

Technical University of Denmark

Written examination, date: 9. December 2014

Page 1 of 16 pages Enclosure: 12 pages

Course name: Multivariate Statistics

Course number: 02409

Aids allowed: All

Exam duration: 4 hours

Weighting: The questions are given equal weight

This exam is answered by:

(name)

(signature)

(study no.)

There is a total of 30 questions for the 6 problems. The answers to the 30 questions must be written into the table below.

Problem	1	1	1	1	1	1	2	2	2	2
Question	1.1	1.2	1.3	1.4	1.5	1.6	2.1	2.2	2.3	2.4
Answer										

Problem	2	3	3	3	3	3	4	4	4	4
Question	2.5	3.1	3.2	3.3	3.4	3.5	4.1	4.2	4.3	4.4
Answer										

Problem	5	5	5	5	5	6	6	6	6	6
Question	5.1	5.2	5.3	5.4	5.5	6.1	6.2	6.3	6.4	6.5
Answer										

The possible answers for each question are numbered from 1 to 6. If you enter a wrong number, you may correct it by crossing the wrong number in the table and writing the correct answer immediately below. If there is any doubt about the meaning of a correction then the question will be considered not answered.

Only the front page must be returned. The front page must be returned even if you do not answer any of the questions or if you leave the exam prematurely. Drafts and/or comments are not considered, only the numbers entered above are registered.

A correct answer gives 5 points, a wrong answer gives – 1 point. Unanswered questions or a 6 (corresponding to “don’t know”) give 0 points. The total number of points needed for a satisfactorily answered exam is determined at the final evaluation of the exam. Especially note that the grade 10 may be given even if only one answer is wrong or unanswered.

Remember to write your name, signature, and study number on the front page.

Please note, that there is one and only one correct answer to each question. Furthermore, some of the possible alternative answers may not make sense. When the text refers to SAS-output the values may be rounded to fewer decimal places than in the output itself. The enclosures do not necessarily contain all the output generated by the given SAS programs. Please check that all pages of the exam paper and the enclosures are present.

Problem 1

We consider the model

$$\begin{bmatrix} X_1 & Y_1 & Z_1 \\ X_2 & Y_2 & Z_2 \\ X_3 & Y_3 & Z_3 \\ X_4 & Y_4 & Z_4 \\ X_5 & Y_5 & Z_5 \end{bmatrix} = \begin{bmatrix} 1 & -2 & 2 \\ 1 & -1 & -1 \\ 1 & 0 & -2 \\ 1 & 1 & -1 \\ 1 & 2 & 2 \end{bmatrix} \begin{bmatrix} \alpha_x & \alpha_y & \alpha_z \\ \beta_x & \beta_y & \beta_z \\ \gamma_x & \gamma_y & \gamma_z \end{bmatrix} + \begin{bmatrix} \varepsilon_1 & \delta_1 & \varphi_1 \\ \varepsilon_2 & \delta_2 & \varphi_2 \\ \varepsilon_3 & \delta_3 & \varphi_3 \\ \varepsilon_4 & \delta_4 & \varphi_4 \\ \varepsilon_5 & \delta_5 & \varphi_5 \end{bmatrix}$$

Where the error terms $[\varepsilon_i \quad \delta_i \quad \varphi_i]', i = 1,2,3,4,5$, are independent and normally distributed $N_3(\mathbf{0}, \Sigma)$, and where Σ is the unknown dispersion matrix

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} & \sigma_{xz} \\ \sigma_{xy} & \sigma_y^2 & \sigma_{yz} \\ \sigma_{xz} & \sigma_{yz} & \sigma_z^2 \end{bmatrix}$$

We assume that we obtained the following observations

$$\begin{bmatrix} 2 & 4 & 6 \\ 1 & 4 & 5 \\ 3 & 2 & 2 \\ 4 & 3 & 5 \\ 0 & 2 & 2 \end{bmatrix}$$

With the usual notation we have

$$\mathbf{x}'\mathbf{x} = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 14 \end{bmatrix}$$

The problem continues on the next page

Question 1.1.

The maximum likelihood estimator for α_x becomes

- 1 1
- 2 2
- 3 3
- 4 4
- 5 5
- 6 Don't know

Question 1.2.

The covariance between the maximum likelihood estimators for α_x and β_x becomes

- 1 σ_{xy}
- 2 0
- 3 $\frac{1}{10}$
- 4 σ_{xz}
- 5 $\frac{1}{5}$
- 6 Don't know

Question 1.3.

The covariance between the maximum likelihood estimators for α_x and α_y becomes

- 1 $\frac{1}{10}$
- 2 0
- 3 $\frac{1}{5}\sigma_{xy}$
- 4 $\frac{1}{10}\sigma_{xy}$
- 5 $\frac{1}{5}\sigma_{xz}$
- 6 Don't know

The problem continues on the next page

Question 1.4.

We now want to test the hypothesis

$$H_0 : \begin{bmatrix} \beta_y & \beta_z \\ \gamma_y & \gamma_z \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

against all alternatives. This hypothesis may also be written

$$H_0: \mathbf{A} \begin{bmatrix} \alpha_x & \alpha_y & \alpha_z \\ \beta_x & \beta_y & \beta_z \\ \gamma_x & \gamma_y & \gamma_z \end{bmatrix} \mathbf{B}' = \mathbf{C} \text{ against } H_1: \mathbf{A} \begin{bmatrix} \alpha_x & \alpha_y & \alpha_z \\ \beta_x & \beta_y & \beta_z \\ \gamma_x & \gamma_y & \gamma_z \end{bmatrix} \mathbf{B}' \neq \mathbf{C}$$

Here the matrix \mathbf{A} is:

1 $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$

2 $\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$

3 $[0 \ 1 \ 1]$

4 $\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

5 $[0 \ 1 \ 0]$

6 Don't know.

Question 1.5.

The matrix \mathbf{B} is:

1 $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

2 $\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

3 $\begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$

4 $[0 \ 1 \ 0]$

5 $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$

6 Don't know.

The problem continues on the next page

Question 1.6.

If the hypothesis H_0 is true then the distribution of the usual test statistic is:

- 1 $U(2, 2, 1)$
- 2 $U(1, 1, 2)$
- 3 $U(1, 2, 1)$
- 4 $U(2, 2, 2)$
- 5 $U(2, 2, 5)$
- 6 Don't know.

