ASSIGNMENT 2

Due date: Friday, 12 November 2021.

The data files are available in Moodle

Submit your homework to Moodle.

1. Work out the following formulae for the simple linear regression in Lecture Note 3 and 4 from the corresponding formulae given in matrix terms:

   (a) the normal equations,

   (b) $\widehat{\beta}_0$,

   (c) $\widehat{\beta}_1$,

   (d) $\mathbf{var}(\widehat{\beta}_0)$,

   (e) $\mathbf{var}(\widehat{\beta}_1)$.

2. Use the method of least squares to fit a straight line to the five data points:

x	-2	-1	0	1	2
y	4	3	3	1	-1

   (a) Construct Y and X matrices for the data.

   (b) Find $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}^T\mathbf{Y}$.

   (c) Find the least squares estimates $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.

   (d) Give the prediction equation.

3. Do the data given in Exercise 2 provide sufficient evidence to indicate that x contributes information for the prediction of y? Test $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$ using $\alpha = .05$.

4. Refer to Exercise 2. Find a 90% confidence interval for E(y) when $x = 1$. Interpret the interval. Suppose you plan to observe y for x = 1. Find a 90% prediction interval for that value of y. Interpret the interval.

5. An experiment was conducted to investigate the effect of extrusion pressure P and temperature T on the strength y of a new type of plastic. Two plastic specimens were prepared for each of five combina- tions of pressure and temperature. The specimens were then tested in random order, and the break- ing strength for each specimen was recorded. The independent variables were coded to simplify com- putations, that is,

$$x_1 = \frac{P - 200}{10}, x_2 = \frac{T - 400}{25}$$

   The $n = 10$ data points are listed in the table

y	x_1	x_2
5.2; 5.0	-2	2
.3; -.1	-1	-1
-1.2; -1.1	0	-2
2.2; 2.0	1	-1
6.2; 6.1	2	2

(a) Give the **Y** and **X** matrices needed to fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$.

(b) Find the least squares prediction equation.

(c) Find SSE and s^2.

(d) Does the model contribute information for the prediction of y? Test using $\alpha = .05$.

(e) Find R^2 and interpret its value.

(f) Test the null hypothesis that $\beta_1 = 0$. Use $\alpha = .05$. What is the practical implication of the test?

(g) Find a 90% confidence interval for the mean strength of the plastic for $x_1 = -2$ and $x_2 = 2$.

(h) Suppose a single specimen of the plastic is to be installed in the engine mount of a Douglas DC-10 aircraft. Find a 90% prediction interval for the strength of this specimen if $x_1 = -2$ and $x_2 = 2$.

6. **Deep space survey of quasars.** A quasar is a distant celestial object (at least 4 billion light-years away) that provides a powerful source of radio energy. The Astronomical Journal (July 1995) reported on a study of 90 quasars detected by a deep space survey. The survey enabled astronomers to measure several different quantitative characteristics of each quasar, including redshift range, line flux (erg/$cm^2\cdot$ s), line luminosity (erg/s), AB1450 magnitude, absolute magnitude, and rest frame equivalent width. The data for a sample of 25 large (redshift) quasars is listed in the below table

(a) Hypothesize a first-order model for equivalent width, y, as a function of the first four variables in the table.

