Statistics 379 Midterm exam Sp 2000

Note: Only brief answers required..

1 a. Draw a Venn diagram of Y, a dependent variable, as a function of X1 and X2. Both X1 and X2 are correlated with Y and with each other.
b. Show on a Venn diagram the variance in Y contributed uniquely by X1, over and above X2.
c. Using the Venn diagram, illustrate the semipartial correlation of Y on X1.
d. Add X3 to the set of predictors. Draw a second Venn diagram in which X3 is a classic suppressor variable.
e. What are the changes that occur to R^2 and to all the B-weights when X3 is added - that is, what are the changes from Y'= A+B1X1+B2X2 to Y'= A+B1X1+B2X2+B3X3?

2 a. Describe, briefly, the condition under which a setwise regression should be used.
b. Describe, briefly, the condition under which a sequential regression should be used.
c. Describe, briefly, the condition under which a standard regression should be used.
d. Given the equation $SS_Y = SS_{REG} + SS_{RES}$, describe, in words, the meaning of each of the terms.
e. Given the regression equation Y' = A+B1X and the formula $CL_B = B \pm SE_B * t$:
If B1 = .416 and its standard error = .646, what are the 95% confidence limits for B1? Is B1 significantly different from zero and, if not, why not?

3. a. When two or more IVs are so highly correlated that the integrity of the regression solution is damaged, the condition is called —?—.
b. How can you determine that these unacceptably high intercorrelations exist among the IVs in your data set?
c. Give one "rule of thumb" for determining how many cases (i.e., N) a study should have.
d. (double credit) Crossvalidation: Assume that you run a regression with three IVs using N=400. After you split the sample into halves randomly (i.e., subsets A and B), and run separate regressions on each half, you find that $R^2_{YA} = .52$ and $R^2_{YB} = .48$. Complete this crossvalidation.

4. Do the following regression using the information given to you and creating additional information as needed. X1 = age, X2 = income, and X3 = the interaction of AGE x INCOME.

Y	X1	X2	X3
1	1	2	
2	2	3	
1	2	2	
5	4	5	
6	3	4	

a. What are the X3 scores for each of the 5 cases?
b. Interpret the information on <u>each line</u> of the SPSS regression output below.

Listwise Deletion of Missing Data

Multiple R	.92442
R Square	.85455
Adjusted R Square	.41818

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	3	18.80000	6.26667
Residual	1	3.20000	3.20000

F = 1.95833 p = .4736

---------- Variables in the Equation ----------

IV	B	SE B	Beta	T	p =
X1	1.400000	3.555278	.680641	.394	.7612
X2	3.200000	3.370460	1.779070	.949	.5165
X3	-.500000	.894427	-1.555050	-.559	.6755
A	-6.200000	7.547185		-.821	.5622

Statistics 379 Midterm Exam Spring 1997

Begin the answer for each part of each question on a new line.

For Questions 1 and 2: On several pages attached to this exam are analyses of a DV called COMPREH (reader's comprehension score) as a function of three IVs: COPY (ability to copy printed words rapidly), WRITE (writing ability), and VERBAL (verbal ability). The first pages give you summary statistics of each variable alone, including the frequency histograms. Then a correlation matrix is given, followed by two standard multiple regressions. The first involves all three IVs; the second involves only two - the IV called VERBAL was dropped in the second analysis.

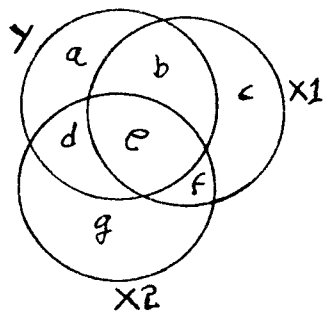
1. Evaluate the assumptions of (1a) normality, (1b) linearity, (1c) homoscedasticity, (1d) multicollinearity, and (1e) ratio of cases to variables, as best you can with the limited information that is provided. For each of these evaluations, list any additional information that you would you like to see if it was available.

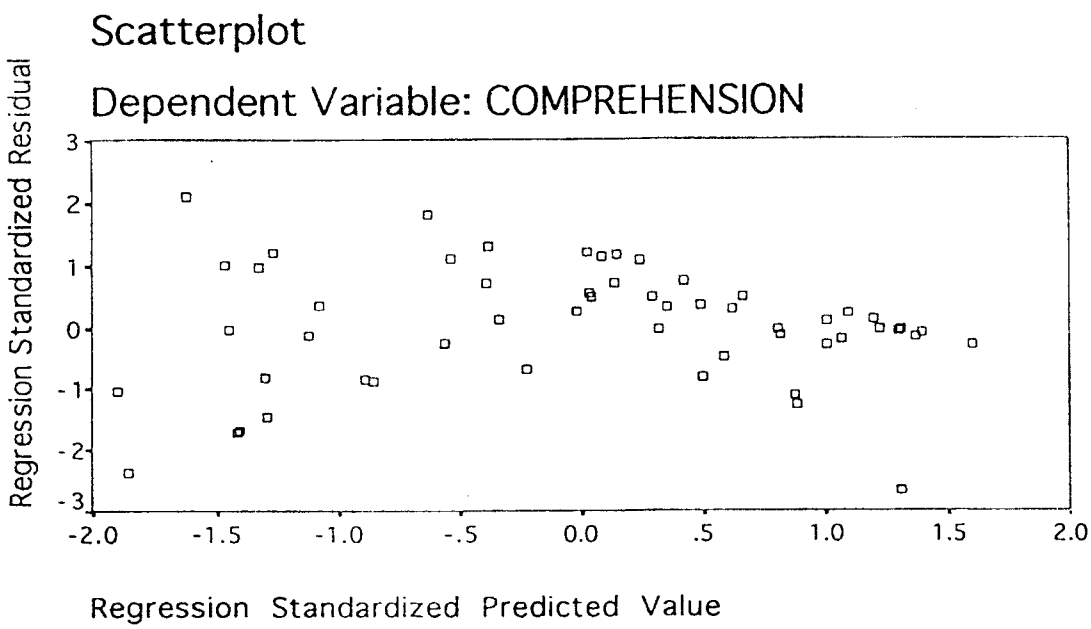
2. Examine the first multiple regression output. (Don't worry about the assumptions).
2a. How much COMPREH variance is accounted for?
2b. How stable is that estimate?
2c. What, specifically, does the term "Standard Error" mean here?
2d. Calculate the significance test for R^2. Show your calculations.
2e. Calculate the variance of COMPREH. (Sufficient information exists on the output).
2f. Interpret the relative importances of the 3 IVs in accounting for COMPREH.
2g. Using the second MR, test for the incremental significance of VERBAL over-and-above the contributions of the other two IVs. Show your calculations.

