MS Applied Statistics
Comprehensive Examination
August, 1999

Number:

Instructions

1. Please write your number at the top of each page.
2. Please do all your work in the space provided. If you need more space, please note on your paper that you continued the problem on the back of the **previous** page. If you need extra paper or if are missing a table you need, please inform the proctor who will take care of the matter for you.

GOOD LUCK

1. Match the following with the single best answer (a) - (n) below:

[20 points]

(i) In any experimental study, _____ are needed in order to obtain an estimate of the experimental error variance.

(ii) The simplest of all experimental designs; it should be considered when some experimental units are likely to be lost or become missing._____

(iii) The smallest subdivision of experimental material that can receive a treatment independently. _____

(iv) The estimated standard error of the difference between two means where each mean is based on a sample of size n. _____

(v) An experimental design where experimental units are grouped together to remove a source of variation between groups. _____

(a)	Duncan's multiple range test	(h)	randomized complete block design
(b)	simple correlation coefficient	(i)	$\sqrt{\frac{2MSError}{n}}$
(c)	replications	(j)	least significant difference procedure
(d)	experimental unit	(k)	ANOVA F-test
(e)	completely randomized design	(l)	simple linear regression
(f)	$\sqrt{\frac{MSError}{n}}$	(m)	the Tukey LSD value
(g)	test statistic	(n)	t-statistic

2. Given the sample set of six numbers $Y_1 = 6, Y_2 = 9, \cdots, Y_6 = 6$ where $\sum_{i=1}^{6} Y_i = 37$ and $\sum_{i=1}^{6} Y_i^2 = 251$,

   (a) Construct a 95% confidence interval for the mean of the population of Y values assuming the population is Normal.

   [7 points]

   (b) Suppose we know that the population of Y is Normal with mean $\mu = 8$ and variance $\sigma^2 = 25$. If we sample 10 more values in addition to the six in a), find the conditional probability, given the first six observed values, that the sample mean ($\overline{Y}$) of the 16 values will be greater than 9.5

   [6 points]

(c) Define the sum of the last 4 observations to be $\ell = Y_{13} + Y_{14} - Y_{15} - Y_{16}$. Find $Pr(17 \leq \ell \leq 47)$.

[7 points]

3. A manufacturer claims his company produces ball bearings with an exact diameter of 5 mm. To test his claim, we draw a sample of $n = 25$. Our calculations produced $\overline{Y} = 4.6$ mm and $s = 0.75$ mm. Is there sufficient evidence to reject the manufacturer's claim? Answer this question on the basis of an appropriate p-value.

[15 points]

4. It is conjectured that students who attend class lectures and also study the textbook perform better than students who study the textbook but occasionally miss class. On a recent test, 35 students, swearing they never miss class and practice good study habits scored $\overline{Y}_1 = 92.5$ with $S_1^2 = 6.5$. Thirty students who admit to missing class "now-and-then" but study the text, scored $\overline{Y}_2 = 90.2$ with $S_2^2 = 10.2$. Construct an appropriate confidence bound that can be used to check the validity of the conjecture at the 1% significance level. Use the calculated bound to draw a conclusion about the conjecture.

[15 points]

5. A candidate for student body president claims she has at least 75% of the votes in an upcoming election. A random sample of 60 students showed 15 favored her opponent and 45 favored the candidate. Do we believe the candidate?

[15 points]

6. Four brands (A, B, C and D) of hand-held calculators are to be compared for the time it takes to compute the sum of 25 numbers. Thirty-two (32) students were selected at random with 8 assigned to each brand of calculator. A partial ANOVA table is

Source	d.f.	Sum of Squares	M.S.
Among Calculator Brands			
Within Brands		15.40	

(a) Write a model for this experiment defining all terms.

[5 points]

(b) The brand averages in minutes are $\overline{Y}_A = 4.52, \overline{Y}_B = 3.08, \overline{Y}_c = 3.92$, and $\overline{Y}_D = 4.75$. Calculate the Sum of Squares Among Calculator Brands and test the equality of the 4 brands at the $\alpha = 0.01$ level.

[10 points]

(c) It is claimed that, on the average, the time for brand A is more than the mean of the times for brands C and D. On the basis of the given data, is there sufficient evidence to support this claim?