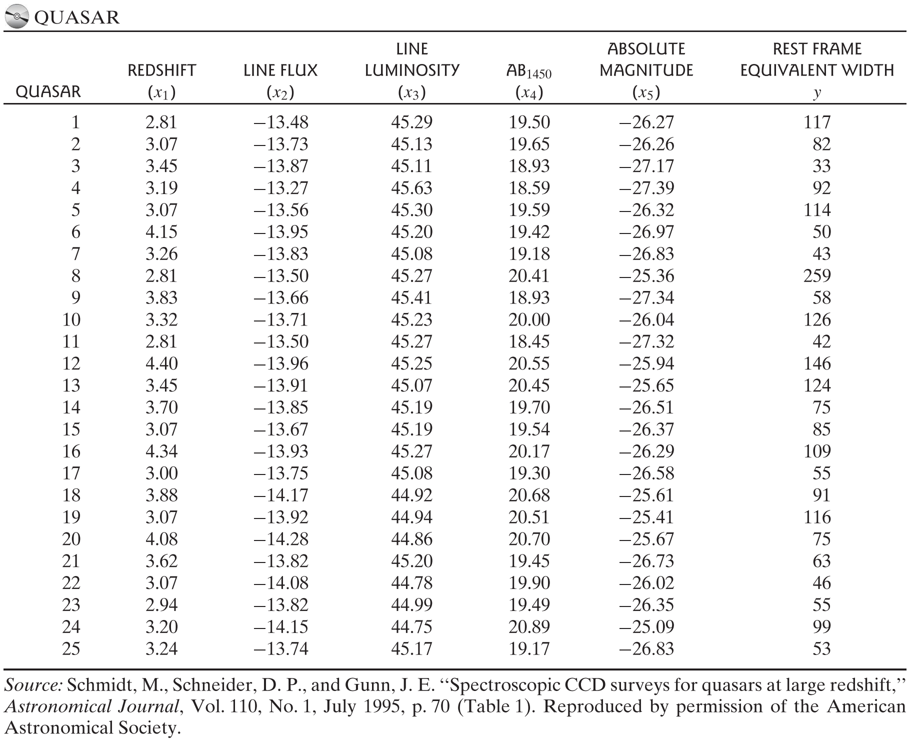
(b) Fit the first-order model for the data. Give the least squares prediction equation.

(c) Interpret the $\boldsymbol{\beta}$ estimates in the model.

(d) Test to determine whether redshift (x1) is a useful linear predictor of equivalent width (y), using $\alpha = .05$.

(e) Locate R^2 and R_a^2 . Interpret these values. Which statistic is the preferred measure of model fit? Explain.

(f) Locate the global F -value for testing the overall model on the SPSS printout. Use the statistic to test the null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_4 = 0$.



QUASAR

QUASAR	REDSHIFT (x_1)	LINE FLUX (x_2)	LINE LUMINOSITY (x_3)	AB_{1450} (x_4)	ABSOLUTE MAGNITUDE (x_5)	REST FRAME EQUIVALENT WIDTH y
1	2.81	−13.48	45.29	19.50	−26.27	117
2	3.07	−13.73	45.13	19.65	−26.26	82
3	3.45	−13.87	45.11	18.93	−27.17	33
4	3.19	−13.27	45.63	18.59	−27.39	92
5	3.07	−13.56	45.30	19.59	−26.32	114
6	4.15	−13.95	45.20	19.42	−26.97	50
7	3.26	−13.83	45.08	19.18	−26.83	43
8	2.81	−13.50	45.27	20.41	−25.36	259
9	3.83	−13.66	45.41	18.93	−27.34	58
10	3.32	−13.71	45.23	20.00	−26.04	126
11	2.81	−13.50	45.27	18.45	−27.32	42
12	4.40	−13.96	45.25	20.55	−25.94	146
13	3.45	−13.91	45.07	20.45	−25.65	124
14	3.70	−13.85	45.19	19.70	−26.51	75
15	3.07	−13.67	45.19	19.54	−26.37	85
16	4.34	−13.93	45.27	20.17	−26.29	109
17	3.00	−13.75	45.08	19.30	−26.58	55
18	3.88	−14.17	44.92	20.68	−25.61	91
19	3.07	−13.92	44.94	20.51	−25.41	116
20	4.08	−14.28	44.86	20.70	−25.67	75
21	3.62	−13.82	45.20	19.45	−26.73	63
22	3.07	−14.08	44.78	19.90	−26.02	46
23	2.94	−13.82	44.99	19.49	−26.35	55
24	3.20	−14.15	44.75	20.89	−25.09	99
25	3.24	−13.74	45.17	19.17	−26.83	53

Source: Schmidt, M., Schneider, D. P., and Gunn, J. E. "Spectroscopic CCD surveys for quasars at large redshift," *Astronomical Journal*, Vol. 110, No. 1, July 1995, p. 70 (Table 1). Reproduced by permission of the American Astronomical Society.

7. **Cooling method for gas turbines.** Refer to the Journal of Engineering for Gas Turbines and Power (January 2005) study of a high pressure inlet fogging method for a gas turbine engine. Recall that the heat rate (kilojoules per kilowatt per hour) was measured for each in a sample of 67 gas turbines augmented with high pressure inlet fogging. In addition, several other variables were measured, including cycle speed (revolutions per minute), inlet temperature , exhaust gas temperature , cycle pressure ratio, and air mass flow rate (kilograms per second). The data are saved in the GASTURBINE file. (The first and last five observations are listed in the below table.)

