.2	700	6.2
14	46.8	760	6.4
20	61.8	930	6.4
17	55.8	690	6.4
13	37.3	800	5.4
21	54.2	650	6.4
11	32.5	530	5.4
19	56.3	680	6.7
17	52.8	620	6.7
15	47.0	900	5.9
16	53.0	620	6.9
16	50.3	730	6.9
14	50.5	680	6.9
22	57.7	480	7.9

(e) For the model in (b), compute the standard error of prediction at the combinations:

x_1	10	80	75
x_2	2,500	6,000	25,000

Does this reveal anything regarding the relative merit of the full (x_1, x_2, x_3) and reduced (x_1, x_2) models for prediction? Explain.

3.8 In a project to study age and growth characteristics of selected mussel species from Southwestern Virginia, the data below were taken from two distinct locations. Fit a regression with weight as the response, age as the independent variable, and location as a categorical variable. Does it appear that location is significant as a categorical variable?

Compute a pure error mean square and perform an F-test for lack of fit. Draw conclusions regarding the adequacy of the simple linear regression.

The categorical variable model assumes, of course, that the slopes of the regression lines are equal. Test the hypothesis that the slopes of the regression lines are equal for the two locations. Draw appropriate conclusions. Is one regression slope an appropriate mode, or is there evidence that one needs to separate regression lines?

Location 1

Age	Weight (g)	Age	Weight (g)
3	0.44	11	3.96
3	0.50	11	3.84
3	0.66	12	5.58
3	0.78	12	5.64
4	1.20	12	4.26
4	1.18	13	6.00
4	1.08	13	2.54
6	1.12	13	3.82
6	1.72	14	4.50
7	1.04	14	5.18
7	1.66	14	4.78
7	1.70	14	5.34
8	2.62	14	4.04
9	1.88	15	6.38
10	2.26	15	4.08
11	4.10	16	4.56
11	4.56	22	6.44
11	2.12		

Location 2

Age	Weight (g)	Age	Weight (g)
3	0.76	8	2.52
4	1.38	8	3.90
5	1.20	10	3.94
5	1.76	10	6.22
6	2.60	10	4.96
6	2.16	13	9.02
6	2.64	13	8.20
6	2.52	13	8.26
6	3.08	14	6.40
6	2.12	15	10.06
7	2.72	15	8.60
7	2.96	18	11.06
8	4.54	19	10.78
8	5.26	22	12.04
8	5.60	24	13.92

3.10 An experiment was designed to study hydrogen embrittlement properties based on electrolytic hydrogen pressure measurements. The solution used was 0.1N NaOH with the material being a certain type of stainless steel. The cathodic charging current density was controlled and varied at four levels. The effective hydrogen pressure was observed as the response. The data follow:

Run	Charging Current Density (x) (ma/cm^2)	Effective Hydrogen Pressure (y) (atm)
1	0.5	86.1
2	0.5	92.1
3	0.5	64.7
4	0.5	74.7
5	1.5	223.6
6	1.5	202.1
7	1.5	132.9
8	2.5	413.5
9	2.5	231.5
10	2.5	466.7
11	2.5	365.3
12	3.5	493.7
13	3.5	382.3
14	3.5	447.2
15	3.5	563.8

(a) Run a simple linear regression of y against x.

(b) Compute the *pure error* sum of squares and make a test for lack of fit.

(c) Does the information in part (b) indicate a need for a model in x beyond a *first order regression*? Explain.

3.11 In an effort to model executive compensation for the year 1979, 33 firms were selected, and data were gathered on compensation, sales, profits, and employment. The data in the following table were gathered for the year 1979.