3. (3a). What are the reasons for carrying out a crossvalidation of R^2 and the B-weights? (3b) Describe a complete crossvalidation procedure, step by step.

4. Short answer. (3a) Contrast Experimental and Nonexperimental design. (3b) Define a residual. (3c) What statistic can be used to determine the degree of covariation (or its opposite, independence) between two variables if the variables are qualitative rather than quantitative? (3d) List 3 conditions that can lead to a "deflated" correlation coefficient. (3e) Contrast the conditions under which you would employ Standard MR and Stepwise MR. (3f) Contrast the utilities of B-weights and Beta-weights.(That is, when would you use one instead of the other?)

5. Using the Venn diagram below, in which each letter represents the variance of its bounded area, provide the ratio of areas that represents:
(5a) $R^2_{Y.12}$ (5b) the unique effect of X1 (5c) r^2_{12} (5d) $R^2_{Y.12} - R^2_{Y.1}$





Number of valid observations (listwise) = 55.00

Variable COMPREH

Mean	8.221	Std Dev	1.855
Kurtosis	.871	S.E. Kurt	.634
Skewness	-1.256	S.E. Skew	.322
Minimum	2.60	Maximum	10.00

Valid observations - 55 Missing observations - 0

Variable COPY

Mean	49.636	Std Dev	18.274
Kurtosis	-.801	S.E. Kurt	.634
Skewness	-.180	S.E. Skew	.322
Minimum	12.00	Maximum	89.00

Valid observations - 55 Missing observations - 0

Variable VERBAL

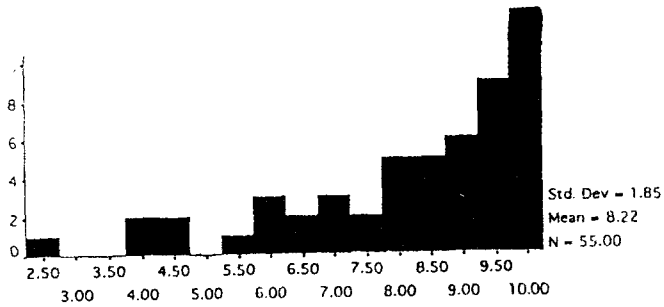
Mean	52.964	Std Dev	19.817
Kurtosis	-.608	S.E. Kurt	.634
Skewness	-.455	S.E. Skew	.322
Minimum	7.00	Maximum	88.00

Valid observations - 55 Missing observations - 0

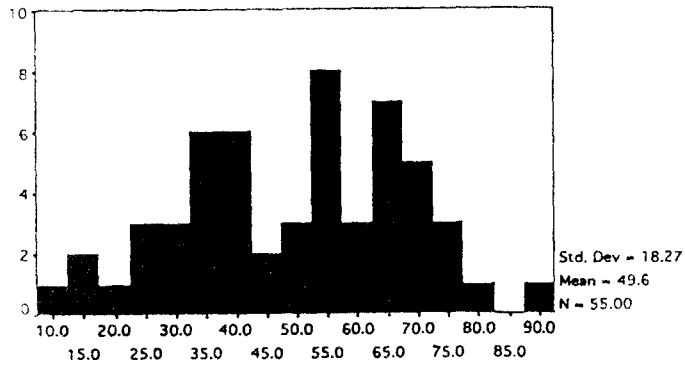
Variable WRITE

Mean	57.164	Std Dev	22.315
Kurtosis	-1.017	S.E. Kurt	.634
Skewness	-.156	S.E. Skew	.322
Minimum	12.00	Maximum	96.00

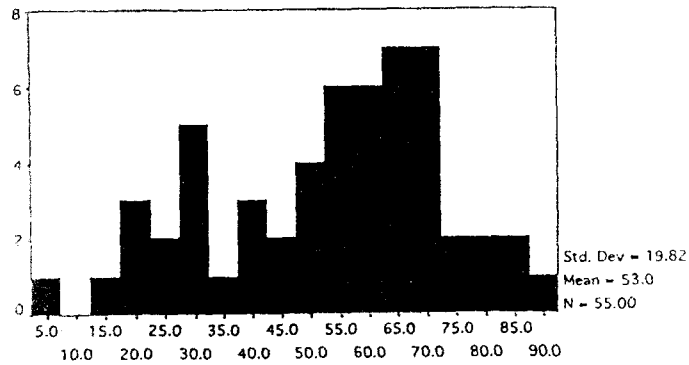
Valid observations - 55 Missing observations - 0



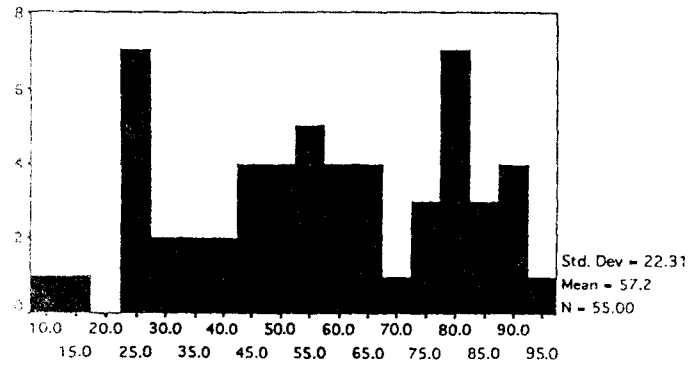
COMPREHENSION



Copy Speed



Verbal ability



Writing Ability

-- Correlation Coefficients --

	COMPREH	COPY	VERBAL	WRITE
COMPREH	1.0000	.4123	.6767	.6278
	(55)	(55)	(55)	(55)
	p= .	p= .002	p= .000	p= .000
COPY	.4123	1.0000	.3266	.4783
	(55)	(55)	(55)	(55)
	p= .002	p= .	p= .015	p= .000
VERBAL	.6767	.3266	1.0000	.6093
	(55)	(55)	(55)	(55)
	p= .000	p= .015	p= .	p= .000
WRITE	.6278	.4783	.6093	1.0000
	(55)	(55)	(55)	(55)
	p= .000	p= .000	p= .000	p= .

