Lab Assignment No. 2: Answer Key

1) Using the dataset (t5_2.sas7bdat), what is the correlation matrix between all response variables?

Pearson Correlation Coefficients, N = 89

	sshist	verbal	science
sshist	1.00000	0.30025	0.53493
	0.0043	<.0001	
verbal	0.30025	1.00000	0.33390
	0.0043		0.0014
science	0.53493	0.33390	1.00000
	<.0001	0.0014	

Based on the computed correlation matrix all three variables appear to be correlated with one another at the $p < .01$ level. Given that these data are analyzed in the multivariate environment some underlying communality is expected. In fact these three assessments are subcomponents of a larger overall assessment. Therefore significant correlations would be expected.

2) Evaluate univariate normality. Take appropriate steps. Justify your decisions.

Tests for Normality for sshist

Test		--Statistic---	-----p Value------	
Shapiro-Wilk	**W**	**0.987798**	**Pr < W**	**0.5781**
Kolmogorov-Smirnov	D	0.093009	Pr > D	0.0566

For the *sshist* variable we observe a Shapiro-Wilk statistic of $W = .99$, $p > .05$. We therefore conclude that the variable appears to be normally distributed.

Tests for Normality for verbal

Test		--Statistic---	-----p Value------	
Shapiro-Wilk	**W**	**0.653503**	**Pr < W**	**<0.0001**
Kolmogorov-Smirnov	D	0.169537	Pr > D	<0.0100

For the *verbal* variable we observe a Shapiro-Wilk statistic of $W = .65$, $p < .001$. We therefore conclude that the variable appears to be non-normally distributed. Further investigation of the corresponding extreme observations table, box plot,

and stem-and-leaf plot found a single observation (id = 89) that was unrepresentatively larger than the remaining bulk of the data. Because of the extremely unlikely (and in fact impossible) score the observation was deleted from the dataset and excluded from further analyses. The normality test was rerun.

Tests for Normality for verbal

Test		--Statistic---		-----p Value------
Shapiro-Wilk	**W**	**0.967803**	**Pr < W**	**0.0274**
Kolmogorov-Smirnov	D	0.084575	Pr > D	0.1214

The deletion of the extreme outlier improved the variables distribution. However, it did not make it normal. The extreme outlier table was also consulted:

Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
25	50	70	83
28	66	71	63
28	48	72	10
32	17	73	73
35	59	75	51

Screening the five lowest and highest remaining values in the non-normally distributed variable did not provide further evidence against any one observation. Since multivariate normality is not guaranteed by univariate normality the variable was no further altered.

Tests for Normality for science

Test		--Statistic---		-----p Value------
Shapiro-Wilk	**W**	**0.985687**	**Pr < W**	**0.4386**
Kolmogorov-Smirnov	D	0.079914	Pr > D	>0.1500

For the *science* variable we observe a Shapiro-Wilk statistic of $W = .98$, $p > .05$. We therefore conclude that the variable appears to be normally distributed.

3) Evaluate multivariate normality (Leverage Values, Mahalonobis Distances).

Either leverage values (h_{ii})or Mahalonobis distances (*MD*) can be used to investigate multivariate outliers. Since one can be expressed in terms of the other

$$MD = (N-1) \times \left(h_{ii} - \frac{1}{N} \right)$$

it is a personal preference which one is used. For the sake of this assignment both were considered. Leverage values ranges from .0118 to .1151 with a mean of .0341, standard deviation of .0202 and a median of .0294. The Mahalonobis distances ranged from .0403 to a maximum of 9.0250, with a mean of 1.9773, standard deviation of 1.7557 and a median of 1.5741. The *proc iml* derived critical values where as follows.

For Leverage:

$$\frac{2 \times p}{N} = .0682$$

or

$$2 \times \frac{p+1}{N} = .0909$$

or

$$\frac{\chi^2_{(df=p,\ \alpha=.01)}}{N-1} + \frac{1}{N} = .1418$$

For Mahalonobis Distances:

