Applied Statistics Comprehensive Examination

Regression Methods & Linear Statistical Models

1. (20 pts) Consider an experiment to compare the effectiveness of two different pesticides, A and B, both applied in two different forms: spray (A_1 and B_1) and powder (A_2 and B_2). A control treatment (i.e., no pesticide), C, is also included in the experiment for comparison purposes. Thus, altogether there are 5 treatments (A_1, A_2, B_1, B_2 and C) and each treatment is randomly applied to n uniformly infested plots of land. Obtain a complete set of meaningful orthogonal contrasts and describe what each contrast is testing.

2. (50 pts) The SAS output on the following four pages comes from fitting a model of the form $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ to nine observations of the form (x_1, x_2, y).

    (a) (5 pts) State the hypothesis being tested by the model F statistic in the ANOVA table and interpret the result.

    (b) (5 pts) State the hypotheses being tested by the t statistics associated with the individual parameter estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ and interpret the results.

    (c) (5 pts) Interpret the R-Square and adjusted R-Sq values.

    (d) (5 pts) Determine if the model is useful for prediction and justify your answer.

    (e) (5 pts) Determine if multicollinearity is a problem and justify your answer.

    (f) (5 pts) Identify all observations with high leverage and justify your answer.

    (g) (5 pts) Identify all observations with high influence and justify your answer.

    (h) (5 pts) Identify all observations with large residuals and justify your answer.

    (i) (5 pts) Determine if normality of the residuals is a problem and justify your answer.

    (j) (5 pts) Determine if constant variance is a problem and justify your answer.

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read   9
Number of Observations Used   9

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	593237	296619	2.33	0.1783
Error	6	763736	127289		
Corrected Total	8	1356974			

Root MSE	356.77641	R-Square	0.4372
Dependent Mean	383.77778	Adj R-Sq	0.2496
Coeff Var	92.96432		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-39.11056	327.25257	-0.12	0.9088	0
x1	1	19.41399	9.66288	2.01	0.0913	1.14666
x2	1	0.14590	7.63111	0.02	0.9854	1.14666

The REG Procedure
Model: MODEL1
Dependent Variable: y

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D	RStudent	Hat Diag H
1	100.0000	240.6981	217.5771	-140.6981	282.8	-0.498	|	0.049	-0.4639	0.3719
2	388.0000	335.9863	143.2906	52.0137	326.7	0.159	|	0.002	0.1456	0.1613
3	755.0000	1029	322.2149	-274.3560	153.2	-1.791	***|	4.730	-2.3965	0.8156
4	1288	526.2280	136.3997	761.7720	329.7	2.311	|****	0.305	6.3565	0.1462
5	230.0000	275.7746	212.7894	-45.7746	286.4	-0.160	|	0.005	-0.1462	0.3557
6	0	163.2584	225.0062	-163.2584	276.9	-0.590	*|	0.077	-0.5546	0.3977
7	345.0000	237.6444	226.1216	107.3556	276.0	0.389	|	0.034	0.3597	0.4017
8	0	173.7550	153.6745	-173.7550	322.0	-0.540	*|	0.022	-0.5050	0.1855
9	348.0000	471.2990	144.6200	-123.2990	326.2	-0.378	|	0.009	-0.3493	0.1643

Output Statistics

Obs	Cov Ratio	DFFITS	Intercept	x1	x2
			--------- DFBETAS ---------		
1	2.4244	-0.3570	-0.3348	-0.0007	0.2794
2	2.0343	0.0639	0.0496	0.0033	-0.0343
3	0.9449	-5.0408	2.3021	-4.2627	-0.2906
4	0.0027	2.6299	0.2145	1.2572	-0.1887
5	2.6479	-0.1087	-0.0990	-0.0078	0.0866
6	2.3988	-0.4507	0.0218	0.3087	-0.3215
7	2.6751	0.2947	-0.0584	-0.1680	0.2338
8	1.8275	-0.2410	-0.1623	0.1404	0.0058
9	1.9235	-0.1549	0.0345	-0.0123	-0.0771