Multiple R	.73731
R Square	.54362
Adjusted R Square	.51677
Standard Error	1.28938

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	3	100.99567	33.66522
Residual	51	84.78779	1.66251

F = 20.24969 Signif F = .0000

Variable	B	SE B	Beta	T	Sig T
COPY	.012610	.010947	.124236	1.152	.2547
VERBAL	.043143	.011181	.460931	3.859	.0003
WRITE	.023898	.010686	.287506	2.236	.0297
(Constant)	3.943884	.611958		6.445	.0000

Multiple R	.64061
R Square	.41039
Adjusted R Square	.38771
Standard Error	1.45140

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	2	76.24276	38.12138
Residual	52	109.54070	2.10655

F = 18.09658 Signif F = .0000

Variable	B	SE B	Beta	T	Sig T
COPY	.014742	.012307	.145236	1.198	.2364
WRITE	.046408	.010079	.558321	4.605	.0000
(Constant)	4.836316	.637760		7.583	.0000

Stat 379 Mid-term Xm Spring, 1996

Question #1 is worth 40 points. Questions 2 - 4 are worth 20 points each.

1. Below is a table of results from a standard regression. (The data are from Table 5.18 in your Tabachnik & Fidell textbook). The DV, LTIMEDRS, is the "log of number of visits to a health professional". The IV LPHYHEAL is the log transform of a physical health index. SSTRESS is the square root of an index of stress. MENHEAL is an untransformed measure of mental health.

1.a What kinds of raw score frequency distributions would require the kinds of data transforms that were used?

1.b Confirm, using the correlation data in the table and the ß-weights that the multiple $R^2 = .38$. Show your calculations.

1.c Interpret the results of the regression. What variables are useful in accounting for LTIMEDRS? What does sr^2 tell you about the two important predictors? How would you improve on the accuracy of the sr^2 values for these data without running more subjects?

1.d Write the raw score prediction equation (using B-weights that are given). Write the standardized prediction equation (using ß-weights). How could the precision of the equation be improved (without running additional subjects).

1.e What are the results in the table that tell you which IV is the most important in accounting for the DV? (List all the sources that provide this information.)

1.f What does the "Adjusted R^2" in the table tell you?

TABLE 5.18 STANDARD MULTIPLE REGRESSION OF HEALTH AND STRESS VARIABLES ON NUMBER OF VISITS TO HEALTH PROFESSIONALS

Variables	LTIMEDRS (DV)	LPHYHEAL	SSTRESS	MENHEAL	*B*	β	sr^2 (unique)
LPHYHEAL	.59				1.040**	0.52	.19
SSTRESS	.36	.32			0.016**	0.19	.03
MENHEAL	.36	.51	.38		0.002	0.02	
				Intercept =	− 0.155		
Means	0.74	0.65	13.40	6.12			
Standard deviations	0.42	0.21	4.97	4.19			
						R^2 =	.38
						Adjusted R^2 =	.37
						R =	.61**

** $p < .01$.

2. Given that $r^2_{Y1} = .30$, $r^2_{Y2} = .25$, $r^2_{12} = .60$, $R^2_{Y.12} = .32$, N = 60.

2.a Perform a hierarchical regression, in the order X1, X2. List R^2 and its test of significance, for both steps. At Step 2, also present sr^2 and test its significance. (You should use the F-table in the back of your text, but do not open the book to any other section).

2.b Draw a Venn diagram for each of the two steps. Indicate, on the appropriate diagram, $R^2_{Y.1} = .30$ and $R^2_{Y.12} = .32$.

3.

3.a Set up the residual scatterplot in order to observe whether homoscedasticity, normality and linearity hold for the data. Make an assessment of these assumptions "by eye", without using statistical tests. Specify the reasons for your assessment.

The best-fitting straight line is

$$Y' = 10 + 2X.$$

The actual data are:

Subject	Y	X
1	140	60
2	170	70
3	110	50
4	130	70
5	120	60
6	90	40
7	80	30
8	60	30

3.b Why does r_{xe} equal zero? (r_{xe} is the correlation between the residuals and x.)

4. Short answers.

4.a List (no description necessary) 3 factors that limit the maximum size of any r_{XY}.

4.b When is it appropriate to use a biserial correlation coeficient?

4.c What is a Type I statistic? (For example: a Type I Sum of Squares or a Type I semipartial correlation.)

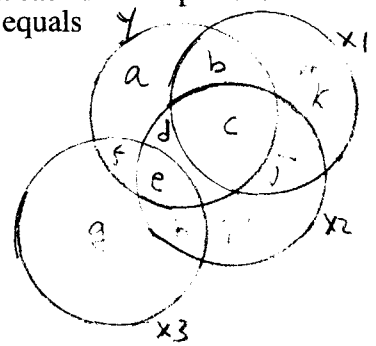
4.d What is the pattern of intercorrelated variables that produces a suppressor IV? Answer by drawing the Venn diagram for this pattern and label the variables appropriately.

Midterm xm Stat 379 Mar 95

1. a) Which of the problem conditions given below can affect the *size* of the correlation coeficient, r_{xy} ?
b) For each of your answers, describe briefly whether that condition *inflates* or *deflates* r_{xy} and *why* it has that effect.

heteroscedasticity in Y on X
↓ restricted variance of X
↓ different shapes for the X and Y frequency distributions
↔ unreliability of X or Y
; nonlinear X-Y relationship

2. Below is a Venn diagram in which each letter represents the variance of its bounded area. Write the ratio of letters that equals
a) $R^2_{Y.123}$
b) $R^2_{Y.123} - R^2_{Y.12}$
c) sr^2_3
d) r^2_{12}



3. A "Statistical Regression" stepwise procedure is run on the following model: Y = X1 X2 X3 X4 X5. N = 100. The R^2 at each step was as follows.

