MONTCLAIR STATE UNIVERSITY
DEPARTMENT OF MATHEMATICAL SCIENCES

MASTER OF SCIENCE
COMPREHENSIVE EXAMINATION

STATISTICS

. . . . . 1 Hour . . . . . .

STUDY GUIDE:

The following questions are representative of the type of content that can be expected in this ONE (1) hour examination. Note that some questions may combine concepts from more than one course.

Instructions:
In this comprehensive examination you must answer a total of THREE (3) questions. Probability Tables, Graphs and SAS / JMP output will be provided as needed.

Methodology: Significance testing, confidence intervals, analysis of variance, contrasts, multiple comparisons, assumptions, blocking designs, model evaluation, residual diagnostics, correlation, simple and multiple linear regression, polynomial regression, sequential and partial sums-of-squares (Type I and III analyses), collinearity, interpretation of output from statistical software such as SAS, JMP and graphing calculators.

Courses: STAT 541, STAT 548

References:

The **candidate's** lecture notes and **textbook** from each of the required courses

Moore and McCabe, *Introduction to the Practice of Statistics*
Ott, *An Introduction to Statistical Methods and Data Analysis*
Kutner, Nachtsheim and Neter, *Applied Linear Regression Models*
Draper and Smith, *Applied Regression Analysis*
Montgomery, *Design and Analysis of Experiments*

Question 1

The following ANOVA table was obtained from a *balanced* completely randomized design:

Source	df	SS	MS	F
Treatment		126		
Error	20		16	
Total	23			

(a) By completing the ANOVA table above determine:

(i) The number of treatments.

(ii) The number of replications per treatment.

(b) Hence state the model and the assumptions needed for inference purposes.

(c) Test whether there is a difference between the true mean responses to the treatment.

(d) Determine the $P-$value of this test and explain what it represents.

Question 2

In a study of fitness, researchers were interested in the effects of weight and oxygen capacity on the time to run 1.5 miles in working adults. Data were collected on 31 individuals. Weight was measured in kg and Runtime in minutes. Higher values of oxygen capacity (Oxy) are associated with better aerobic fitness.

(a) Figure 1 and Table 1 give the results from fitting different Simple Linear Regressions (SLR's). What conclusions are suggested by each SLR ?

(b) Based on your conclusions from Figure 1 and Table 1, find a 95% confidence interval for the mean time for an individual whose oxygen capacity is 51 to run 1.5 miles.

(c) Can the correlation between `Runtime` and `Oxy` be determined from Table 1 ?
If so, compute it ; if not, explain why.

(d) If both weight and oxygen capacity are used as explanatory variables, would you expect your conclusions as to the usefulness of weight and oxygen capacity to explain average running time to be the same as those in part (a) ? Explain.