Problem 2

We consider a normally distributed random variable

$$\begin{bmatrix} X \\ Y \end{bmatrix}$$

that represents a measurement of a two-dimensional property on subjects that belong to one of two populations π_1 or π_2 . The dispersion matrix is equal to

$$\Sigma = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$$

The mean value depends on which population the subject belongs to:

$$E \left(\begin{bmatrix} X \\ Y \end{bmatrix} \right) = \begin{cases} \boldsymbol{\mu}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} & \text{if } \pi_1 \text{ is true} \\ \boldsymbol{\mu}_2 = \begin{bmatrix} 4 \\ 4 \end{bmatrix} & \text{if } \pi_2 \text{ is true} \end{cases}$$

We finally assume that the prior probabilities of belonging to either population is 0.5. We immediately obtain

$$\Sigma^{-1} = \begin{bmatrix} \frac{4}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{4}{3} \end{bmatrix}$$

$$\boldsymbol{\mu}'_1 \Sigma^{-1} \boldsymbol{\mu}_1 = \frac{4}{3} \quad , \quad \boldsymbol{\mu}'_2 \Sigma^{-1} \boldsymbol{\mu}_2 = \frac{64}{3} \quad , \quad (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = 12$$

The problem continues on the next page

Question 2.1.

The linear discriminant function Z for distinguishing between the two populations is:

- 1 $-2X + Y$
- 2 $-2X - 2Y + 10$
- 3 $X + Y - 12$
- 4 $X - 12$
- 5 $-2X - 2Y$
- 6 Don't know

Question 2.2.

If π_2 is true then the mean of Z is:

- 1 -6
- 2 -3
- 3 0
- 4 3
- 5 6
- 6 Don't know.

Question 2.3.

The variance of Z is:

- 1 6
- 2 $\frac{64}{3}$
- 3 36
- 4 12
- 5 $\frac{4}{3} \times \frac{64}{3}$
- 6 Don't know.

The problem continues on the next page

Question 2.4.

Let $\Phi(x)$ be the cumulative distribution function for a $N(0,1)$ -distributed random variable. If π_2 is true then the probability of misclassification is :

- 1 0.05
- 2 $\Phi(12)$
- 3 $\Phi(\sqrt{3})$
- 4 $1 - \Phi(\sqrt{3})$
- 5 0.95
- 6 Don't know.

Question 2.5.

If the prior probability for π_2 is 0.9 (and 0.1 for π_1) the the above probability of misclassification becomes:

- 1 0.023
- 2 $1 - \Phi(2.37)$
- 3 $\Phi(2.37)$
- 4 $\Phi(10.2)$
- 5 0.0000
- 6 Don't know.

Problem 3

Enclosure A belongs to this problem. The data are the first 20 records of gene expression data taken from www.biostat.jhsph.edu/~ririzarr/Teaching/649/. It is not necessary to be familiar with gene expression data in order to solve the problems. The variables are called X and Y and we wish to see how well $\ln(Y)$ may be predicted based on $\ln(X)$.

We do a regression analysis with $\ln(Y) = \ln y$ as dependent variable and $\ln(X) = \ln x$ as independent variable.

The problem continues on the next page

Question 3.1.

The 95% confidence interval for the coefficient to $\ln x$ is:

- 1 $[0.96429 - t(18)_{0.975} \times 0.03336, 0.96429 + t(18)_{0.975} \times 0.03336]$
- 2 $[0.96429 - t(18)_{0.975} \times \sqrt{0.01014}, 0.96429 + t(18)_{0.975} \times \sqrt{0.01014}]$
- 3 $[0.94929 - t(18)_{0.975} \times 0.01014, 0.94929 + t(18)_{0.975} \times 0.01014]$
- 4 $[0.96429 - t(18)_{0.975} \times \sqrt{0.03336}, 0.96429 + t(18)_{0.975} \times \sqrt{0.03336}]$
- 5 $[0.94929 - 1.96 \times 0.1007, 0.94929 + 1.96 \times 0.1007]$
- 6 Don't know.

Question 3.2.

The length of the 95% prediction interval for the last observation is:

- 1 $2 \times \sqrt{0.0230^2 + 0.01014}$
- 2 $2 \times t(18)_{0.975} \times \sqrt{0.0230^2 + 1}$
- 3 $2 \times t(18)_{0.975} \times \sqrt{0.0230 + 0.01014}$
- 4 $2 \times t(18)_{0.975} \times \sqrt{0.0230^2 + 0.01014}$
- 5 $2 \times t(18)_{0.975} \times \sqrt{0.01014}$
- 6 Don't know.

Question 3.3.

We now remove each of the observations one at a time. The observation that - when removed- will cause the numerically least change in its predicted value is no:

- 1 2
- 2 6
- 3 10
- 4 15
- 5 20
- 6 Don't know.

Question 3.4.

The number of observations that have as well an extreme RStudent residual as an extreme (i.e. large) leverage is:

- 1 0
- 2 1
- 3 2
- 4 3
- 5 4
- 6 Don't know.

Question 3.5.

For which observation do we have the worst prediction of lny?

- 1 2
- 2 6
- 3 9
- 4 14
- 5 18
- 6 Don't know.

Problem 4

Enclosure B belongs to this problem. The data are measurements of the average number of miles per gallon gasoline a number of cars could drive under controlled circumstances. The original data source is R.M. Heavenerich, J.D. Murrell, and K.H. Hellman, Light Duty Automotive Technology and Fuel Economy Trends Through 1991, U.S. Environmental Protection Agency, 1991 (EPA/AA/CTAB/91-02). The number of cases is 82 and the variables measured were:

1. vol: Cubic feet of cabin space
2. hp: Engine horsepower
3. mpg: Average miles per gallon
4. sp: Top speed (mph)
5. wt: Vehicle weight (100 lb)

The problem continues on the next page

The fuel consumption (mpg) is the independent variable, and in order to improve linearity we have taken the logarithm of mpg. Furthermore we have added the logarithm and the square of the independent variables to the data set. Firstly we consider the model that contains ‘vol lnhp sp wt’.

Question 4.1.

The diagnostic plots show that one might consider adding two quadratic terms to the model. The variables whose squares should be added are:

- 1 (vol, lnhp)
- 2 (vol, sp)
- 3 (vol, wt)
- 4 (lnhp, sp)
- 5 (lnhp, wt)
- 6 Don't know.

Question 4.2.

How many residuals (RStudent) fall outside the usual [-2, 2] interval:

- 1 1
- 2 5
- 3 9
- 4 13
- 5 17
- 6 Don't know.

The problem continues on the next page

Question 4.3.

We now consider the model that contains ‘wt lnhp sqsp sqhp’ and we want to make a single test in order to see whether we may assume that the coefficients to sqsp and sqhp can be assumed to be simultaneously equal to 0. The usual test statistic is:

1 $\frac{(0.74641 - 0.44873)/1}{0.44873/79}$

2 $\frac{(0.56239 - 0.44873)/2}{0.44873/79}$

3 $\frac{(0.56239 - 0.41289)/2}{0.44873/77}$

4 $\frac{(0.56239 - 0.74641)/2}{0.44873/77}$

5 $\frac{(0.56239 - 0.44873)/2}{0.44873/77}$

6 Don't know.

Question 4.4.