(1) $R^2_{Y.2} = .50$
(2) $R^2_{Y.21} = .60$
(3) $R^2_{Y.13} = .64$
(4) Stepping stopped.

a) Describe the selection and deletion events that occurred at each step.
b) What statistical tests were involved in selection and deletion?
c) Describe the selection process that chose X2 at the first step. Present the formula for the statistical test that was used to select X2 and calculate the F-ratio for its significance. DO NOT look up the statistical significance of the F-ratio (i.e., do not look up the p- value; trust me, it is significant).

Stat 379 Mid xm Sp 95

4. For each of the following situations described below:

1) Name the kind of multiple regression procedure you would use to answer the question.

2) What are the DV and IVs for that regression?

3) What statistics would you examine in order to assess your hypothesis? (A partial, i.e., incomplete, list of possibly relevent statistics includes r^2_{12} , $R^2_{Y.123}$, sr^2 , B- weights, beta-weights, etc.).

a) You are an anthropologist who is interested in the nutritional habits of a hunter-gatherer group. You think the number of calories expended each day determines the number of calories consumed the next day. However, a person's status in the group may also affect the number of calories consumed and you would like to remove the effect of this "nuisance variable" to assess, conservatively, the hypothesis that was stated above (in the second sentence).

b) In a survey of people's automobile-buying behavior, you collect 32 measures of "buyer characteristics" that may or may not relate to the cost of the automobile they purchased. How would you create a useful subset of these measures that will have a reasonable ability to predict how much money a potential buyer will spend? What factors will determine how many measures you use, ultimately?

c) You are interested in whether Gender and Number of Friends affect the amount of time people spend each week in listening to their friend's problems. You measure this "total weekly empathy time" for every person. You also record how many friends each person has and, of course, the person's gender.

5. Brief answers:

a) Interpret the following results.

DV: Y N = 60 $R^2_{Y.123}$ = .40 F(3,56) = 37.38, p< .001

IV	B	ß	t	p≤
X1	.202	.30	3.5	.001
X2	.101	.08	1.3	.20
X3	.002	.15	2.5	.01

b) multivariate outlier
c) homoscedasticity
d) multicolineratity
e) kurtosis

Julia Irwin

Spring, 1993 Statistics 379 Mid-term Xm

Begin each answer on a new line.

Use this list of regression procedures for the first 4 questions.

~~A. standard (simultaneous)~~
B. hierarchical, all IVs continuous
C. all IVs categorical = ANOVA
D. mixed continuous and categorical = ANCOVA
E. stepwise (specify forward, backward, or stepwise)
F. setwise

Comp + phon ⇒ DV

Questions 1 to 4: For each of the situations described below:

(a) list the letter of one appropriate analysis (more than one may be appropriate, but list only one)
(b) list the IVs for the analysis
(c) describe, simply, the hypothesis (or hypotheses) you want to test
(d) what statistics, specifically, you would examine in order to assess your hypotheses (e.g., R^2, b-weights, Type I or Type III SS, semi-partial correlation, etc.)

1. A child's reading comprehension is thought to be a function of both the child's ability to *sound out* printed words (as measured by a test in which the child reads a list of nonwords, like "brutoom") and the child's ability to comprehend what he or she *hears* (as measured by a test of the child's comprehension after listening to a story). Reading comprehension is measured by giving the child 5 paragraphs to read and testing the child with questions on the content of each paragraph; a single summary score is calculated.

2. You believe that the average number of calories a person consumes each day may be explained by the number of calories he expends. You collect these measures on a diverse sample of men who live in Mansfield. The sample varies quite a lot with regard to age, and you have reason to suspect that age is somewhat correlated with both of the other variables. You wish to remove any influence of age on your interpretation.

3. You would like to find out if men and women differ in the average amount of time spent in listening to their friends' problems. You measure "empathy time" for your sample of men and women. However, you find that the number of friends per subject varies substantially (although it is about the same overall for men and women) and this can reduce the power of the analysis.

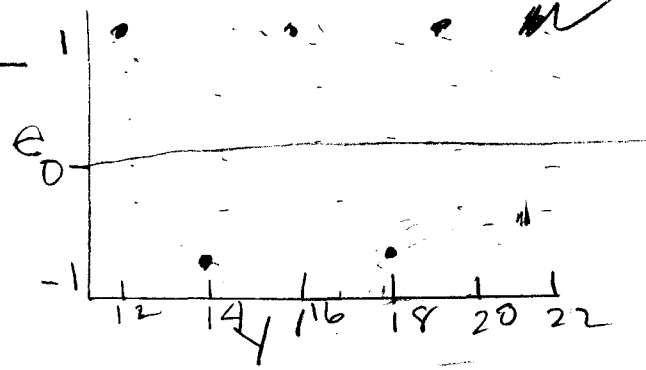
4. In a large-scale survey of people's automobile-purchasing behavior, you collect over 30 measures of "buyer characteristics" that may or may not have some relation to whether a person bought a luxury car in the past 5 years. How would you create a subset of these 30 measures that will have reasonable predictive ability?

5. Given that $Y' = 10 + 2X$

(a) residualize Y on X and plot the error scores against Y'.

(b) Evaluate homoscedasticity. (c) Evaluate the correlation between X and the residuals.

Sub	Y	X	Y'	e
1	13	1	12	1
2	17	3	16	1
3	13	2	14	-1
4	17	4	18	-1
5	21	5	20	1
6	21	6	22	-1



6. Describe briefly:

(a) multicolinearity, singularity

(b) partial regression coeficient

(c) bivariate normal distribution

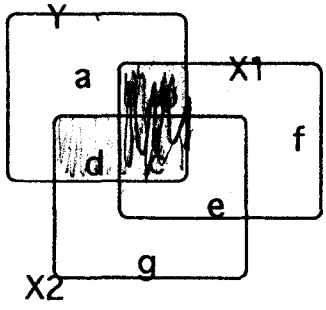
(d) classical suppressor variable

(e) Why assess shrinkage in R-square?

7. (a) Given $A = \begin{matrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{matrix}$ and $B = \begin{matrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{matrix}$

find the product **AB**.

(b) What areas of the Venn diagram below illustrate the squared semipartial correlation between Y and X2?