Figure 1: Plots for Question 2

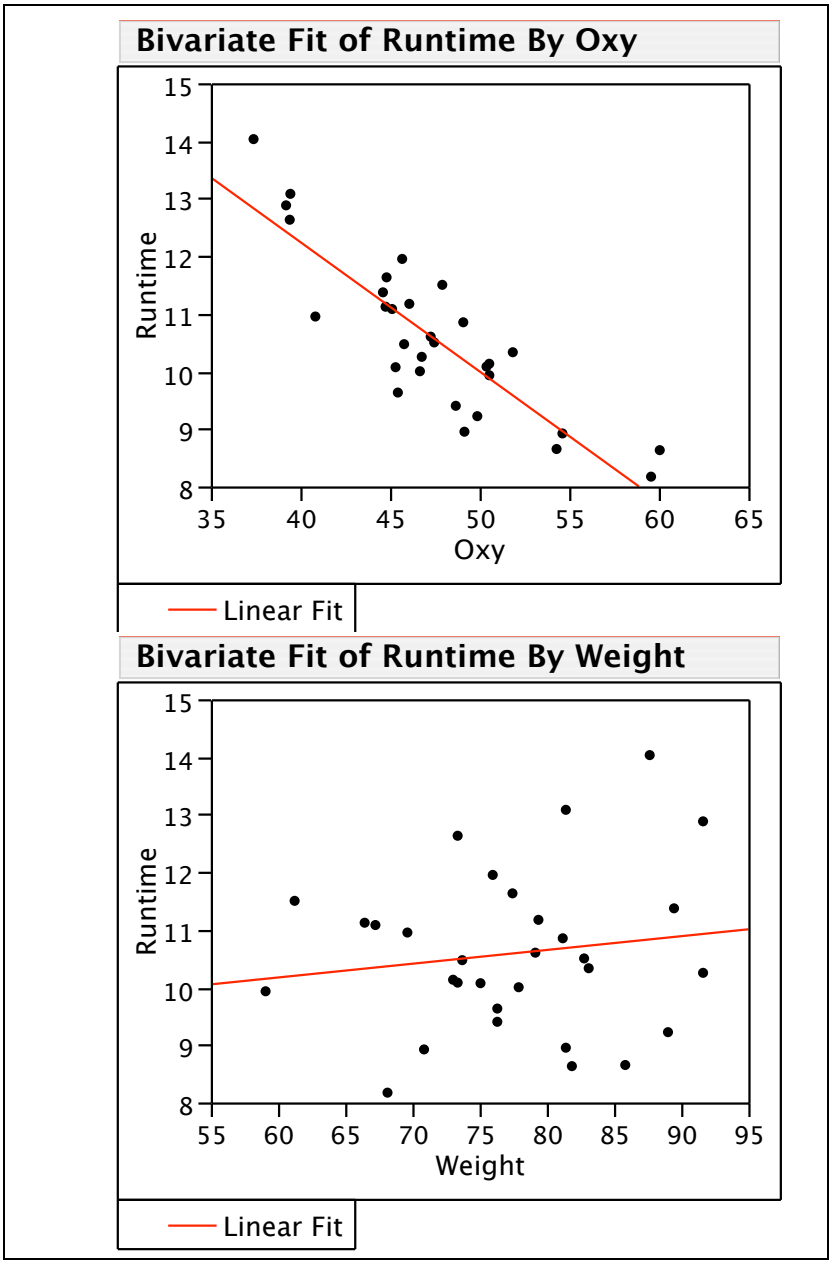

Table 1: SAS Output for Question 2

Dependent Variable: Runtime

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	42.92405	42.92405	83.97	<.0001
Error	29	14.82349	0.51115		
Corrected Total	30	57.74754			

Root MSE	0.71495	R-Square	0.7433
Dependent Mean	10.58613	Adj R-Sq	0.7345
Coeff Var	6.75366		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	21.22219	1.16775	18.17	<.0001
Oxy	1	-0.22450	0.02450	-9.16	<.0001

Dependent Variable: Runtime

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.18928	1.18928	0.61	0.4412
Error	29	56.55826	1.95028		
Corrected Total	30	57.74754			

Root MSE	1.39653	R-Square	0.0206
Dependent Mean	10.58613	Adj R-Sq	-0.0132
Coeff Var	13.19204		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	8.73472	2.38411	3.66	0.0010
Weight	1	0.02391	0.03061	0.78	0.4412

Question 3

A randomized complete block design was employed to study the effects of three diets on the *reduction* of total lipid (fat) level in blood plasma. Data and SAS output from fifteen male subjects who were within 20% of their ideal body weight is given in **Table 2** below.

Table 2: Lipid Readings

Age Group	Fat Content of Diet		
	Extremely Low	Fairly Low	Moderately Low
15 - 24	0.73	0.67	0.15
25 - 34	0.86	0.75	0.21
35 - 44	0.94	0.81	0.26
45 - 54	1.40	1.32	0.75
55 - 64	1.62	1.41	0.78

Source	DF	Sum of Squares	Mean Square
AGE	4	1.41896	0.354740
FAT	2	1.32028	0.660140
Error	8	0.01932	0.002415
Corrected Total	14	2.75856	

Level of AGE	N	Y Mean	Y SD
15-24	3	0.51666667	0.31895663
25-34	3	0.60666667	0.34789845
35-44	3	0.67000000	0.36097091
45-54	3	1.15666667	0.35444793
55-64	3	1.27000000	0.43714986

Level of FAT	N	Y Mean	Y SD
ext_low	5	1.11000000	0.38078866
fair_low	5	0.99200000	0.34557199
mod_low	5	0.43000000	0.30846394

(a) Write the model and assumptions for this experiment.

(b) Use the SAS output to explain why *two* of the assumptions appear to be satisfied.

(c) Use the ANOVA results to state "appropriate" conclusions about the two factors.

(d) Give two orthogonal contrasts that could be used to compare the `FAT` levels.

Question 4

The number of pounds of steam used per month at a plant is thought to be related to the average monthly ambient temperature. The past year's usages and temperature follow:

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Temperature	21	24	32	47	50	59	68	74	62	50	40	30
Usage	19	21	29	42	46	54	62	67	56	45	37	28

(a) Fit a simple linear regression model to the data (with Usage as the response).

(b) Is there a linear association between Temperature and Usage?

(c) Management believes that an increase in average ambient temperature of 1 degree will increase average monthly steam consumption by 1 (thousand) lbs.