The distribution under the null hypothesis of the statistic above is:

1 t(79)

2 F(1, 79)

3 F(2, 79)

4 U(2, 2, 79)

5 F(2, 77)

6 Don't know.

Problem 5

Enclosure C with SAS program and SAS output belongs to this problem. The measurements are estimated correlations between 176 observed values of different properties of offspring of alfalfa (lucerne). The data is described in Example 5.4, p. 238 in the course book Ersbøll & Conradsen (2012). The variables measured are also given in the table below.

The problem continues on the next page

Variable no. & name	Unit of measure	Explanation
x1: Type of growth	Grade 1 - 9	1 = growth is lying down, 9 = growth is upright
x2: Regrowth after winter	Grade 1 - 9	1 = worst, 9 = best
x3: Ability to creep	Grade 1 - 9	1 = no runners, 9 = most runners
x4: Activity	Grade 1 - 9	1 = weakest, 9 = strongest
x5: Time of blooming	Grade 1 - 9	1 = latest blooming, 9 = earliest blooming
x6: Plant height	cm	
x7: Seed weight	g per plant	
x8: Plant weight	g per plant after drying	
x9: Percent seed	%	Calculated per plant by means of (7) and (8)

Question 5.1.

How many principal components must we include if we want to explain at least 80% of the total variation ?

1 2

2 3

3 4

4 5

5 6

6 Don't know

Question 5.2.

The degrees of freedom for the usual test statistic for testing the hypothesis that the smallest 3 eigenvalues are equal against all alternatives is:

1 5

2 10

3 15

4 20

5 25

6 Don't know.

The problem continues on the next page

Question 5.3.

We now consider a factor analysis with three factors of the data considered above. Consider the following statements on a plant having a large value of an arbitrary factor

- A. A small plant with an upright growth.
- B. A plant that immediately looks healthy: Early blooming, good height and weight, lot of seeds - absolute and relative.
- C. A plant with very little seed.
- D. A plant with good 'dynamic' properties: A heavy plant that grows fine after winter, that has many runners and shows good activity.
- E. A plant that irrespective of whether it is upright or lying down looks good and has good 'dynamic' properties.

For (VARIMAX rotated factor 1, VARIMAX rotated factors 2, VARIMAX rotated factors 3) the following characterization is adequate:

- 1 (E, B, D)
- 2 (D, C, A)
- 3 (A, B, C)
- 4 (A, D, C)
- 5 (B, D, A)
- 6 Don't know.

Question 5.4.

What fraction of the total variance will be explained by the first VARIMAX rotated factor?

- 1 $0.17187 + 0.13527 + 0.79486$
- 2 0.3475
- 3 0.679644^2
- 4 $2.7197/9$
- 5 $2.5733320^2 + 1.3926576^2$
- 6 Don't know.

The problem continues on the next page

Question 5.5.

What fraction of the variation of the first variable x_1 is explained by the third VARIMAX rotated factor?

- 1 0.79486
- 2 0.679644
- 3 0.679644^2
- 4 0.79486^2
- 5 $1.3926576/9$
- 6 Don't know.

Problem 6

Enclosure D belongs to this problem. The data are due to Littell, Freund and Spector (1991). They are measurements on the effect of three different weightlifting programs. There were taken measurements at 7 different time points, the same for the three treatments. They were

- RI: The number of repetitions of weightlifting was increased as subjects became stronger.
- WI: The amount of weight was increased as subjects became stronger.
- CONT: Control group with no training.

Since there may be large individual differences between the strength of the subjects participating in the study we have normalized the last two measurements by dividing them with the first measurement yielding the new variables RS6 and RS7. We shall only consider these relative measures of the strength of the subjects.

Question 6.1.

We now want to compare the three treatments with a multivariate analysis of variance of the relative strengths at times 6 and 7. Then the null hypothesis distribution of the usual test statistic Wilk's Lambda is:

- 1 $U(2, 1, 55)$
- 2 $U(2, 2, 54)$
- 3 $U(2, 3, 56)$
- 4 $F(2, 54)$
- 5 $F(3, 56)$
- 6 Don't know.

The problem continues on the next page

Question 6.2.

Under the usual assumptions the estimate for the variance of RS6 is equal to:

1 $\sqrt{0.0059531967/2}$

2 $0.0059531967/2$

3 0.0349065

4 $0.0349065/54$

5 $\sqrt{0.0349065/54}$

6 Don't know.

Question 6.3.

We now only consider the active treatment groups WI and RI, and we want to test whether the means of $[RS6 \quad RS7]'$ are the same for the two treatment groups. The usual test statistic becomes:

1 $\frac{16+21-2-1}{2(16+21-2)} \frac{16 \times 21}{16+21} 0.16700$

2 0.0068386375

3 $\frac{16+21-2-1}{2(16+21-2)} \frac{16 \cdot 21}{16+21} 0.14851$

4 $\frac{57-2}{56 \times 2} 0.14851^2$

5 $\frac{16+21-2-1}{2(16+21-2)} \frac{16 \times 21}{16+21} 0.81667748^2$

6 Don't know.

Question 6.4.

The distribution under the null hypothesis of the test statistic above is:

1 $t(54)$

2 $F(1, 54)$

3 $F(2, 34)$

4 $U(2, 2, 34)$

5 $F(2, 54)$

6 Don't know.

Question 6.5.

We are now interested in using the values of $[RS6 \quad RS7]'$ in distinguishing between the two training programs with a decision function of the form $[RS6 \quad RS7]' \begin{bmatrix} a \\ b \end{bmatrix} > c$. Rounded to integers the values of a, b and c should be chosen as

1 27, -32, -6

2 21, 28, 33

3 24, -4, 21

4 20, 18, 8

5 16, 12, 8

6 Don't know.

Enclosure A – SAS program

Page 1 of 6

```
/*The gene dataset consists of the first 20 observations from the sasuser.gene  
dataset taken from (http://www.biostat.jhsph.edu/~ririzarr/Teaching/649)*/;
```

```
data gene;  
set sasuser.gene;  
if _n_<21;  
run;  
data gene;  
set gene;  
lny = log(y);  
lnx = log(x);  
run;  
  
Title 'The observations in the gene expression dataset';  
proc print data=gene; run;  
  
ods graphics on;  
  
proc reg data=gene plots(label)=all;  
model lny=lnx/r influence;  
run;  
  
ods graphics off;
```

Enclosure A – SAS Output

Page 2 of 6

The observations in the gene expression dataset

Obs	y	x	lny	lnx
1	1012.92	403.10	6.92060	5.99917
2	1777.01	888.12	7.48269	6.78911
3	1838.03	1015.10	7.51645	6.92274
4	1878.87	1039.59	7.53843	6.94658
5	2481.68	1352.98	7.81669	7.21007
6	718.38	375.99	6.57700	5.92957
7	631.74	334.46	6.44848	5.81251
8	801.87	418.60	6.68694	6.03692
9	735.21	287.91	6.60016	5.66266
10	1154.28	547.40	7.05124	6.30517
11	876.57	400.69	6.77602	5.99319
12	1764.01	794.54	7.47535	6.67776
13	2068.58	915.77	7.63462	6.81976
14	9284.50	4295.62	9.13610	8.36535
15	2157.77	1100.30	7.67683	7.00334
16	4633.86	2245.81	8.44114	7.71682
17	1647.83	878.70	7.40722	6.77845
18	2891.35	1549.99	7.96948	7.34600
19	858.48	439.49	6.75517	6.08562
20	1838.49	895.51	7.51670	6.79739