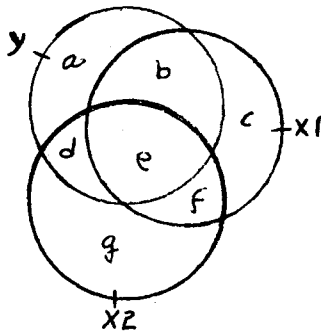


$R^2_{Y12} - R^2_{Y1} =$

Statistics 379 Midterm exam Spring, 1991

1. Only brief (1 to 3 line) answers are to be given to this question which concerns data screening. (a) List two examples of causes of data inaccuracy. (b) Define homoscedasticity. (c) What is a multivariate outlier? (d) List three conditions that artificially limit the zero-order correlation coefficient and one condition that inflates it.

2. Using the Venn diagram below, in which each letter represents the variance of its bounded area, answer what ratio of areas (i.e., what combination of letters) represents:
(a) $R^2y.12$ (b) the unique effect of X1 (c) r^2_{12} (d) $R^2y.12 - R^2y.1$



3. In both of the following examples, (1) List the multiple regression technique that you think is most appropriate, (2) briefly (2 to 4 lines) give the question or questions that motivated your choice and (3) list the statistics you would exam to assess your hypothesis. Choose two different kinds of MR for the two parts of this question.
Ss= 300 children in grades 2 through 6. DV= reading ability.
(a) X1= grade, X2= age, X3= sex.
(b) X1 through X18= a collection of demographic variables, such as family income, parents' education, number of telephones in home, distance of home from city center, etc.

4. In both of the following examples, (1) List the multiple regression technique that you think is most appropriate, (2) briefly (2 to 4 lines) give the question or questions that motivated your choice and (3) list the statistics you would exam to assess your hypothesis. Choose two different kinds of MR for the two parts of this question.
Ss= 100 senior year university students from 100 different universities.
DV= National Graduate Record Exam score on "Verbal". (a) X1= Private or Public university, scored as zero or one, X2= family income, X3= IQ.
(b) X1= Private or Public university, scored as zero or one, X2= family income, X3= IQ. [Yes, this is the same as part (a)].

Midterm Xm Stat 379 Sp 90

1. Define briefly: a. residual sum of squares (SS_{res}) b. SS_{reg}
   c. r_{xy} in terms of standard scores (Z scores)
   d. r_{xy} in terms of SS_{res} e. General Linear Model

2. Draw the Venn diagram and estimate $R^2_{y.12}$ for each. If you do not have enough information to make an exact estimate, make as precise a statement as you can under the circumstances.
   a. $r^2_{y.1} = .30$ $r^2_{y.2} = .30$ $r^2_{12} = 0$
   b. $r^2_{y.1} = .30$ $r^2_{y.2} = .30$ $r^2_{12} = .15$
   c. $r^2_{y.1} = .30$ $r^2_{y.2} = .30$ $r^2_{12} = 1.0$
   d. $r^2_{y.1} = .30$ $r^2_{y.2} = 0$ $r^2_{12} = .30$

3. We measure Y, X1, and X2 on 20 subjects: $r^2_{y1} = .45$ $R^2_{y.12} = .51$
   a. Test the significance of $r^2_{y(2.1)}$
   b. List the factors that affect the significance of any correlation.

4. We study the relation between the dv Y (score on social skills), X1 (verbal ability), and X3 (intelligence). We hypothesize that social skills are related to intelligence. However, because verbal ability may affect both the intelligence score and the social skills score, we decide to control for it. In terms of Y, X1, and X2, briefly describe:
   a. primary variance
   b. secondary variance
   c. error variance
   d. control by residualization

5. a. Express the normal (simultaneous) equations below in the form of the matrix and vectors in $Rb = r$. (That is, simply transcribe the numbers in the equations into their matrix form.)

$$.3 + (.6)\,.4 = .5$$
$$(.6)\,.3 + \quad .4 = .5$$

   b. List the two assumptions about Y, X1, and X2 that lead, necessarily, to the normal equations.

6. Describe briefly:

a. blockwise selection in stepwise regression

b. Significance Level to Stay

c. reliability

d. validity

e. the interpretation of a regression weight, ß

7. Consider the following field experiment. Last year, I regressed reading ability on age, sex, and verbal ability for a random sample of 30 third grade children.

a. Which predictor would you expect to be weakest? Why?

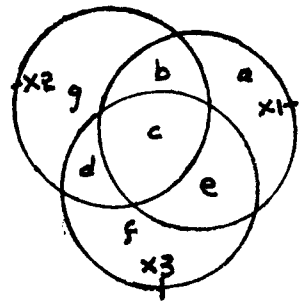
b. This year, I repeated the experiment with a new random sample. Do you think the R^2 stayed about the same or was reduced? Why?

c. I could improve the significance of the R^2 by increasing the number of children sampled. Based on standard recommendations, how many children should have been tested for this experiment?

Statistics 379 Mid-term exam Spring, 1989

1. List (no description necessary): (a) 3 factors that affect the maximum size of r_{xy} and (b) 3 factors that affect the precision (statistical significance) of r_{xy}.

2. In the Venn diagram below, each letter represents the variance of its bounded area. What ratio of areas (i.e., what ratio of letters) represents: (a) $R^2_{1.23}$ (b) $r^2_{12.3}$ (c) $r^2_{1(2.3)}$



3. Given: r^2_{y1}= .09 r^2_{y2}= 0 r^2_{12}= .25 $R^2_{y.12}$= .12 $ß_1$= .4 $ß_2$= -.2
(a) Draw the Venn diagram that illustrates the 3 correlations
(b) Which variable is the suppressor and why?
(c) Find the predicted score for S_1, who has Z_1= 1.0 and Z_2= -1.0 and S_2, who has Z_1= 1.0 and Z_2= 1.0 (No explanation needed).
(d) Is X1 or X2 more important in accounting for Y? How much more?

4. Given r^2_{y2}= .4 r^2_{y1}= .3 r^2_{y3}= .05 and r_{12}= r_{13}= r_{23}= 0 : Perform a forward solution for the selection of variables. Use a criterion F-to-enter of 4.00 Show calculations. Use the following format for the answer in your test booklet:

N=15

	Variable Selected	F-ratio	R^2
Step 1			
Step 2			
Step 3			

5. In assessing the contributions of X1 - X4 in accounting for the variance of Y, the following regression equation was obtained:

$$Z' = (.021)Z_1 + (.500)Z_2 + (.300)Z_3 + (.001)Z_4$$

For your information, the intercorrelation matrix is given:

	Y	1	2	3	4
Y	1	.2	.4	.3	.2
1		1	0	0	.6
2			1	0	0
3				1	0
4					1

The ß weights of X1 and X4 were found to be not significant, by t-test. List the steps you would follow to determine a final regression equation.