Sum of Residuals	0
Sum of Squared Residuals	763736
Predicted Residual SS (PRESS)	3242634

y = -39.111 +19.414 x1 +0.1459 x2

N   9
Rsq   0.4372
AdjRsq   0.2496
RMSE   356.78

Normal Quantile

Studentized Residual

$y = -39.111 + 19.414\ x1 + 0.1459\ x2$

N	9
Rsq	0.4372
AdjRsq	0.2496
RMSE	356.78

Predicted Value

Studentized Residual
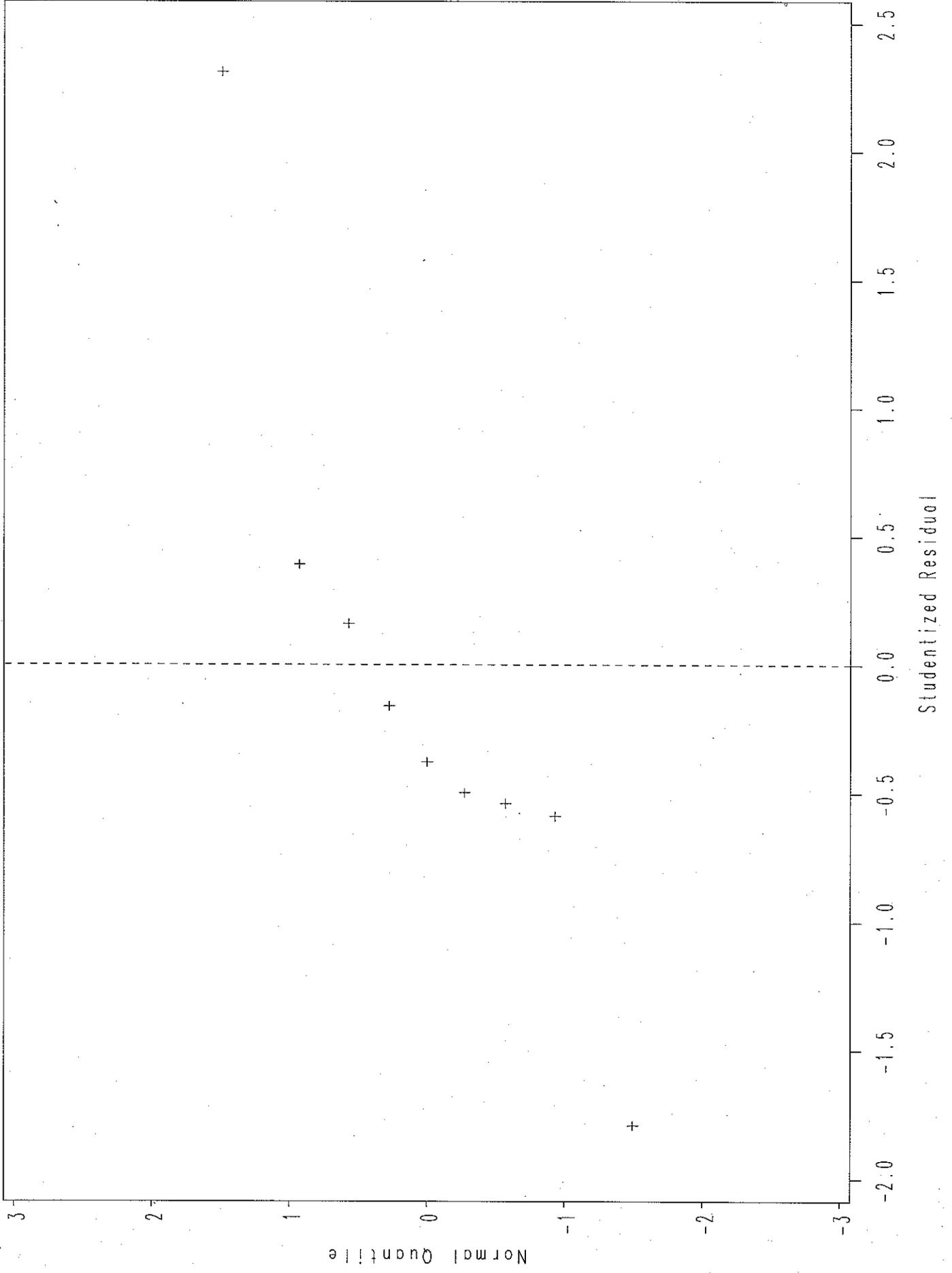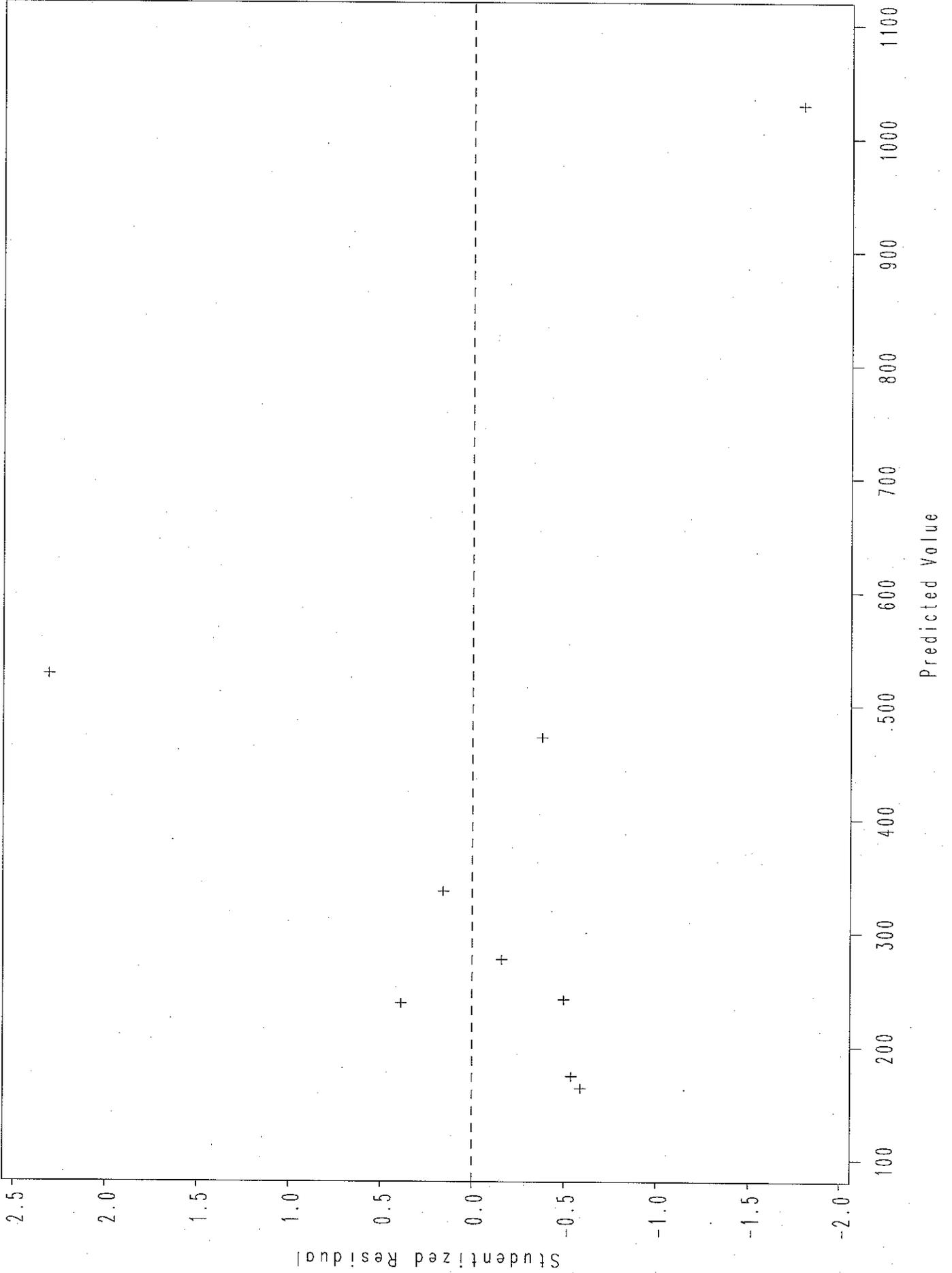
3. (30 pts) An experiment was conducted to compare six different management techniques (such as pruning, spraying and fertilizing) for apple trees with respect to yield. Each apple tree represents an experimental unit and the trial was laid out as a completely randomized design with 5 replications for each management technique. All the trees underwent the same management practice before the trial. For each tree, the yield in bushels (x) for the four year period preceding the trial will be used as a covariate. At the end of the four year experimental period, the yield in pounds of apples (y) is obtained for every tree. Suppose the partial SAS Proc GLM printout is as follows:

Source	Type I SS	Type III SS
Treatments	40	60
Previous yield	26	26
Residual	46	
Total	112	

(a) (10 pts) State the complete mathematical model appropriate for this experiment.

(b) (10 pts) Determine if there are differences among the management techniques after adjusting for previous yield using $\alpha = 0.05$.

(c) (10 pts) Determine if the use of a covariate has been successful in reducing the variance for treatment comparisons using $\alpha = 0.05$.