Enclosure A – SAS Output

The REG Procedure
Model: MODEL1
Dependent Variable: Inv

Number of Observations Read	20
Number of Observations Used	20

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	8.47517	8.47517	835.78	<.0001
Error	18	0.18253	0.01014		
Corrected Total	19	8.65770			

Root MSE	0.10070	R-Square	0.9789
Dependent Mean	7.37136	Adj R-Sq	0.9777
Coeff Var	1.36610		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	0.94929	0.22328	4.25	0.0005
lnx	1	0.96429	0.03336	28.91	<.0001

Enclosure A – SAS Output

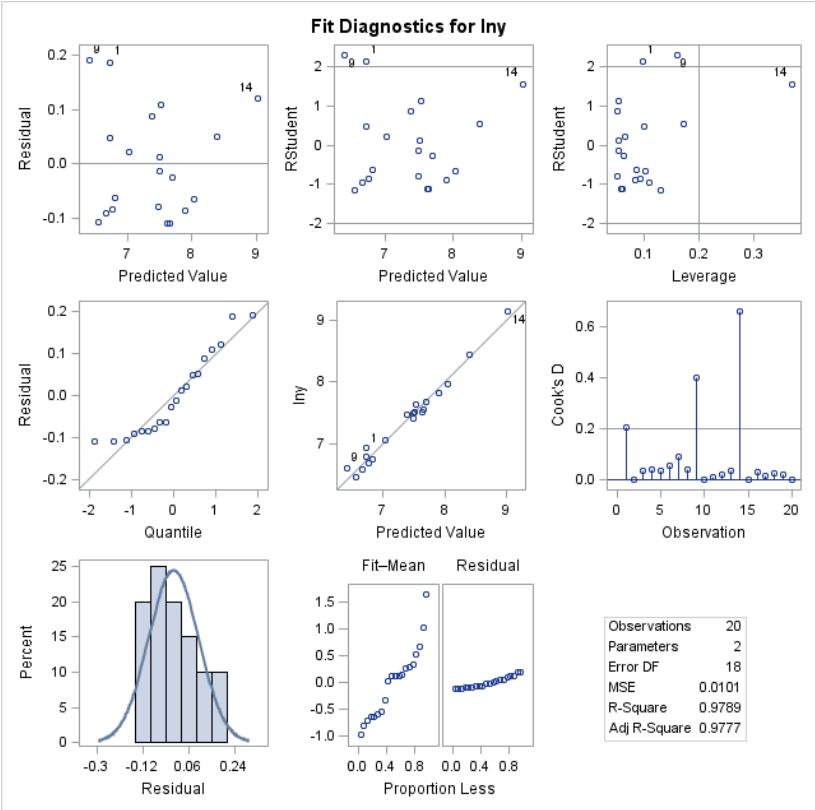
The REG Procedure
Model: MODEL1
Dependent Variable: Inv

Output Statistics										
Obs	Dep. Variable	Pred. Value	Std Err. Mean Predict	Residual	Std Err. Residual	Student Residual	-2-1 0 1 2	Cook's D	Hat Diag H	DFFITS
1	6.9206	6.7342	0.0315	0.1864	0.0956	1.949	***	0.206	0.0979	0.7023
2	7.4827	7.4960	0.0229	-0.0133	0.0981	-0.135		0.001	0.0518	-0.0308
3	7.5165	7.6248	0.0242	-0.1084	0.0978	-1.108	**	0.038	0.0576	-0.2758
4	7.5384	7.6478	0.0245	-0.1094	0.0977	-1.120	**	0.039	0.0590	-0.2825
5	7.8167	7.9019	0.0290	-0.0852	0.0964	-0.883	*	0.035	0.0832	-0.2645
6	6.5770	6.6671	0.0332	-0.0901	0.0951	-0.948	*	0.055	0.1085	-0.3297
7	6.4485	6.5542	0.0361	-0.1057	0.0940	-1.125	**	0.094	0.1288	-0.4360
8	6.6869	6.7706	0.0306	-0.0837	0.0959	-0.872	*	0.039	0.0926	-0.2767
9	6.6002	6.4097	0.0402	0.1904	0.0923	2.062	****	0.402	0.1591	0.9975
10	7.0512	7.0293	0.0254	0.0219	0.0974	0.225		0.002	0.0638	0.0572
11	6.7760	6.7285	0.0316	0.0476	0.0956	0.498		0.014	0.0988	0.1612
12	7.4753	7.3886	0.0225	0.0868	0.0981	0.884	*	0.021	0.0500	0.2016
13	7.6346	7.5255	0.0231	0.1091	0.0980	1.113	**	0.035	0.0528	0.2647
14	9.1361	9.0159	0.0612	0.1202	0.0800	1.503	***	0.661	0.3691	1.1945
15	7.6768	7.7025	0.0253	-0.0257	0.0975	-0.264		0.002	0.0629	-0.0665
16	8.4411	8.3905	0.0418	0.0506	0.0916	0.553	*	0.032	0.1726	0.2473
17	7.4072	7.4857	0.0229	-0.0785	0.0981	-0.800	*	0.017	0.0515	-0.1845
18	7.9695	8.0330	0.0321	-0.0635	0.0954	-0.665	*	0.025	0.1016	-0.2201
19	6.7552	6.8176	0.0296	-0.0624	0.0963	-0.648	*	0.020	0.0862	-0.1958
20	7.5167	7.5039	0.0230	0.0128	0.0980	0.130		0.000	0.0521	0.0297

Sum of Residuals	0
Sum of Squared Residuals	0.18253
Predicted Residual SS (PRESS)	0.24540