6. Consider a dependent variable Y and 3 predictors, X1,X2,X3. (a) Write the expressions for the general linear model and the minimum least squares error. (b) Write the normal (simultaneous) equations that express the unknown ß weights as a function of the observed correlations between Y,X1,X2, and X3. (c) Re-write the normal equations in matrix notation. (d) Re-write the matrix equation into the form used to solve for the ß weights (e.g., ß= ...).

Statistics 379 Mid-term Xm Spring, 1988

1. For the following data set, the best-fitting straight line (Y on X) has coeficients a= .5 and b= .5
   Calculate and list the residual Y for each pair.

Y	X
1	1
1	2
2	3
3	4

2. How do the frequency distributions and variances of X and Y affect the size of a simple (zero-order) r^2 ?

3. Describe briefly:
   (a) homoscedasticity
   (b) X reliability
   (c) r is a dimensionless quantity

4. Match items between the two columns. The model is Y= a + bX.

(a) MSR	(1) amount of Y variability predicted by x
(b) SS_{res}	(2) variance of observations around the regression line
(c) S_b	(3) $\Sigma(Y-Y')^2$
(d) $r^2\Sigma y^2$	(4) standard error of estimate
	(5) regression coeficient standard error

5. Draw a Venn diagram illustrating the following intercorrelations. $(r_{y1})^2$= .4 $(r_{y2})^2$= .4 $(r_{y(2.1)})^2$= .15 . Label all areas.

6. **Note**: This question has <u>double</u> weight.
   (a) Name 5 analytic procedures that are suitable for selecting a subset of independent variables from a larger original set for a prediction equation.
   (b) Describe each procedure briefly. (Outline the steps of the analysis itself, not the computer program's language).
   (c) List the advantages and disadvantages of each procedure.

7. After obtaining a multiple regression prediction equation, I apply it to an applicant and predict a dependent variable score of 10.00 . The standard error ($S_{y\cdot}$) equals 2.0 . Given that the appropriate F at alpha= .05 is 4.0, calculate the 95% confidence limits for the predicted score.

8. (a) Add: **B + C**. (b) Multiply: **BxC**.

B	C
1 2 3	2 1 3
2 1 1	1 2 1
3 1 2	2 2 1

9. On the attached computer printout, identify and briefly comment on the result for each of the circled terms (a) through (j).

VARIABLE	N	MEAN	STD DEV	SUM	MINIMUM	MAXIMUM
X1	5	3.00000000	1.58113883	15.00000000	1.00000000	5.00000000
X2	5	3.00000000	1.58113883	15.00000000	1.00000000	5.00000000
X3	5	3.00000000	1.58113883	15.00000000	1.00000000	5.00000000

a

PEARSON CORRELATION COEFFICIENTS / PROB > |R| UNDER HO:RHO=0 / N = 5

	X1	X2	X3
X1	1.00000 0.0000	0.70000 0.1881	0.60000 0.2848
X2	0.70000 0.1881	1.00000 0.0000	0.90000 0.0374
X3	0.60000 0.2848	0.90000 0.0374	1.00000 0.0000

b

GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: X1

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F	R-SQUARE	C.V.
MODEL	2	c 4.94736842	2.47368421	d 0.98	0.5053	0.494737	52.9813 e
ERROR	2	5.05263158	2.52631579		ROOT MSE		X1 MEAN
CORRECTED TOTAL	4	10.00000000			1.58943883		3.00000000

SOURCE	DF	TYPE I SS	F VALUE	PR > F	DF	TYPE III SS	F VALUE	PR > F
X2	1	f 4.90000000	1.94	0.2983	1	g 1.34736842	0.53	0.5412
X3	1	h 0.04736842	0.02	0.9036	1	0.04736842	0.02	0.9036

PARAMETER	ESTIMATE	T FOR HO: PARAMETER=0	PR > |T|	STD ERROR OF ESTIMATE
INTERCEPT	0.94736842	0.56	0.6339	1.70253193
X2	0.84210526	0.73	0.5412	1.15310012
X3	-0.15789474 i	-0.14	0.9036	1.15310012 j

Statistics 379 Midterm examination March, 1987

1. Calculate the correlation coeficient r_{xy} for the following standard score data. Don't use a calculator; show your work.

	Z_x	Z_y
S1	1	-1
S2	0	0
S3	-1	1

2. List the measurement criteria that differentiate the use of the following coeficients. When is each appropriate? a) Pearson product-moment b) point-biserial c) biserial d) tetrachoric

3. List 3 factors that affect the precision (i.e., statistical significance) of the regression equation and describe briefly <u>how</u> precision is affected.

4. For r_{yx}, describe briefly with a) scattergram and b) equation(s) what a <u>residual</u> is. c) In words, what does this equality mean and why is it true: $r_{12.3} = r_{e1e2}$?

5. Given: $R^2_{y.1}$= .4536 $R^2_{y.2}$= .1557 $R^2_{y.12}$= .5057 N= 20
Set up an F test to determine if X1 accounts for significant Y variance over and above X2. Specify the appropriate degrees of freedom. (Don't calculate: just put the appropriate numbers in the correct formula).
F(?, ?) = ?

6. Multiple regression theory has two major assumptions: the "linear model" and the requirement that the sum of the squared residuals be as small as possible. Present each assumption in its symbolic (i.e., mathematical) notation.

7. Short answers: a) List 2 types of reliability coeficient b) What are the criteria for a suppressor variable?

8. The normal equations relate the regression weights (betas) to the observed correlations. Given the following information, re-create the original normal equations. r_{12}= .5 r_{y1}= .3 r_{y2}= .4 $R_{ij}\beta_j = r_{yj}$

9. Using Venn diagrams, illustrate (to a reasonable approximation):

a) r^2_{y1}= .50 b) r^2_{y1}= .50 $r^2_{y(2.1)}$= .25 r^2_{12}= .25

c) r^2_{y1}= .50 $r^2_{y(2.1)}$= .25 r^2_{12}= 0

10. a) A Stepwise analysis is run on Model Y = X1 X2 X3 X4 X5. The R^2 at each step is given below. Describe the selection and deletion events that occur at each step. 1) $R^2_{y.2}$= .50 2) $R^2_{y.21}$= .60 3) $R^2_{y.13}$= .64 4) stepping stopped. b) How can the stability of the final prediction equation be assessed?