[8 points]

(d) Check to see if the follwing contrasts are mutually orthogonal.

$$\begin{aligned}\hat{\theta}_1 &= \overline{Y}_A - \frac{1}{2}(\overline{Y}_C + \overline{Y_D}) \\ \hat{\theta}_2 &= \overline{Y}_B - \frac{1}{2}(\overline{Y}_C + \overline{Y_D})\end{aligned}$$

[7 points]

7. Two diagnostic tests (A and B) were to be compared to see if the two tests provided similar results in detecting heartworms in dogs. Duplicate serum samples from 100 dogs were tested where 1 serum sample was tested by A and the other sample by B. The following table lists the results where positive by a test implies heartworm was detected by the test.

		Test A		
		Positive	Negative	Total
Test B	Positive	85	1	86
	Negative	9	5	14
	Total	94	6	100

Is it reasonable to conclude that tests A and B give different results?

[10 points]

8. Table below lists the average hypnotic susceptibility scores of six subjects before and after a course of hypnotic suspecibility training.

subject #	1	2	3	4	5	6
before	10.5	19.5	7.5	4.0	4.5	2.0
after	18.5	24.5	11.0	2.5	5.5	3.5

(a) Suggest a parametric and a nonparametric method for performing a test to see if the hypnotic suscepetibility (as measured by the hypnotic suspecibility score) can be increased by training. Discuss the necessary assumptions needed to justify the tests you suggest.

[5 points]

(b) Perform the nonparametric test you suggest in Part (a) and write your conclusions.

[10 points]

9. Let $Y_1, Y_2, \ldots Y_n$ be random variables with

$$\mathcal{E}(Y_i) = \mu_i, \qquad Var(Y_i) = \sigma^2, \quad \text{, and} \quad Cov(Y_i, Y_j) = \rho\sigma^2, 1 \leq i \neq j \leq n.$$

Find the expected value of $S^2 = (n-1)^{-1} \sum (Y_i - \overline{Y})^2$.

[10 points]

10. Appendix A contains partial results of a multiple regression of the number of minutes required to handle chemical shipments ("MINS") on number of drums in shipment ("DRUMS") and weight of shipment ("WEIGHT"). Use this output to answer the following questions.

(a) What is the meaning of R^2? (Be very specific)

[5 points]

(b) What would the R^2 be for a model that contained DRUMS as the only independent variable?

[10 points]

(c) Does the model containing both DRUMS and WEIGHT fit significantly better than the model containing only WEIGHT? Show the results of a statistical test to support your answer. Use $\alpha = .05$. State your null hypothesis, alternative hypothesis, decision rule and conclusions very clearly.

[10 points]

(d) Prior to conducting the analysis, the investigator had expected the coefficient for DRUMS to be very similar to the coefficient for WEIGHT. Construct a 95% confidence interval for the true difference of these two coefficients. Does the evidence support the investigator's prior hypothesis?

[15 points]

11. For each of the questions below, assume the usual linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{Y}$ is $n \times 1$, $\mathbf{X}$ is $n \times p$, $\boldsymbol{\beta}$ is $p \times 1$ and $\boldsymbol{\epsilon}$ is $n \times 1$.

    (a) Draw a picture of a situation in which the internally studentized and externally studentized residual for a point can differ dramatically. Be sure to identify that point.

    [10 points]

    (b) What are partial regression plots and how are they used?

    [10 points]

(c) For each of the following situations, *briefly* describe how you would determine that this problem exists and *briefly* describe what approach you would take to analyze the data.

i. Autocorrelated errors.

[10 points]

ii. Multicollinearity.

[10 points]

12. Suppose $\mathbf{Y}$ is a $n \times 1$ response vector and $\mathbf{X}_1$, $\mathbf{X}_2$ are two $n \times 1$ vectors of covariates which are orthogonal to each other, that is, $\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_2 = 0$.

    (a) Show that the estimated coefficients of $\mathbf{X}_1$ and $\mathbf{X}_2$ obtained from the multiple linear regression model $E[\mathbf{Y}] = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2$ are identical to the coefficients of $\mathbf{X}_1$ and $\mathbf{X}_2$ from the two simple linear regression models $E[\mathbf{Y}] = \mathbf{X}_1\beta_1^*$ and $E[\mathbf{Y}] = \mathbf{X}_2\beta_2^*$. That is, show that $\widehat{\beta}_1 = \widehat{\beta}_1^*$ and $\widehat{\beta}_2 = \widehat{\beta}_2^*$, where $\widehat{\beta}_1, \widehat{\beta}_2$ are the OLS estimators from the MLR model and $\widehat{\beta}_1^*, \widehat{\beta}_2^*$ are the corresponding estimators from the two SLR models.