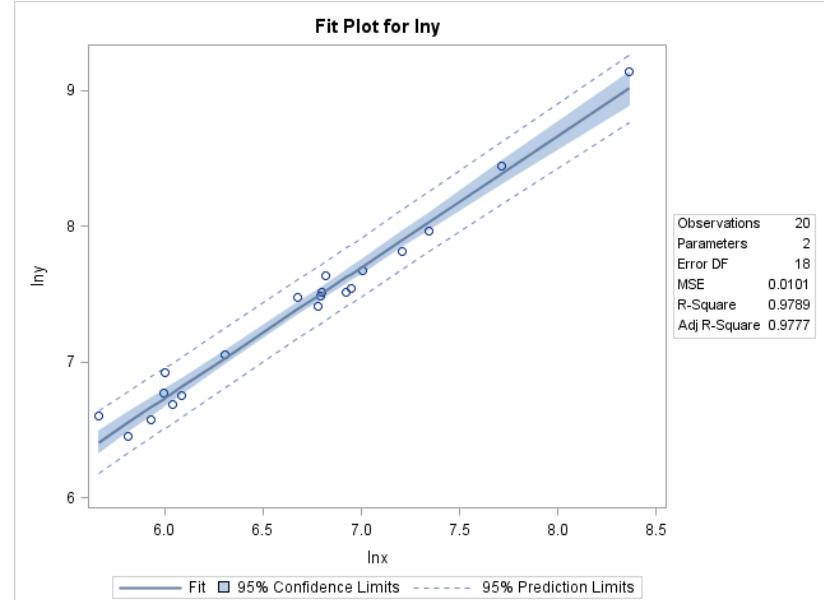
Enclosure A – SAS Output

Page 5 of 6



Enclosure A – SAS Output

Page 6 of 6



Enclosure B – SAS program

Page 1 of 8

/*Passenger Car Mileage data from R.M. Heavenrich, J.D. Murrell, and K.H. Hellman: Light Duty Automotive Technology and Fuel Economy Trends Through 1991, U.S. Environmental Protection Agency, 1991 (EPA/AA/CTAB/91-02). Data available at (<http://lib.stat.cmu.edu/DASL/Datafiles/carmpgdat.html>)*/;

```
data car;  
set sasuser.car;  
lnmpg=log(mpg); lnvol=log(vol); lnhp=log(hp); lnspl=log(sp);  
lnwt=log(wt);  
sqvol=vol*vol; sqhp=hp*hp; sqsp=sp*sp; sqwt=wt*wt;  
run;
```

Title 'The first 20 observations in the car mileage dataset with squares and logarithms added';

```
proc print data=car (obs=20);  
run;
```

ods graphics on;

Title 'Model lnmpg = vol lnhp sp wt';
proc reg plots(label)=all;
model lnmpg = vol lnhp sp wt;
run;

Title 'Stepwise regression on all variables';

```
proc reg plots(label)=all;  
model lnmpg = vol lnvol sqvol hp lnhp sqhp sp lnspl sqsp wt lnwt sqwt /  
r selection=stepwise;  
run;
```

ods graphics off;

Enclosure B – SAS Output

Page 2 of 8

The first 20 observations in the car mileage dataset without squares and logarithms added

Obs	make	mpg	vol	hp	sp	wt
1	GM/GeoMe	65.4	89	49	96	17.5
2	GM/GeoMe	56.0	92	55	97	20.0
3	GM/GeoMe	55.9	92	55	97	20.0
4	SuzukiSw	49.0	92	70	105	20.0
5	Daihatsu	46.5	92	53	96	20.0
6	GM/GeoSp	46.2	89	70	105	20.0
7	GM/GeoSp	45.4	92	55	97	20.0
8	HondaCiv	59.2	50	62	98	22.5
9	HondaCiv	53.3	50	62	98	22.5
10	Daihatsu	43.4	94	80	107	22.5
11	SubaruJu	41.1	89	73	103	22.5
12	HondaCiv	40.9	50	92	113	22.5
13	HondaCiv	40.9	99	92	113	22.5
14	SubaruJu	40.4	89	73	103	22.5
15	SubaruJu	39.6	89	66	100	22.5
16	SubaruJu	39.3	89	73	103	22.5
17	ToyotaTe	38.9	91	78	106	22.5
18	HondaCiv	38.8	50	92	113	22.5
19	ToyotaTe	38.2	91	78	106	22.5
20	FordEsco	42.2	103	90	109	25.0

Enclosure B – SAS Output

Model lnmpg= vol lnhp sp wt

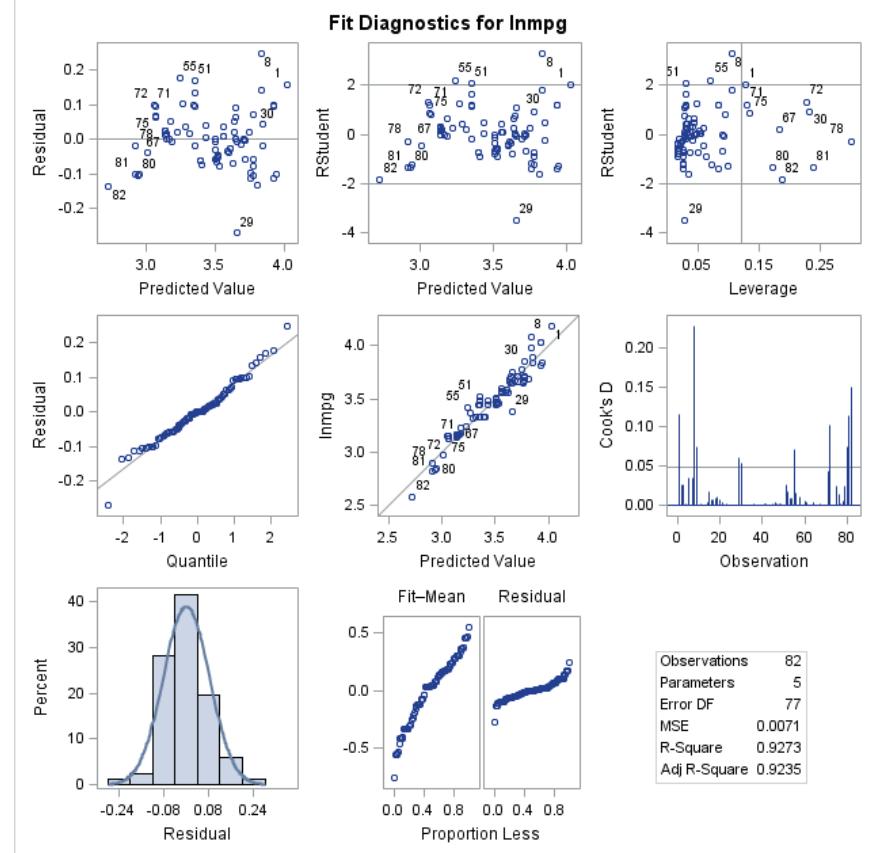
The REG Procedure
Model: MODEL1
Dependent Variable: lnmpg

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	6.95846	1.73961	245.58	<.0001
Error	77	0.54544	0.00708		
Corrected Total	81	7.50390			