Firm	Compensation, y (thousands of dollars)	Sales, x_1 (millions of dollars)	Profits, x_2 (millions of dollars)	Employment, x_3
1	450	4600.6	128.1	48000
2	387	9255.4	783.9	55900
3	368	1526.2	136.0	13783
4	277	1683.2	179.0	27765
5	676	2752.8	231.5	34000
6	454	2205.8	329.5	26500
7	507	2384.6	381.8	30800
8	496	2746.0	237.9	41000
9	487	1434.0	222.3	25900
10	383	470.6	63.7	8600
11	311	1508.0	149.5	21075
12	271	464.4	30.0	6874
13	524	9329.3	577.3	39000
14	498	2377.5	250.7	34300
15	343	1174.3	82.6	19405
16	354	409.3	61.5	3586
17	324	724.7	90.8	3905
18	225	578.9	63.3	4139
19	254	966.8	42.8	6255
20	208	591.0	48.5	10605
21	518	4933.1	310.6	65392
22	406	7613.2	491.6	89400
23	332	3457.4	228.0	55200
24	340	545.3	54.6	7800
25	698	22862.8	3011.3	337119
26	306	2361.0	203.0	52000
27	613	2614.1	201.0	50500
28	302	1013.2	121.3	18625
29	540	4560.3	194.6	97937
30	293	855.7	63.4	12300
31	528	4211.6	352.1	71800
32	456	5440.4	655.2	87700
33	417	1229.9	97.5	14600

Consider the model

$$y_i = \beta_0 + \beta_1 \ln x_{1i} + \beta_2 \ln x_{2i} + \beta_3 \ln x_{3i} + \varepsilon_i \qquad (i = 1, 2, \ldots, 33)$$

(a) Fit the regression with the above model.

(b) Compute the correlation matrix among the regressor variables.

(c) Compute the eigenvalues of the correlation matrix.

(d) Compute the variance inflation factors for the coefficients in the above model.

(e) Make an assessment of the extent of the multicollinearity in this problem.

3.12 The HAT matrix is an idempotent matrix that plays an important role in linear regression analysis. Idempotency of $\mathbf{H}$ implies that $\mathbf{H}^2 = \mathbf{H}$. Show that $\mathbf{I} - \mathbf{H}$ is also idempotent.

3.14 Consider the situation of a simple linear regression with a single categorical variable with two levels; that is, the model is that given in Exercise 2.23.

(a) Write out the model in $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ form, where each x_i is centered with the mean $\bar{x}_j$ for the jth group in question, as given in Exercise 2.23. The vector $\boldsymbol{\beta}$ is a three parameter vector.

(b) Use $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$; write out an estimator for the common slope, and show its equivalence to that given in Exercise 2.23. Is this estimator of common slope intuitively reasonable?

3.15 Consider the data of Exercise 2.19. The chambers in which the experiments were conducted were held under different conditions of temperature and humidity. There is also some evidence that the model should be curvilinear in head diameter.

(a) Write an appropriate model including linear and quadratic terms in head diameter and an indicator variable to account for the nine chambers.

(b) Fit the model created in part (a) and make appropriate conclusions.

3.16 In an experiment conducted in the civil engineering department at Virginia Polytechnic Institute and State University in 1988, a growth of a certain type of algae in water was observed as a function of time and the dosage of copper added to the water. The following data were collected.

y (units of algae)	$\mathbf{x}_1$ (mg copper)	$\mathbf{x}_2$ (days)	**y** (units of algae)	$\mathbf{x}_1$ (mg copper)	$\mathbf{x}_2$ (days)
.3	1	5	.23	1	18
.34	1	5	.23	1	18
.2	2	5	.28	2	18
.24	2	5	.27	2	18
.24	2	5	.25	2	18
.28	3	5	.27	3	18
.2	3	5	.25	3	18
.24	3	5	.25	3	18
.02	4	5	.06	4	18
.02	4	5	.10	4	18
.06	4	5	.10	4	18
0	5	5	.02	5	18
0	5	5	.02	5	18
0	5	5	.02	5	18
.37	1	12	.36	1	25
.36	1	12	.36	1	25
.30	2	12	.24	2	25
.31	2	12	.27	2	25
.30	2	12	.31	2	25
.30	3	12	.26	3	25
.30	3	12	.26	3	25
.30	3	12	.28	3	25
.14	4	12	.14	4	25
.14	4	12	.11	4	25
.14	4	12	.11	4	25
.14	5	12	.04	5	25
.15	5	12	.07	5	25
.15	5	12	.05	5	25

(a) Consider the following model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2} + \varepsilon_i$$

Estimate the coefficients of the model, using multiple linear regression.