   (a) Write a first-order model for heat rate (y) as a function of speed, inlet temperature, exhaust temperature, cycle pressure ratio, and air flow rate.

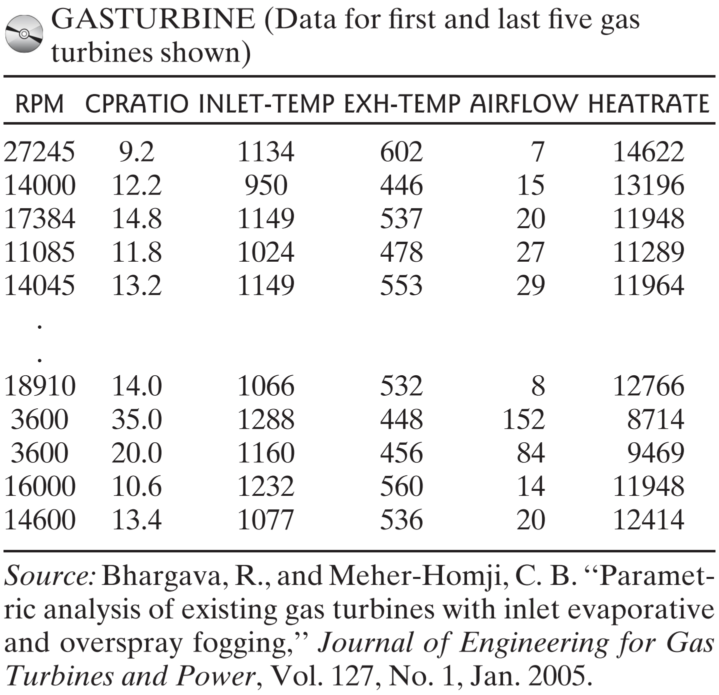
   (b) Fit the model to the data using the method of least squares.

   (c) Give practical interpretations of the β estimates.

   (d) Find the model standard deviation, s, and interpret its value.

   (e) Find the adjusted-R2 value and interpret it.

   (f) Is the overall model statistically useful at predicting heat rate (y)? Test using $\alpha = .01$.



GASTURBINE (Data for first and last five gas turbines shown)

RPM	CPRATIO	INLET-TEMP	EXH-TEMP	AIRFLOW	HEATRATE
27245	9.2	1134	602	7	14622
14000	12.2	950	446	15	13196
17384	14.8	1149	537	20	11948
11085	11.8	1024	478	27	11289
14045	13.2	1149	553	29	11964
.					
.					
18910	14.0	1066	532	8	12766
3600	35.0	1288	448	152	8714
3600	20.0	1160	456	84	9469
16000	10.6	1232	560	14	11948
14600	13.4	1077	536	20	12414

Source: Bhargava, R., and Meher-Homji, C. B. ''Parametric analysis of existing gas turbines with inlet evaporative and overspray fogging,'' *Journal of Engineering for Gas Turbines and Power*, Vol. 127, No. 1, Jan. 2005.

8. **Role of retailer interest on shopping behavior.** Retail interest is defined by marketers as the level of interest a consumer has in a given retail store. Marketing professors at the University of Tennessee at Chattanooga and the University of Alabama investigated the role of retailer interest in consumers' shopping behavior (Journal of Retailing, Summer 2006). Using survey data collected for $n = 375$ consumers, the professors developed an interaction model for $y =$ willingness of the consumer to shop at a retailer's store in the future (called "repatronage intentions") as a function of $x_1 =$ consumer satisfaction and $x2 =$ retailer interest. The regression results are shown below.

VARIABLE	ESTIMATED β	t-VALUE	p-VALUE
Satisfaction (x_1)	.426	7.33	<.01
Retailer interest (x_2) .044	0.85	$>$ .10	
Satisfaction × Retailer interest (x_1x_2)	-.157	-3.09	$<$ .01
$R_2 = .65, F = 226.35, p -$ value $< .001$			

(a) Is the overall model statistically useful for predicting y? Test using $\alpha = .05$.

(b) Conduct a test for interaction at $\alpha = .05$.