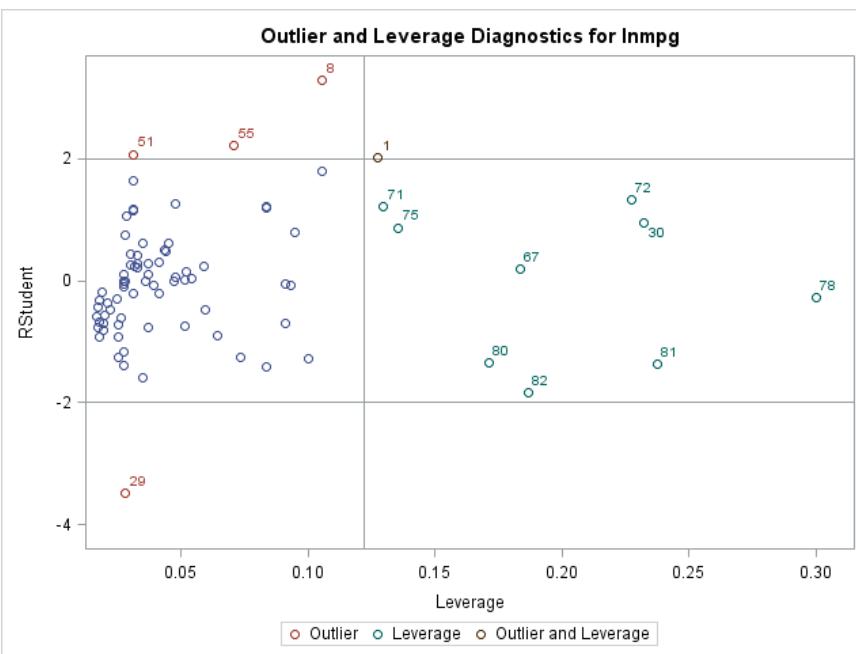
Root MSE	0.08416	R-Square	0.9273
Dependent Mean	3.47571	Adj R-Sq	0.9235
Coeff Var	2.42151		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	5.75631	0.29230	19.69	<.0001
vol	1	-0.00012456	0.00052666	-0.24	0.8137
lnhp	1	-0.45030	0.14998	-3.00	0.0036
sp	1	0.00388	0.00292	1.33	0.1879
wt	1	-0.01943	0.00396	-4.91	<.0001

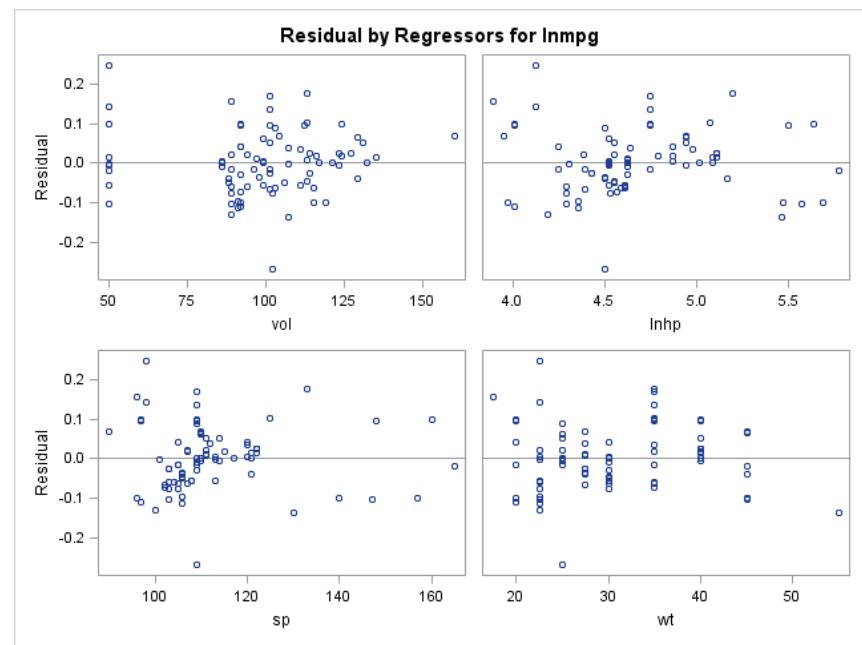
Enclosure B – SAS Output



Enclosure B – SAS Output



Enclosure B – SAS Output



Enclosure B – SAS Output

Stepwise regression on all variables

The REG Procedure
Model: MODEL1
Dependent Variable: Inmpg

Stepwise Selection: Step 1

Variable wt Entered: R-Square = 0.9005 and C(p) = 57.1574

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6.75750	6.75750	724.27	<.0001
Error	80	0.74641	0.00933		
Corrected Total	81	7.50390			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	4.57248	0.04213	109.92158	11781.4	<.0001
wt	-0.03548	0.00132	6.75750	724.27	<.0001

Bounds on condition number: 1, 1

Stepwise Selection: Step 2

Variable Inhp Entered: R-Square = 0.9251 and C(p) = 25.8365

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	6.94151	3.47076	487.54	<.0001
Error	79	0.56239	0.00712		
Corrected Total	81	7.50390			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	5.38578	0.16415	7.66396	1076.56	<.0001
Inhp	-0.24895	0.04897	0.18401	25.85	<.0001
wt	-0.02417	0.00250	0.66353	93.21	<.0001

Bounds on condition number: 4.7282, 18.913

Enclosure B – SAS Output

Stepwise Selection: Step 9

Variable Inhp Removed: R-Square = 0.9450 and C(p) = 4.7654

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	7.09101	1.41820	261.04	<.0001
Error	76	0.41289	0.00543		
Corrected Total	81	7.50390			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	16.83379	2.14108	0.33583	61.82	<.0001
hp	0.01770	0.00511	0.06520	12.00	0.0009
sqhp	-0.00004828	0.00001000	0.12668	23.32	<.0001
sp	-0.15797	0.02794	0.17370	31.97	<.0001
sqsp	0.00060887	0.00010033	0.20007	36.83	<.0001
lnwt	-1.37447	0.22678	0.19957	36.73	<.0001

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	wt		1	0.9005	0.9005	57.1574	724.27	<.0001
2	Inhp		2	0.0245	0.9251	25.8365	25.85	<.0001
3	sqsp		3	0.0022	0.9273	24.8135	2.39	0.1265
4	sqhp		4	0.0129	0.9402	9.2295	16.67	0.0001
5		wt	3	0.0016	0.9386	9.3999	2.06	0.1555
6	hp		4	0.0026	0.9412	7.8355	3.44	0.0675
7	sp		5	0.0019	0.9432	7.1937	2.60	0.1109
8	lnwt		6	0.0023	0.9455	6.0503	3.18	0.0784
9		Inhp	5	0.0005	0.9450	4.7654	0.72	0.3974

Enclosure C – SAS Program

```
/*The data comes from Example 5.4 p. 238 in Ersbøll & Conradsen (2012)*;
```

```
data crop(type=corr);
infile cards;
input _type_ $ _name_ $ x1-x9;
cards;
N 176 176 176 176 176 176 176 176 176
CORR x1 1.000 -0.033 0.116 0.018 0.131 -0.207 0.035 -0.087 0.041
CORR x2 -0.033 1.000 0.711 0.515 0.125 0.199 -0.025 0.348 -0.066
CORR x3 0.116 0.711 1.000 0.440 0.022 0.039 -0.133 0.218 -0.157
CORR x4 0.018 0.515 0.440 1.000 0.201 0.517 0.071 0.689 -0.081
CORR x5 0.131 0.125 0.022 0.201 1.000 0.496 0.487 0.168 0.486
CORR x6 -0.207 0.199 0.039 0.517 0.496 1.000 0.453 0.559 0.367
CORR x7 0.035 -0.025 -0.133 0.071 0.487 0.453 1.000 0.360 0.947
CORR x8 -0.087 0.348 0.218 0.689 0.168 0.559 0.360 1.000 0.128
CORR x9 0.041 -0.066 -0.157 -0.081 0.486 0.367 0.947 0.128 1.000
;
run;
```

```
Title 'The crop data';
proc print data=crop;
run;
```

```
ods graphics on;
proc factor
  data=crop
  rotate=varimax
  plots=(scree initloadings(vector) loadings(vector));
run;
ods graphics off;
```

