/75

STAT 2509 C
Assignment #3
SOLUTION

1. A study was conducted involving the relationship between the selling price (in thousands of dollars) of a house (y) and two independent variables, the number of rooms (x_1) and the number of square feet (x_2). The following data were collected on 22 properties sold in a particular residential area;

House	Selling price (y)	Rooms (x_1)	Sq. Ft (x_2)
1	25.75	5	986
2	37.95	5	998
3	46.45	7	1690
4	46.55	8	1829
5	47.95	6	1186
6	49.95	6	1734
7	52.45	7	1684
8	54.05	7	1846
9	54.85	7	1690
10	52.05	7	1910
11	54.39	7	1784
12	53.45	6	1690
13	59.51	7	1590
14	60.10	8	1855
15	63.85	8	2212
16	62.05	10	2784
17	69.45	7	2190
18	82.30	8	2259
19	81.85	7	1919
20	70.05	7	1685
21	112.45	10	2654
22	127.05	10	2756

Consider the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$.

[5] (a) State the MLR model and all assumptions which are necessary for the statistical inference.

Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, **$n = 22$**

Assumptions: (i) **X_1, X_2 are observed without error**
(ii) **y's (or ε's) are independently distributed with mean** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ **(or** $E(\varepsilon) = 0$**)**
(iii) **variance of y's (or ε's) is constant, σ^2 for all X_1, X_2**
(iv) **$y \sim N(E(y), \sigma^2)$ for any value of X_1, X_2 (or $\varepsilon \sim N(0, \sigma^2)$ for any value of X_1, X_2)**

[5] (b) Use matrices to compute the estimates of the population parameters β_0, β_1, β_2 and hence obtain the fitted least squares prediction line.

$$(X^TX)^{-1}X^TY = \begin{bmatrix} 1.68378472 & -0.399634527 & 0.000681592 \\ -0.399634527 & 0.171636536 & -0.000456131 \\ 0.000681592 & -0.000456131 & 0.00000142 \end{bmatrix} * \begin{bmatrix} 1364.5 \\ 10430.2 \\ 2719760 \end{bmatrix} =$$

$$= \begin{bmatrix} -16.9759 \\ 4.336021 \\ 0.025511 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \hat{\beta}$$

∴ **the least squares fitted regression line is given by:** $\hat{Y} = X\hat{\beta}$, **i.e.**

$\hat{y}$ = **– 16.9759 + 4.336021** x_1 **+ 0.025511** x_2

[20] (c) Set up the ANOVA table and hence test for the significance of the model. Use $\alpha = 0.05$.

$$TSS = Y^TY - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 95497.48 - \frac{(1364.5)^2}{22} =$$

= 95 497.48 – 84 630.01 = 10 867.47

$$SSR = \hat{\beta}^T(X^TY) - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = [-16.9759 \quad 4.336021 \quad 0.025511] * \begin{bmatrix} 1364.5 \\ 10430.2 \\ 2719760 \end{bmatrix} - \frac{(1364.5)^2}{22} =$$

= 91 445.75 – 84 630.01 = 6 815.739

$$SSE = TSS - SSR = \mathbf{4\ 051.73}$$

$$MSR = \frac{SSR}{k} = \frac{6815.739}{2} = \mathbf{3\ 407.869}$$

$$MSE = \frac{SSE}{n-(k+1)} = \frac{4051.73}{19} = \mathbf{213.2489}$$

$$F = \frac{MSR}{MSE} = \mathbf{15.98071}$$

Source	*d.f.*	*SS*	*MS*	*F*
Regression	2	6 815.739	3 407.869	15.98071
Error	19	4 051.73	213.2489	
Total	21	10 867.47		

$H_0 : \beta_1 = \beta_2 = 0$ $\quad \alpha = 0.05$

H_a : *at least one of the* $\beta' s \neq 0$

test-statistics: $F = \dfrac{MSR}{MSE} =$ **15.98071**

R.R: **we reject** H_0 **if** $F > F_{\alpha(k,n-(k+1))} = F_{0.05(2,19)} =$ **3.52**

Since *F* = 15.98 > 3.52, we reject H_0 and conclude that at 5% level of significance there is an evidence to say that a linear relationship between the selling price of a house, The number of rooms and the number of square feet exists.

[7] (d) Test whether x_1 term (i.e. whether the number of rooms) contributes to the given model. Use t-test with α = 0.05.

$H_0 : \beta_1 = 0$ $\quad \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$

$H_a : \beta_1 \neq 0$

test-statistics: $t = \dfrac{\hat{\beta}_1}{\sqrt{v_{11} MSE}} = \dfrac{4.336021}{\sqrt{(0.171636536)(213.2489)}} =$ **0.716709**

v_{11} is the 2nd diagonal element of $(X'X)^{-1}$

R.R: **we reject** H_0 **if** $t < -t_{\alpha/2;n-(k+1)} = -t_{0.025;19} =$ **− 2.093**

or $t > t_{\alpha/2;n-(k+1)} = t_{0.025;19} =$ **2.093**

Since *t* = 0.717 ≯ 2.093, we do not reject H_0 and conclude that at 5% level of significance that there is not enough evidence to say that the X_1 term (i.e. the number of rooms) contributes to the model.