(b) Test

$$H_0 : \beta_{12} = 0$$
$$H_1 : \beta_{12} \neq 0$$

Do you have any reason to change the model from that given in part (a) ?

(c) Show a partitioning of total degrees of freedom into those attributed to regression, pure error, and lack of fit.

(d) Using the model you adopted in part (b), make a test for lack of fit and draw conclusions.

(e) Plot residuals of fitted model against x_1 and x_2, separately, and comment.

3.17 Chemical pruning of fruit trees is a very important activity. A study was done by the horticulture department of Virginia Polytechnic Institute and State University to determine the impact of Terbacil. Three levels of Terbacil, 50 PPM, 75 PPM, and 100 PPM, were used in an experimental apple orchard. The following data set gives the average number of healthy apples per cm^2 of cross sectional area of limb. In another portion of the same orchard, a type of surfactant was added to the Terbacil. This should theoretically reduce the number of heathy apples. The data are as follows:

	Number of Healthy Apples Per cm^2
Control (0 PPM)	7.67
50 PPM	6.26
75 PPM	5.73
100 PPM	5.44
50+ surfactant	4.05
75+ surfactant	3.02
100+ surfactant	2.34

Fit an appropriate linear regression model to the data. Use the surfactant as an indicator variable.

3.18 Consider the punting data in Exercise 2.11. Fit three separate models

$$E(y_i) = \beta_0 + \beta_1 x_{1i} \qquad \text{(model 1)}$$

$$E(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \qquad \text{(model 2)}$$

$$E(y_i) = \beta_0 + \beta_2 x_{2i} \qquad \text{(model 3)}$$

Let x_{1i} be right-leg strength; x_{2i} is left-leg strength.

(a) Plot the residuals for these three models. Plot against $\hat{y}$. Comment on which model seems to be most appropriate.

(b) Test

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0$$

Make appropriate conclusion.

3.19 A scientist collects experimental data on the radius of a propellant grain (y) as a function of powder temperature, x_1, extrusion rate, x_2, and die temperature, x_3.

Grain Radius	Powder Temperature (x_1)	Extrusion Rate (x_2)	Die Temperature (x_3)
82	150	12	220
93	190	12	220
114	150	24	220
124	150	12	250
111	190	24	220
129	190	12	250
157	150	24	250
164	190	24	250

(a) Consider the linear regression model with centered regressors

$$y_i = \beta_0^* + \beta_1(x_{1i} - \bar{x}_1) + \beta_2(x_{2i} - \bar{x}_2) + \beta_3(x_{3i} - \bar{x}_3) + \varepsilon_i$$

Write the vector **y**, the matrix **X**, and the vector **β** in the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

(b) Write out the least squares normal equations

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}$$

Comment on what is special about the $\mathbf{X}'\mathbf{X}$ matrix. What characteristic in this experiment do you suppose produces this special form of $\mathbf{X}'\mathbf{X}$?

(c) Estimate the coefficients in the multiple linear regression model.

(d) Test the hypotheses

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

and make conclusions.

(e) Compute the $100(1 - \alpha)\%$ confidence intervals on $E(y \mid x)$ at each of the locations of x_1, x_2, and x_3 described by the data points. Note the relative widths of these confidence intervals.