(c) Use the β-estimates to sketch the estimated relationship between repatronage intentions (y) and satisfaction (x_1) when retailer interest is $x_2 = 1$ (a low value).

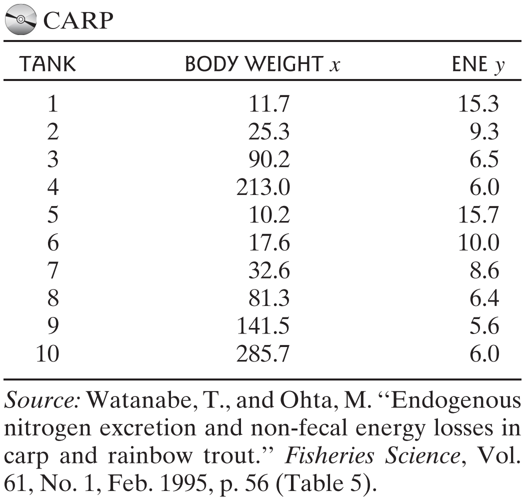
(d) Repeat part c when retailer interest is $x_2 = 7$ (a high value).

(e) Sketch the two lines, parts c and d, on the same graph to illustrate the nature of the interaction.

9. **Carp diet study.** *Fisheries Science* (February 1995) reported on a study of the variables that affect endogenous nitrogen excretion (ENE) in carp raised in Japan. Carp were divided into groups of 2-15 fish, each according to body weight and each group placed in a separate tank. The carp were then fed a protein- free diet three times daily for a period of 20 days. One day after terminating the feeding experiment, the amount of ENE in each tank was measured. The below table gives the mean body weight (in grams) and ENE amount (in mil ligrams per 100 grams of body weight per day for each carp group.

   (a) Graph the data in a scatterplot. Do you detect a pattern?

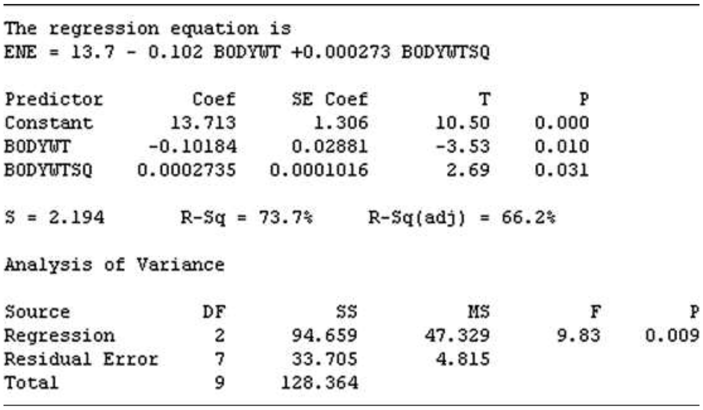
   (b) The quadratic model $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$ was fit to the data using MINITAB The MINITAB printout is displayed below Conduct the test $H_0 0 : \beta - 2 = 0$ against $H_a : \beta_2 \neq 0$ using $\alpha = .10$. Give the conclusion in the words of the problem.



CARP

TANK	BODY WEIGHT x	ENE y
1	11.7	15.3
2	25.3	9.3
3	90.2	6.5
4	213.0	6.0
5	10.2	15.7
6	17.6	10.0
7	32.6	8.6
8	81.3	6.4
9	141.5	5.6
10	285.7	6.0

Source: Watanabe, T., and Ohta, M. "Endogenous nitrogen excretion and non-fecal energy losses in carp and rainbow trout." *Fisheries Science*, Vol. 61, No. 1, Feb. 1995, p. 56 (Table 5).



```
The regression equation is
ENE = 13.7 - 0.102 BODYWT +0.000273 BODYWTSQ

Predictor        Coef     SE Coef        T      P
Constant       13.713       1.306    10.50  0.000
BODYWT       -0.10184     0.02881    -3.53  0.010
BODYWTSQ    0.0002735   0.0001016     2.69  0.031

S = 2.194        R-Sq = 73.7%     R-Sq(adj) = 66.2%

Analysis of Variance

Source            DF        SS        MS       F      P
Regression         2    94.659    47.329    9.83  0.009
Residual Error     7    33.705     4.815
Total              9   128.364
```