    [10 points]

(b) Despite the equivalence of the coefficients from the MLR model with those from the SLR models, what is the advantage of fitting the MLR model? Provide a mathematical justification for your answer.

[15 points]

13. A rehabilitation center researcher was interested in examining the relationship between physical fitness status (below average, average, above average) prior to surgery of persons undergoing corrective knee surgery and time required in physical therapy (Y) until successful rehabilitation. The researcher also wants to use the age (x) of the patient as a concomitant variable. Assume the analysis of covariance (ANCOVA) model $Y = \beta_0 + \beta_1 (x - \bar{x}) + \delta_{20} z_2 + \delta_{30} z_3 + \epsilon$ is appropriate for these data where $z_2 = 1$ if physical fitness is average and zero otherwise and $z_3 = 1$ if physical fitness is above average and zero otherwise. Appendix B displays the data and SAS output corresponding to this model. Use this output to answer the following questions.

    (a) Draw a picture of the relationship between time required in physical therapy and age for each of the physical fitness status groups, as hypothesized by the above ANCOVA model. (Your plot does not have to be drawn to scale, but instead should merely demonstrate an understanding of the implications of the above ANCOVA model regarding the mean structure.)

    [5 points]

(b) What percentage reduction in error sums of squares can be attributed to including age as a covariate?

[10 points]

(c) Conduct a statistical test of the effect of age on time required in physical therapy. Use $\alpha = .05$. State your null hypothesis, alternative hypothesis, decision rule and conclusions very clearly.

[10 points]

(d) Conduct a statistical test of the effect of physical fitness on time required in physical therapy after adjusting for age. Use $\alpha = .05$. State your null hypothesis, alternative hypothesis, decision rule and conclusions very clearly.

[10 points]

(e) Construct a table showing the adjusted (for age) mean time required in physical therapy for each of the fitness status groups. Comment intuitively on the pattern of adjustments. The following data summaries should assist with your calculations.

Table 1: Data summaries for physical therapy data

Fitness status	$\bar{y}_i$	$\bar{x}_i$
Below average	38	26.187
Average	32	22.630
Above average	24	21.667

[10 points]

Appendix A

X'X Inverse Matrix

	INTERCEPT	DRUMS	WEIGHT
INTERCEPT	0.3066618451	-0.032952127	0.01486704
DRUMS	-0.032952127	0.0119556559	-0.011997386
WEIGHT	0.01486704	-0.011997386	0.0140366384

General Linear Models Procedure

Dependent Variable: MINS

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	40496.482	20248.241	641.64	0.0001
Error	17	536.468	31.557		
Corrected Total	19	41032.950			

R-Square	C.V.	Root MSE	MINS Mean
0.986926	5.947652	5.6176	94.450

Source	DF	Type I SS	Mean Square	F Value	Pr > F
DRUMS	1	38658.278	38658.278	1225.03	0.0001
WEIGHT	1	1838.204	1838.204	58.25	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
DRUMS	1	1187.6097	1187.6097	37.63	0.0001
WEIGHT	1	1838.2041	1838.2041	58.25	0.0001

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	3.324292153	1.07	0.3002	3.11083818
DRUMS	3.768110016	6.13	0.0001	0.61423456
WEIGHT	5.079587210	7.63	0.0001	0.66554757

Appendix B

General Linear Models Procedure

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1081.8343	360.6114	1169.72	0.0001
Error	20	6.1657	0.3083		
Corrected Total	23	1088.0000			

R-Square	C.V.	Root MSE	Y Mean
0.994333	1.735114	0.5552	32.000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
AGECENTR	1	835.75055	835.75055	2710.95	0.0001
Z2	1	34.49993	34.49993	111.91	0.0001
Z3	1	211.58378	211.58378	686.32	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
AGECENTR	1	409.83425	409.83425	1329.39	0.0001
Z2	1	12.77841	12.77841	41.45	0.0001
Z3	1	211.58378	211.58378	686.32	0.0001

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	34.95046430	163.79	0.0001	0.21338079
AGECENTR	1.16728639	36.46	0.0001	0.03201483
Z2	-1.84737866	-6.44	0.0001	0.28694289
Z3	-8.72289277	-26.20	0.0001	0.33296397