3.20 The following is a data set generated by the Environmental Protection Agency and analyzed by the Statistical Consulting Centre at Virginia Polytechnic Institute and State University. The amount of magnesium uptake is measured at several levels of time. It is anticipated that the two treatments used may results in different regression equations. The data are as follows:

Magnesium (y)	Time (x)	Treatment
2.08	0	1
2.07	0	1
2.09	0	1
2.06	0	1
2.17	1	1
2.15	1	1
2.16	1	1
2.15	1	1
2.08	2	1
2.02	2	1
2.06	2	1
2.04	2	1
1.97	3	1
1.96	3	1
1.98	3	1
1.96	3	1
1.89	4	1
1.87	4	1
1.88	4	1
1.89	4	1
1.81	5	1
1.72	5	1
1.81	5	1
1.75	5	1
1.66	6	1
1.64	6	1
1.61	6	1
1.59	6	1

Magnesium (y)	Time (x)	Treatment
2.08	0	2
2.09	0	2
2.07	0	2
2.04	0	2
2.03	1	2
2.04	1	2
2.01	1	2
2.08	1	2
1.88	2	2
1.86	2	2
1.96	2	2
1.91	2	2
1.77	3	2
1.60	3	2
1.78	3	2
1.80	3	2
1.57	4	2
1.47	4	2
1.59	4	2
1.56	4	2
1.52	5	2
1.43	5	2
1.48	5	2
1.52	5	2
1.53	6	2
1.40	6	2
1.98	6	2
1.45	6	2

(a) A model is postulated in which calcium uptake is regressed against time in a quadratic regression. In other words

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 z$$

where z is an indicator variable accounting for the treatment. Fit this regression model.

(b) We need to determine if the simple indicator variable is actually appropriate. Suppose we rewrite the model

$$E(y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 \qquad \text{(treatment 1)}$$
$$= \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2 \qquad \text{(treatment 2)}$$

Set up the general linear hypothesis for testing

$$H_0 : \beta_1 = \gamma_1$$
$$H_0 : \beta_2 = \gamma_2$$

(If this hypothesis is rejected, two separate regressions are necessary.) That is, determine **C**, $\boldsymbol{\beta}$, and **d** for $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$.

(c) Test the above hypothesis. The unrestricted model contains six parameters and the restricted model contains four parameters.

3.21 Consider the following data set:

y	$\mathbf{x}_1$	$\mathbf{x}_2$
3.9	1.5	2.2
7.5	2.7	4.5
4.4	1.8	2.8
8.7	3.9	4.4
9.6	5.5	4.3
19.5	10.7	8.4
29.3	14.6	14.6
12.2	4.9	8.5

(a) Fit a multiple linear regression model.

(b) Use the general linear hypothesis to test

$$H_0 : \beta_1 = \beta_2 = 1$$

and make appropriate conclusions. Use full and restricted model residual sums of squares.

3.22 Consider the following data set. The response involves percent yield.

		Humidity x_2		
		10%	20%	30%
Temperature, x_1	100°	74.5	72.7	71.5
	150°	81.7	80.4	79.5
	200°	95.1	94.2	90.7

(a) Fit a multiple linear regression with the model

$$E(y_i) = \beta_0 + \beta_1 x_1 + \beta_{11} x_1^2 + \beta_2 x_2 + \beta_{22} x_2^2$$

Also compute R^2, $R(\beta_1, \beta_{11} \mid \beta_0, \beta_2, \beta_{22})$, and test $H_0 : \beta_1 = \beta_{11} = 0$.

(b) Rewrite the model using temperature and humidity as indicator variables. Set up the model in $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ form. Fit the model to the set of data.

(c) For the model in part (b) compute R^2 and test the hypothesis that both coefficients of the temperature indicator variables are zero (jointly). Use sums of squares for the full and reduced model.

(d) Are you surprised at the comparison between the results of parts (b) and (c) ?

3.23 Consider the general linear regression model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ and the least squares estimate $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Show that

$$\mathbf{b} = \beta + \mathrm{R}\varepsilon \qquad \text{where } \mathbf{R} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$