Statistics 370 Make-up Mid-term exam

Spring, 1987

Calculators may be used but show all calculations.

1. Calculate $R^2_{1.23}$ given $r^2_{1(2.3)}= .36$, $r^2_{1(3.2)}= .19$, $r^2_{12}= .45$, $r^2_{13}= .28$, $r^2_{23}= .02$

2. Given $r^2_{y1}= .3$, $r^2_{y2}= .4$, $r_{12}= 0$ and N= 33:
   a) calculate $R^2_{y.12}$ b) test its significance

3. Calculate the beta weights for X1 and X2 using the following:
   $r_{12}= .3$, $r_{y1}= .5$, $r_{y2}= .3$

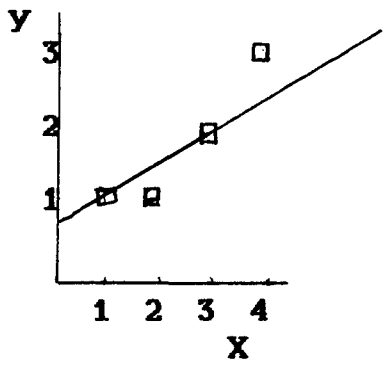
$$\mathbf{R}^{-1}= \begin{bmatrix} 1.10 & -.33 \\ -.33 & 1.10 \end{bmatrix}$$

4. Using the formulas for variance and covariance, demonstrate that variance is a special case of covariance.

5. Brief answers: Describe a) singularity b) identity matrix c) inverse matrix d) matrix of sums and cross products.

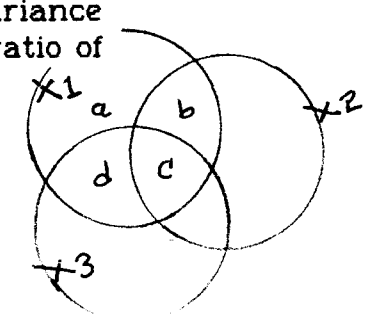
6. a) Distinguish between secondary variance and error variance.
   b) How can you control for secondary variance in
   1) experimental research and 2) nonexperimental research?
   Give an example for each.

7. For the scattergram below, the best-fitting straight line has the coeficients: a= .5 and b= .5 Using the definitional formula for residual variance, calculate the mean square residual (MSR).



The (x,y) data pairs are: (1,1), (2,1), (3,2), (4,3)

8. In the Venn diagram below, each letter represents the variance of (only) its bounded area. What ratio of areas (i.e., what ratio of letters) gives a) $r^2_{1(2.3)}$ and b) $r^2_{12.3}$



STATISTICS 309

MID-TERM EXAM OCT 86

1. Brief answers:

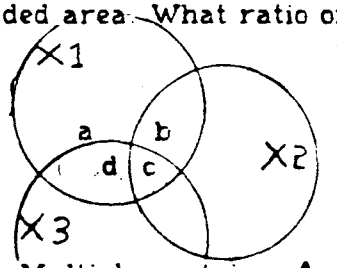
a. What is the purpose and method of double cross-validation?

b. Why is r_{xy} a "dimensionless number", i.e., how does correlation relate two variables that may be on different scales of measurement?

c. Define homoscedasticity.

2. a. The correlation between age and reading speed for the students in a third grade class (N= 100) is found to be moderate, r= .40. We then determine the correlation again, but based on all grades (1 through 5) together (N= 20 in each of the 5 grades). How and why is the correlation likely to change?

b. In the Venn diagram below, each letter represents the variance of its bounded area. What ratio of areas (i.e., what ratio of letters) gives $r^2_{12.3}$?



a = $\chi_{1.23}$

b = $\chi_{12.3}$

c = $\chi_{1.23}$

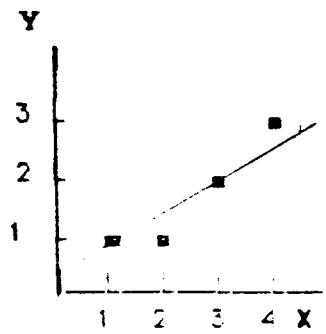
d = $\chi_{12.3}$

c. Multiply matrices **A** and **B** :

A		
1	2	3
4	5	6

B	
1	2
3	4
5	6
7	8

3. For the scattergram below, the best-fitting line has the coeficients a=.5 and b= .5. Using the definitional formula, calculate MSR (i.e., $S^2_{y.x}$).



(x,y) : (1, 1), (2, 1), (3, 2), (4, 3)

$SS_{res} = \Sigma(y - \hat{y})$

$MSR = \frac{SS_{res}}{n - k - 1}$

4. The following ordered regression is done N= 3

$R^2_{y.1}$= .360

$R^2_{y.14}$= .413

$R^2_{y.142}$= .430

$R^2_{y.1423}$= .439

$F = \frac{R^2_{y.14} - R^2_{y.1} / 3-0}{1 - R^2_{y.14} / 30}$ (2)

a. Does X4 account for a significant amount of variance in Y over and above X1? Make the test, specifying the appropriate degrees of freedom for the F ratio.

b. $r^2_{y(3.124)}$= ?

Tabled F values: F(1,25)= 4.25 F(2,25)= 3.38 F(4,25)= 2.76
F(1,27)= 4.21 F(2,27)= 3.35 F(4,27)= 2.73
F(1,28)= 4.20 F92,28)= 3.34 F(4,28)= 2.71

5. Some social scientists are interested in whether TV political advertising can affect viewer's attitude toward the public funding of religious schools. In a laboratory experiment, they show three kinds of political propaganda films (mild/ moderately strong/ strong) to subjects, but only one of the three films to each subject. Subjects are then tested for attitude toward public funding for religious schools.

a. Give a brief definition of primary, secondary, and error variance together with an example in this experiment. Each example should name a variable that could reasonably be expected to be involved.

b. In this experiment, specifically, how can we control unwanted variance?

c. After running the lab experiment, it was decided to take the study "into the field" and the films were run on television. Give an example of secondary variance in this situation and how it can be controlled, statistically.