[6] (e) Find the values of the coefficient of determination, r^2, and the adjusted r^2 and interpret their meanings in this problem.

$r^2 = \dfrac{SSR}{TSS} =$ **0.627169 ≅ 62.72%**

i.e. approximately 62.72% of the total variation in the data is explained by the regression line (and 37.28% is due to error).

$$r^2_{adj} = 1 - \frac{SSE/n-(k+1)}{TSS/n-1} = 1 - \frac{MSE}{TSS/n-1} = 1 - \frac{213.2489}{10\,867.47/21} = 1 - 0.412076 = \mathbf{0.587924}$$

$\cong$ **58.79%**

Since both r^2 and r^2_{adj} are quite low (i.e. around 60%) and since the X_1 term does not contribute to the model, we can conclude that the model is not good.

(f) Run SAS to verify your above results and also use the SAS output to answer part (d) using partial F-test with $\alpha = 0.05$.

$H_0 : \beta_1 = 0$ $\quad \alpha = 0.05$

$H_a : \beta_1 \neq 0$

- **full model:** $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$
- **reduced model:** $y = \beta_0 + \beta_2 x_2 + \varepsilon$

SSR_f = 6 816.77693 (d.f. = 2) **SSR_r = 6 707.23479 (d.f. = 1)**

SSE_f = 4 050.68890 (d.f. = 19) **SSE_r = 4 160.23105 (d.f. = 20)**

test-statistics :

$$F_{part} = \frac{[SSR_f - SSR_r]/[df_{SSR_f} - df_{SSR_r}]}{SSE_f/df_{SSE_f}} = \frac{(6816.77693 - 6707.23479)/(2-1)}{4050.68890/19} =$$

$$= \frac{109.5421/1}{213.1942} = \mathbf{0.513814}$$

or equivalently,

$$F_{drop} = \frac{[SSE_r - SSE_f]/[df_{SSE_r} - df_{SSE_f}]}{SSE_f/df_{SSE_f}} = \frac{(4160.23105 - 4050.68890)/(20-19)}{4050.68890/19} =$$

$$= \frac{109.5422/1}{213.1942} = \mathbf{0.513814}$$

R.R: we reject H_0 if $F_{part} > F_{\alpha(1,19)} = F_{0.05(1,19)} = 4.38$

Since F_{part} = 0.514 $\ngtr$ 4.38, we do not reject H_0 and conclude that at 5% level of significance there is not enough evidence to say that the X_1 term (i.e. the number of rooms) contributes to the model.

See SAS output on pages 5 & 6 for results.

The REG Procedure
Model: MODEL1
Dependent Variable: Price

Number of Observations Read	22
Number of Observations Used	22

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	6816.77693	3408.38847	15.99	<.0001
Error	19	4050.68890	213.19415		
Corrected Total	21	10867			

SSR, MSR, SSE, MSE, TSS, F-test

Root MSE	14.60117	R-Square	0.6273
Dependent Mean	62.02273	Adj R-Sq	0.5880
Coeff Var	23.54164		

r^2, r^2_{adj}

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-16.97598	18.94658	-0.90	0.3815
Rooms	1	4.33606	6.04912	0.72	0.4822
SqFt	1	0.02551	0.01738	1.47	0.1585

$\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, t-test

- **Full model:** $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

3

Name, student #

The REG Procedure
Model: MODEL2
Dependent Variable: Price

Number of Observations Read	22
Number of Observations Used	22

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6707.23479	6707.23479	32.24	<.0001
Error	20	4160.23105	208.01155		
Corrected Total	21	10867			

Root MSE	14.42261	R-Square	0.6172
Dependent Mean	62.02273	Adj R-Sq	0.5980
Coeff Var	23.25374		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-6.87999	12.51767	-0.55	0.5887
SqFt	1	0.03703	0.00652	5.68	<.0001

- **Reduced model:** $y = \beta_0 + \beta_2 x_2 + \varepsilon$

3

Name, student #

```
Footnote 'Name, student #';
Data HousePrice;
Input Price Rooms SqFt @@;
Cards;
   25.75 5 986 37.95 5 998 46.45 7 1690 46.55 8 1829 47.95 6 1186
   49.95 6 1734 52.45 7 1684 54.05 7 1846 54.85 7 1690 52.05 7 1910
   54.39 7 1784 53.45 6 1690 59.51 7 1590 60.10 8 1855 63.85 8 2212
   62.05 10 2784 69.45 7 2190 82.30 8 2259 81.85 7 1919 70.05 7 1685
   112.45 10 2654 127.05 10 2756
Run;
Proc Reg;
      Model Price=Rooms SqFt;
      Model Price=SqFt;
Run;
```