10. **RNA analysis of wheat genes.** Engineers from the Department of Crop and Soil Sciences at Washington State University used regression to estimate the number of copies of a gene transcript in an aliquot of RNA extracted from a wheat plant (Electronic Journal of Biotechnology, April 15, 2004). The proportion (x_1) of RNA extracted from a cold-exposed wheat plant was varied, and the transcript copy number (y, in thousands) was measured for each of two cloned genes: Mn Superoxide Dismutose (MnSOD) and Phospholipose D (PLD). The data are listed in the accompanying table. Letting x_1 = RNA proportion and $x_2 = \{1$ if MnSOD, 0 if PLD$\}$, consider the second-order model for number of copies (y):

$$\mathrm{E}(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_1 x2 + \beta_5 x_1^2 x_2$$

(a) MINITAB was used to fit the model to the data. Locate the least squares prediction equation for y on the printout shown below.

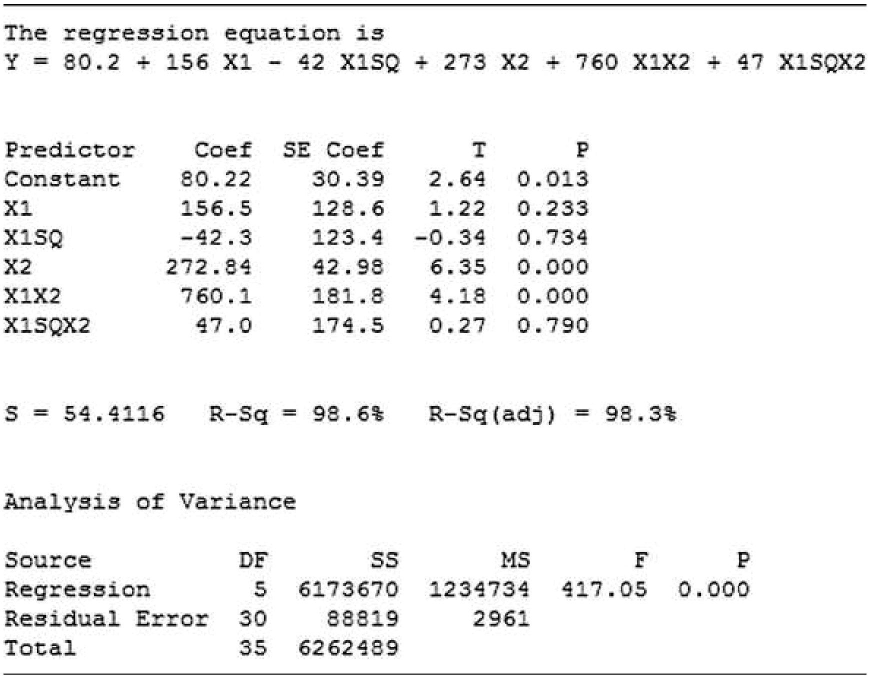
(b) Is the overall model statistically useful for predicting transcript copy number (y)? Test using $\alpha = .01$.

(c) Based on the MINITAB results, is there evidence to indicate that transcript copy number (y) is curvilinearly related to proportion of RNA $(x1)$? Explain.