6. What is the effect on b_1 in the model $y = a + b_1X_1$ of the omission of X_2 from the model when:

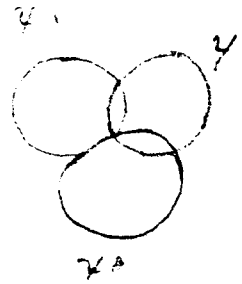
a. $r_{y2} > 0$, $r_{12} = 0$

b. $r_{y2} > 0$, $r_{12} > 0$

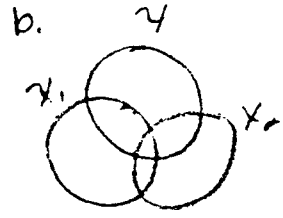
c. $r_{y2} = 0$, $r_{12} > 0$

true $y = a + b_1X_1 + b_2X_2$

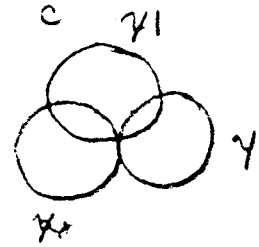
a



b.



c



Statistics 379
Spring 1985
Mid-term Exam

1. Short answers to the following:

   (a) Define secondary variance.

   (b) Contrast experimental and nonexperimental research.

   (c) How can you control secondary variance in experimental research?

   (d) How will you know a supressor variable when you see one?

2. Statistical control of secondary variance can be accomplished by residualizing.

   (a) Give a concrete example (invent one) in which spurious effects can be removed by partialing. (Just name three variables and state which ones are partialed and why.)

   (b) Demonstrate what a residual is, using a scattergram.

   (c) Given the partial correlation $r_{y1.2} = .6$, what can you say about $r_{2\varepsilon_y}$ and $r_{\varepsilon_1\varepsilon_y}$? Why?

3. An ordered regression is run on 4 IVs with the following results:

	STEP	
N = 60	1 $R^2_{y.1}$	= .30
	2 $R^2_{y.12}$	= .45
	3 $R^2_{y.12.3}$	= .55
	4 $R^2_{y.1234}$	= .60

   (a) Set up the significance test, using the appropriate actual numbers, for $R^2_{y.1234}$. (Don't look up the significance level.)

   (b) Assume that the test in part (a) was significant. What can you say about the effect of X_1 and what other information would be helpful for interpreting its effect?

4. List the similarities and differences between b weights and beta weights.

5. Given the following data: Draw a Venn diagram (circle diagram) illustrating the intercorrelations.

$r^2_{y1} = .4 \quad r^2_{y2} = .4 \quad r^2_{y(2.1)} = .15$

Label the above areas.

6. Given the information.

$\beta_1 + .5\beta_2 + .6\beta_3 = .2$
$.5\beta_1 + \beta_2 + .7\beta_3 = .3$
$.6\beta_1 + .7\beta_2 + \beta_3 = .4$

Rewrite, in matrix form suitable for solution, the above set of normal equations.

7. (a) Under what circumstances is a stepwise regression an appropriate procedure? Why? Give an example.

   (b) Describe the selection procedure for the variables in a stepwise regression. When does the procedure stop?

8. (a) In the predictive use of MR, applying the prediction formula to a new sample of subjects will not usually give us predictions as close as those we obtained on the orgininal sample. Why not?

   (b) One method of estimating the shrinkage in prediction is double cross-validation. Describe this procedure.

KAI ± N=27

Statistics 379 Mid-term xm Spring, 1984

1. Calculate the beta weights for X1 and X2 using the following:

$r_{12} = .3 \quad r_{1y} = .5 \quad r_{2y} = .3$

$$R^{-1} = \begin{bmatrix} 1.1 & -.33 \\ -.33 & 1.1 \end{bmatrix} \qquad \beta_j = R^{-1} r_{jy}$$

$$\begin{bmatrix} 1.1 & -.33 \\ -.33 & 1.1 \end{bmatrix} \begin{bmatrix} .5 \\ .3 \end{bmatrix}$$

$\beta_1 = 1.1(.5) - .33(.3) = .55 - .099 = .451$

$\beta_2 = (-.33).5 + 1.1(.3) = -.165 + .33 = .165$

2. Using the formulas for variance and covariance, demonstrate that variance is a special case of covariance.

3. Given $r_{1y} = .3 \quad r_{2y} = .4 \quad r_{12} = 0 \quad N = 33$

a) Calculate R^2

$R^2 = (.3)^2 + (.4)^2 = .09 + .16 = .25$

b) Test its significance

$F(2,30) = \frac{.25/2}{.75/30} = 4.95.$

4. What correlation coeficient is used in each of the following situations (list its name only)?

a) two continuous variables Pearson.

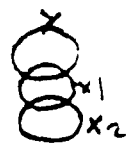
b) one true dichotomy, one continuous r_{pb}

c) one arbitray (artificial) dichotomy, one continuous r_{bis}

d) two true dichotomous variables r_ϕ

e) two arbitrary dichotomous variables r_{tet}

5. Draw a Venn diagram (i.e., a circle represents total, unit, variance) for a classical suppressor variable (X2), another IV (X1) and the DV, Y.



*** NOTE: Answer questions 6-8 as a block; do not separate.

6. How is control of secondary variance achieved in experimental ("laboratory") research?

7. In nonexperimental ("field") research, control of secondary variance is accomplished by means of partialing or semipartialing. Describe verbally (without formulas or diagrams) what is accomplished by a) partial correlation and b) semipartial correlation. That is, why would they be used? Give an example of each use.

8. a) Show, by means of a Venn diagram, how the semipartial correlation controls for secondary variance.

b) Demonstrate, using the multiple correlation coeficient, how to control for X1 and X2 in the set of variables Y, X1-X5. That is, how would you determine the proportion of Y variance that is unique to the combination of X3, X4, X5?

$R^2_{y.1234} - R^2_{y.12}$

9. a) Under what circumstances would you elect to use a stepwise regression technique? Prediction only (hypothesis is not necessary here since RSQUARE)

b) Describe, in words, the procedure for selecting variables in the following variable selection techniques?

1) forward
2) backward
3) stepwise