Do the data support this statement ?

(d) Construct a 95% prediction interval on steam usage in a month with average ambient temperature of 58 degrees.

Question 5

SAS output from fitting three regression models to the data below is given in **Table 3**.

Y	X_1	X_2
28	3	49
33	4	41
42	7	33
55	8	31
75	16	7
91	21	5

(a) From the SAS output in **Table 3**, write down the three regression models that were fitted to the dataset above.

(b) (i) For each regression, what are your conclusions ?

Be specific: discuss the quality of the regression, the overall significance, and the significance of the explanatory variable(s) in the regression model.

(ii) What "conflict" appears to arise between the three regression results ?

(c) (i) Do a rough sketch of X_1 versus X_2

(ii) Why does this plot help explain the apparent conflict between the results ?

Table 3: SAS Output for Question 5

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3004.41465	3004.41465	177.81	0.0002
Error	4	67.58535	16.89634		
Corrected Total	5	3072.00000			

Root MSE	4.11052	R-Square	0.9780
Dependent Mean	54.00000	Adj R-Sq	0.9725

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	20.23610	3.03764	6.66	0.0026
X1	1	3.43362	0.25749	13.33	0.0002

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2913.36198	2913.36198	73.46	0.0010
Error	4	158.63802	39.65950		
Corrected Total	5	3072.00000			

Root MSE	6.29758	R-Square	0.9484
Dependent Mean	54.00000	Adj R-Sq	0.9355

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	91.17851	5.04245	18.08	<.0001
X2	1	-1.34380	0.15679	-8.57	0.0010

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3009.92646	1504.96323	72.73	0.0029
Error	3	62.07354	20.69118		
Corrected Total	5	3072.00000			

Root MSE	4.54876	R-Square	0.9798
Dependent Mean	54.00000	Adj R-Sq	0.9663

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	33.93210	26.74829	1.27	0.2941
X1	1	2.78476	1.28906	2.16	0.1195
X2	1	-0.26442	0.51232	-0.52	0.6414

Question 6

A study to investigate possible causes for high serum phosphate levels in adults with sickle cell anemia included three groups: five sickle cell patients with high serum phosphate levels ($>$ 4.5 mg/dliter), four sickle cell patients with normal serum phosphate levels (3 - 4.5 mg/dliter), and 7 controls. The response variable was tubular reabsorptive capacity for phosphate (TRCP). [TRCP is a measure of the phosphate reabsorption rate in the kidney.]

(a) In planning your pre-ANOVA analysis determine appropriate orthogonal contrasts that would help experimenters analyze the experiment. State the contrasts in words and give the coefficients for the contrast.

(b) State the ANOVA model and state the necessary assumptions.

(c) What methods would you use to test the assumptions of part b).
[**Note:** Do not perform any tests, just discuss the methods.]

(d) What options would you consider if these assumptions are violated ?

(e) Use the SAS Output given in **Table 4** to analyze the data. Summarize the results and conclusions of your analysis.

Table 4: SAS Output for Question 6

Group	N	Mean	Std Dev	Minimum	Maximum
Control NonS	7	3.3071429	0.5498398	2.6000000	4.0800000
High Serum	5	4.6420000	0.4506884	4.1200000	5.3600000
Normal Serum	4	3.5750000	0.2998333	3.1800000	3.8900000

Dependent Variable: TRCP

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5.44645214	2.72322607	12.22	0.0010
Error	13	2.89612286	0.22277868		
Corrected Total	15	8.34257500			

Table 4 Continued: SAS Output for Question 6

R-Square	C.V.	Root MSE	TRCP Mean
0.652850	12.44957	0.471994	3.791250

Source	DF	Anova SS	Mean Square	F Value	Pr > F
GROUP	2	5.44645214	2.72322607	12.22	0.0010

Analysis of Variance Procedure

Level of GROUP	N	TRCP Mean	TRCP SD
Control NonS	7	3.30714286	0.54983980
High Serum	5	4.64200000	0.45068836
Normal Serum	4	3.57500000	0.29983329

Duncan's Multiple Range Test for variable: TRCP

NOTE: This test controls the type I comparisonwise error rate, not the experimentwise error rate

Alpha= 0.05 df= 13 MSE= 0.222779
WARNING: Cell sizes are not equal.
Harmonic Mean of cell sizes= 5.060241

Number of Means	2	3
Critical Range	.6411	.6714

Means with the same letter are not significantly different.

Duncan Grouping	Mean	N	GROUP
A	4.6420	5	High Serum
B	3.5750	4	Normal Serum
B			
B	3.3071	7	Control NonS