Enclosure C – SAS Output

The crop data

Obs	_type_	_name_	x1	x2	x3	x4	x5	x6	x7	x8	x9
1	N		176	176	176	176	176	176	176	176	176
2	CORR	x1	1.000	-0.033	0.116	0.018	0.131	-0.207	0.035	-0.087	0.041
3	CORR	x2	-0.033	1.000	0.711	0.515	0.125	0.199	-0.025	0.348	-0.066
4	CORR	x3	0.116	0.711	1.000	0.440	0.022	0.039	-0.133	0.218	-0.157
5	CORR	x4	0.018	0.515	0.440	1.000	0.201	0.517	0.071	0.689	-0.081
6	CORR	x5	0.131	0.125	0.022	0.201	1.000	0.496	0.487	0.168	0.486
7	CORR	x6	-0.207	0.199	0.039	0.517	0.496	1.000	0.453	0.559	0.367
8	CORR	x7	0.035	-0.025	-0.133	0.071	0.487	0.453	1.000	0.360	0.947
9	CORR	x8	-0.087	0.348	0.218	0.689	0.168	0.559	0.360	1.000	0.128
10	CORR	x9	0.041	-0.066	-0.157	-0.081	0.486	0.367	0.947	0.128	1.000

The FACTOR Procedure

Input Data Type	Correlations
N Set/Assumed in Data Set	176
N for Significance Tests	176

The FACTOR Procedure
Initial Factor Method: Principal Components

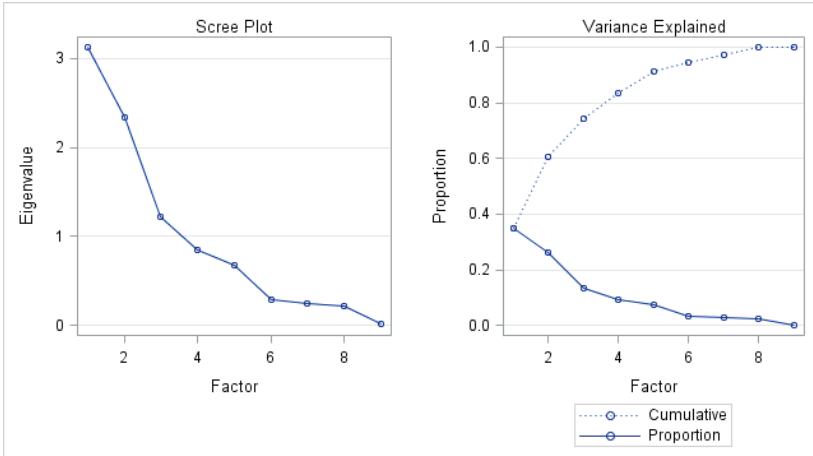
Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 9 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	3.12713262	0.78562655	0.3475	0.3475
2	2.34150607	1.12443569	0.2602	0.6076
3	1.21707038	0.37454816	0.1352	0.7429
4	0.84252222	0.15878292	0.0936	0.8365
5	0.68373929	0.38994087	0.0760	0.9124
6	0.29379842	0.04096571	0.0326	0.9451
7	0.25283270	0.03277307	0.0281	0.9732
8	0.22005963	0.19872097	0.0245	0.9976
9	0.02133867		0.0024	1.0000

3 factors will be retained by the MINEIGEN criterion.

Enclosure C – SAS Output

Page 3 of 6



Factor Pattern			
	Factor1	Factor2	Factor3
x1	-0.03215	-0.02210	0.82348
x2	0.47109	0.66538	0.21979
x3	0.30601	0.71369	0.39707
x4	0.66989	0.55564	-0.11180
x5	0.62336	-0.32613	0.26750
x6	0.79859	-0.07914	-0.33676
x7	0.68698	-0.62850	0.11367
x8	0.73621	0.26529	-0.28829
x9	0.56702	-0.70865	0.19869

Variance Explained by Each Factor		
Factor1	Factor2	Factor3
3.1271326	2.3415061	1.2170704

Final Communality Estimates: Total = 6.685709								
x1	x2	x3	x4	x5	x6	x7	x8	x9
0.679644	0.712957	0.760665	0.769980	0.566497	0.757409	0.879880	0.695504	0.863172

Enclosure C – SAS Output

Page 4 of 6

The FACTOR Procedure
Rotation Method: Varimax

	1	2	3
1	0.74239	0.59769	-0.30269
2	-0.63256	0.77418	-0.02277
3	0.22073	0.20837	0.95282

Rotated Factor Pattern			
	Factor1	Factor2	Factor3
x1	0.17187	0.13527	0.79486
x2	-0.02264	0.84248	0.05167
x3	-0.13662	0.81816	0.26946
x4	0.12117	0.80725	-0.32194
x5	0.72812	0.17583	0.07362
x6	0.56859	0.34586	-0.56079
x7	0.93267	-0.05229	-0.08533
x8	0.31511	0.58534	-0.50358
x9	0.91307	-0.16832	0.03382

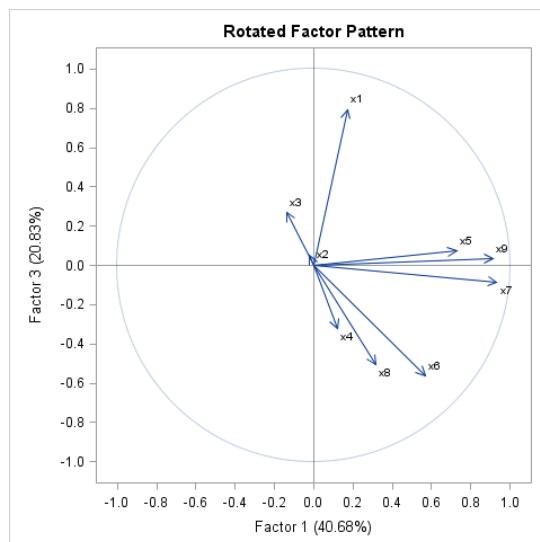
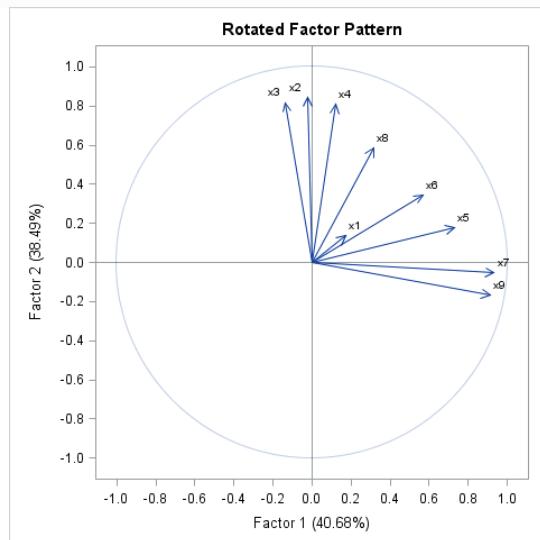
Variance Explained by Each Factor		
Factor1	Factor2	Factor3
2.7197196	2.5733320	1.3926576