Q.1 (2)

```
Footnote 'Name, student #';
Data Drug;
Input dose X2 X3 potency;
   X1=log(dose);
   interact12=X1*X2;
   interact13=X1*X3;
Cards;
    0.2 0 0 2.0
    0.4 0 0 4.3
    0.8 0 0 6.5
    1.6 0 0 8.9
    0.2 1 0 1.8
    0.4 1 0 4.1
    0.8 1 0 4.9
    1.6 1 0 5.7
    0.2 0 1 1.3
    0.4 0 1 2.0
    0.8 0 1 2.8
    1.6 0 1 3.4
Run;
Proc Reg;
   Model potency=X1 X2 X3 interact12 interact13;
   Model potency=X1 X2 X3;
Run;
```

Q.2 (2)

[8]

2. Consider the following model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \varepsilon$$

where $x_2 = \begin{cases} 1, & \text{if drug } B \\ 0, & \text{otherwise} \end{cases}$ $x_3 = \begin{cases} 1, & \text{if drug } C \\ 0, & \text{otherwise} \end{cases}$

$x_1 = \ln(dose)$ y = potency of drug

Run **SAS** to test whether the 3 lines are parallel, i.e. test whether the slopes of these 3 lines are the same. Use $\alpha = 0.05$.

- **full model:** $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \varepsilon$ (1)

if drug A: $y = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3(0) + \beta_4 x_1(0) + \beta_5 x_1(0) + \varepsilon$

or $y = \beta_0 + \beta_1 x_1 + \varepsilon$

if drug B: $y = \beta_0 + \beta_1 x_1 + \beta_2(1) + \beta_3(0) + \beta_4 x_1(1) + \beta_5 x_1(0) + \varepsilon$

or $y = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) x_1 + \varepsilon$

if drug C: $y = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3(1) + \beta_4 x_1(0) + \beta_5 x_1(1) + \varepsilon$

or $y = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) x_1 + \varepsilon$

to test whether the 3 drug lines are parallel is the same as to test whether their slopes are the same, i.e. whether β_4 and β_5 = 0

$H_0 : \beta_4 = \beta_5 = 0$ $\alpha = 0.05$

H_a : *at least one of the* β's $\neq 0$ (1)

- **reduced model:** $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ (1)

SSR_f = 55.29350 (d.f. = 5)
SSE_f = 0.68900 (d.f. = 6)

SSR_r = 48.84417 (d.f. = 3)
SSE_r = 7.13833 (d.f. = 8)

test-statistics :

$$F_{drop} = \frac{[SSE_r - SSE_f]/[df_{SSE_r} - df_{SSE_f}]}{SSE_f / df_{SSE_f}} = \frac{(7.13833 - 0.68900)/(8-6)}{0.68900/6} =$$

$$= \frac{3.224665}{0.114833} = \underline{\underline{28.08126}}$$

or equivalently,

$$F_{part} = \frac{[SSR_f - SSR_r]/[df_{SSR_f} - df_{SSR_r}]}{SSE_f / df_{SSE_f}} = \frac{(55.29350 - 48.84417)/(5-3)}{0.68900/6} =$$

$$= \frac{3.224665}{0.114833} = \underline{\underline{28.08126}}$$

R.R: we reject H_0 if $F_{drop} > F_{\alpha(2,6)} = F_{0.05(2,6)} =$ **5.14**

Since F_{drop} = 28.08126 > 5.14, we reject H_0 and conclude that at 5% level of significance there is an evidence to say that the slopes of the 3 drug lines are not parallel (i.e. they differ).

See SAS output on pages 10 & 11 for results.

The REG Procedure
Model: MODEL1
Dependent Variable: potency

Number of Observations Read	12
Number of Observations Used	12

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	55.29350	11.05870	96.30	<.0001
Error	6	0.68900	0.11483		
Corrected Total	11	55.98250			

Root MSE	0.33887	R-Square	0.9877
Dependent Mean	3.97500	Adj R-Sq	0.9774
Coeff Var	8.52505		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7.30722	0.21029	34.75	<.0001
X1	1	3.30377	0.21864	15.11	<.0001
X2	1	-2.15481	0.29740	-7.25	0.0004
X3	1	-4.34865	0.29740	-14.62	<.0001
interact12	1	-1.50040	0.30920	-4.85	0.0028
interact13	1	-2.27946	0.30920	-7.37	0.0003

- **Full model:** $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \varepsilon$

3

name, st.#

The REG Procedure
Model: MODEL2
Dependent Variable: potency

Number of Observations Read	12
Number of Observations Used	12

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	48.84417	16.28139	18.25	0.0006
Error	8	7.13833	0.89229		
Corrected Total	11	55.98250			

Root MSE	0.94461	R-Square	0.8725
Dependent Mean	3.97500	Adj R-Sq	0.8247
Coeff Var	23.76382		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.58940	0.51309	12.84	<.0001
X1	1	2.04382	0.35187	5.81	0.0004
X2	1	-1.30000	0.66794	-1.95	0.0875
X3	1	-3.05000	0.66794	-4.57	0.0018

- **Reduced model:** $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$

3

name, st.#