WHEATRNA

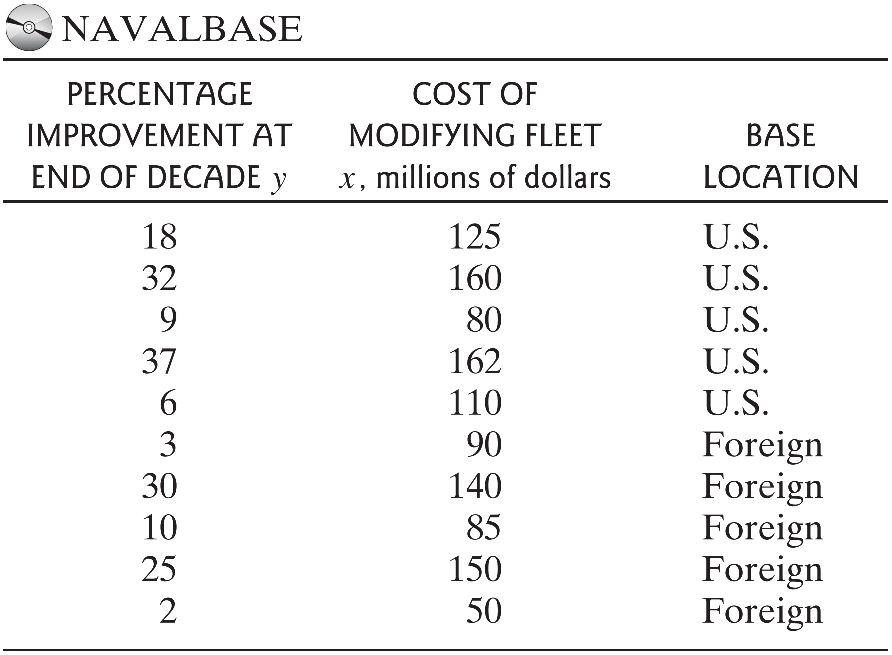
RNA PROPORTION	NUMBER OF COPIES (Y, THOUSANDS)	
(x_1)	*MNSOD*	*PLD*
0.00	401	80
0.00	336	83
0.00	337	75
0.33	711	132
0.33	637	148
0.33	602	115
0.50	985	147
0.50	650	142
0.50	747	146
0.67	904	146
0.67	1007	150
0.67	1047	184
0.80	1151	173
0.80	1098	201
0.80	1061	181
1.00	1261	193
1.00	1272	187
1.00	1256	199

Source: Baek, K. H., and Skinner, D. Z. "Quantitative real-time PCR method to detect changes in specific transcript and total RNA amounts," *Electronic Journal of Biotechnology*, Vol. 7, No. 1, April 15, 2004 (adapted from Figure 2).



```
The regression equation is
Y = 80.2 + 156 X1 - 42 X1SQ + 273 X2 + 760 X1X2 + 47 X1SQX2


Predictor     Coef  SE Coef      T      P
Constant     80.22    30.39   2.64  0.013
X1           156.5    128.6   1.22  0.233
X1SQ         -42.3    123.4  -0.34  0.734
X2          272.84    42.98   6.35  0.000
X1X2         760.1    181.8   4.18  0.000
X1SQX2        47.0    174.5   0.27  0.790


S = 54.4116   R-Sq = 98.6%   R-Sq(adj) = 98.3%


Analysis of Variance

Source          DF       SS       MS       F      P
Regression       5  6173670  1234734  417.05  0.000
Residual Error  30    88819     2961
Total           35  6262489
```

11. **Cost of modifying a naval air base.** A naval base is considering modifying or adding to its fleet of 48 standard aircraft. The final decision regarding the type and number of aircraft to be added depends on a comparison of cost versus effectiveness of the modified fleet. Consequently, the naval base would like to model the projected percentage increase y in fleet effectiveness by the end of the decade as a function of the cost x of modifying the fleet. A first proposal is the quadratic model $\mathrm{E}(y) = \beta_0 + \beta_1 x + \beta_2 x^2$ The data provided in the table were collected on 10 naval bases of similar size that recently expanded their fleets.



NAVALBASE

PERCENTAGE IMPROVEMENT AT END OF DECADE y	COST OF MODIFYING FLEET x, millions of dollars	BASE LOCATION
18	125	U.S.
32	160	U.S.
9	80	U.S.
37	162	U.S.
6	110	U.S.
3	90	Foreign
30	140	Foreign
10	85	Foreign
25	150	Foreign
2	50	Foreign

(a) Fit the quadratic model to the data.

(b) Interpret the value of R_a^2 on the printout.

(c) Find the value of s and interpret it.

(d) Perform a test of overall model adequacy. Use $\alpha = .05$.

(e) Is there sufficient evidence to conclude that the percentage improvement y increases more quickly for more costly fleet modifications than for less costly fleet modifications? Test with $\alpha = .05$.

(f) Now consider the complete second-order model

$$\mathrm{E}(y) = \beta_0 + \beta_1 x - 11 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_1 x_2 + \beta_5 x_1^2 x_2$$

where x_1 = Cost of modifying the fleet. $x_2 = 1$ if U.S. base, else $x_2 = 0$ if foreign base. Fit the complete model to the data. Is there sufficient evidence to indicate that type of base (U.S. or foreign) is a useful predictor of per- centage improvement y? Test using $\alpha = .05$.