Final Communality Estimates: Total = 6.685709								
x1	x2	x3	x4	x5	x6	x7	x8	x9
0.679644	0.712957	0.760665	0.769980	0.566497	0.757409	0.879880	0.695504	0.863172

The FACTOR Procedure
Rotation Method: Varimax

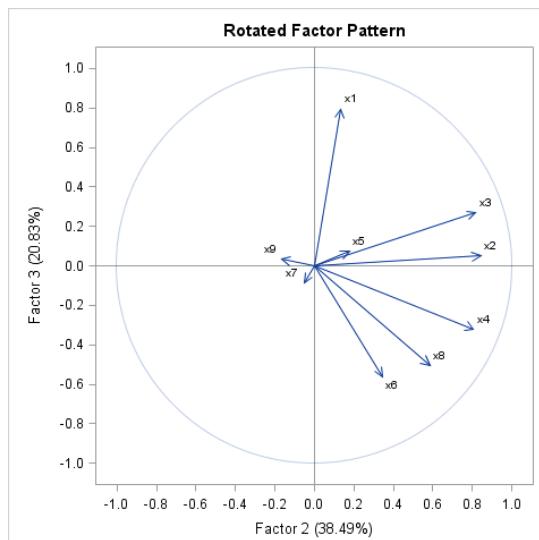
Enclosure C – SAS Output

Page 5 of 6



Enclosure C – SAS Output

Page 6 of 6



Enclosure D – SAS Program

Title 'Strength data from Littell, Freund & Spector (1991);

data weights;

input subject treatment \$ s1 s2 s3 s4 s5 s6 s7;

cards;

1 cont 85 85 86 85 87 86 87

2 ri 79 79 79 80 80 78 79

20 wi 80 81 80 79 80

16 rs 84 85 85 85 83 82

1 wi 84 85 84 83 83 83 84

21 rs 80 81 80 81 81 82 83

;

/*We add two new relative measures rs6 and rs7*/;

data weights;

set weights;

rs6=s6/s1; rs7=s7/s1; **run**;

/*We discard the control group*/;

data weightsriwi;

set weights;

if treatment NE 'cont'; **run**;

Title 'Proc glm on the relative measures for all three treatments';

proc glm data=weights;

class treatment;

model rs6 rs7=treatment;

manova h=_all_ /printe printh;

run;

Title 'Proc discrim on the relative measures for all three treatments';

proc discrim data=weights pool=yes pcov;

class treatment;

var rs6 rs7;

run;

Title 'Proc discrim on the relative measures for the ri and the wi treatment';

proc discrim data=weightsriwi pool=yes pcov;

class treatment;

var rs6 rs7;

run;

Page 1 of 4

Enclosure D – SAS Output

Proc glm on the relative measures for all three treatments

The GLM Procedure
Multivariate Analysis of Variance

E = Error SSCP Matrix		
	rs6	rs7
rs6	0.0349065144	0.0320103992
rs7	0.0320103992	0.0389930628

H = Type III SSCP Matrix for treatment		
	rs6	rs7
rs6	0.0059531967	0.0068386375
rs7	0.0068386375	0.0080267802

Characteristic Roots and Vectors of: E Inverse * H, where H = Type III SSCP Matrix for treatment E = Error SSCP Matrix			
Characteristic Root	Percent	Characteristic Vector V'EV=1	
		rs6	rs7
0.20681714	93.39	0.7631015	4.4249594
0.01463059	6.61	-10.7384604	9.1744414

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall treatment Effect
H = Type III SSCP Matrix for treatment
E = Error SSCP Matrix

S=2 M=.5 N=25.5					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.81667748	2.82	4	106	0.0285
Pillai's Trace	0.18579367	2.77	4	108	0.0311
Hotelling-Lawley Trace	0.22144772	2.92	4	62.575	0.0281
Roy's Greatest Root	0.20681714	5.58	2	54	0.0062

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

NOTE: F Statistic for Wilks' Lambda is exact.

Enclosure D – SAS Output

Proc discrim on the relative measures for all three treatments**The DISCRIM Procedure**

Total Sample Size	57	DF Total	56
Variables	2	DF Within Classes	54
Classes	3	DF Between Classes	2

Number of Observations Read	57
Number of Observations Used	57

Class Level Information					
treatment	Variable Name	Frequency	Weight	Proportion	Prior Probability
cont	cont	20	20.0000	0.350877	0.333333
ri	ri	16	16.0000	0.280702	0.333333
wi	wi	21	21.0000	0.368421	0.333333

Pooled Within-Class Covariance Matrix, DF = 54		
Variable	rs6	rs7
rs6	0.0006464169	0.0005927852
rs7	0.0005927852	0.0007220938

Generalized Squared Distance to treatment			
From treatment	cont	ri	wi
cont	0	0.62023	1.03154
ri	0.62023	0	0.14851
wi	1.03154	0.14851	0

Linear Discriminant Function for treatment			
Variable	cont	ri	wi
Constant	-789.97582	-821.58820	-828.29052
rs6	1119	1141	1120
rs7	463.43389	473.18529	500.89391

Enclosure D – SAS Output

Proc discrim on the relative measures for the ri and the wi treatment**The DISCRIM Procedure**

Total Sample Size	37	DF Total	36
Variables	2	DF Within Classes	35
Classes	2	DF Between Classes	1

Number of Observations Read	37
Number of Observations Used	37

Class Level Information					
treatment	Variable Name	Frequency	Weight	Proportion	Prior Probability
ri	ri	16	16.0000	0.432432	0.500000
wi	wi	21	21.0000	0.567568	0.500000

Pooled Within-Class Covariance Matrix, DF = 35		
Variable	rs6	rs7
rs6	0.0007003206	0.0006643357
rs7	0.0006643357	0.0007811916

Generalized Squared Distance to treatment		
From treatment	ri	wi
ri	0	0.16700
wi	0.16700	0

Linear Discriminant Function for treatment		
Variable	ri	wi
Constant	-748.96607	-754.69084
rs6	1125	1098
rs7	346.41638	378.60181