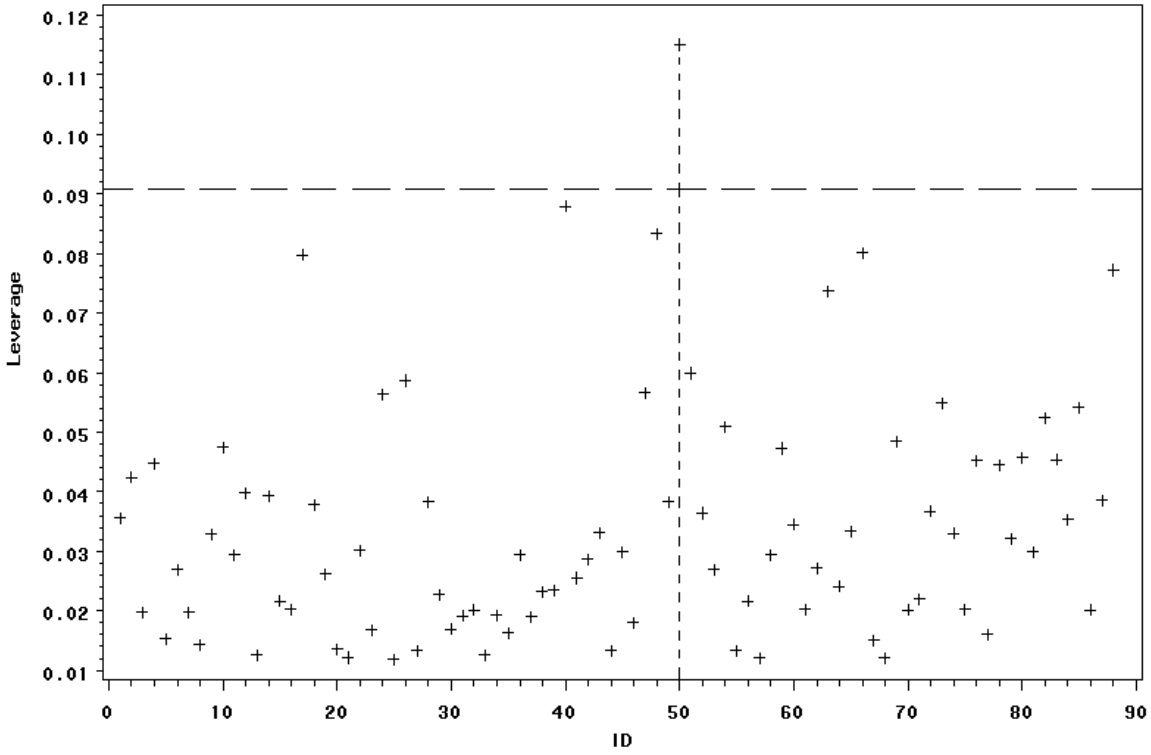
$$\chi^2_{(df=p,\ \alpha=.05)} = 2 \times \Gamma_{(1-\alpha,\ p/2)} = 7.8147$$

or

$$\chi^2_{(df=p,\ \alpha=.01)} = 2 \times \Gamma_{(1-\alpha,\ p/2)} = 11.3449$$

As can be seen from the Figure 1 the distribution of leverage values across the observations in the dataset could be plausibly be regarded as random. Given that there are several methods/several criterions for determining if any one observation's leverage value or Mahalonobis distance should be considered an outlier it is up to the discretion of the analyst to make a decision which cut of rule to use. Figure 1 shows the second/moderate cut of line for leverage values (horizontal line) by which one observation would fall in the rejection region. That observation, case number 50 (as indicated by the projected vertical line) would be considered a multivariate outlier and be deleted from the dataset. Consistent with the authors of our textbook (Raykov & Marcoulides, 2008) not only should we consider the chi-square based cut off region for the leverage values, but also at a more conservative alpha level of .01. In such an event, case number 50 would not be considered an outlier and would be left in the analysis; as was done here.

Figure 1. Leverage values by case order



4) What are the Mardia's coefficients for these data? Interpret. Take appropriate steps. Justify your decisions.

From the SAS macro multnorm Mardia's skeweness and kurtosis coefficients were computed. The macro also provides the univariate tests of normality, which can be used to verify the earlier estimates. In the case of our dataset with 88 observations neither the multivariate skeweness or kurtosis deviate from an expected multivariate normal distribution.

Variable	n	Test	Multivariate Skewness & Kurtosis	Test Statistic Value	p-value
sshist	88	Shapiro-Wilk	.	0.9885	0.63350
verbal	88	Shapiro-Wilk	.	0.9678	0.02745
science	88	Shapiro-Wilk	.	0.9859	0.46178
	88	**Mardia Skewness**	**0.8659**	**13.3573**	**0.20437**
	88	**Mardia Kurtosis**	**13.6712**	**-1.1379**	**0.25514**

Compare these findings against those of the Mardia estimates when the previous univariate outlier would not have been deleted.

Variable	n	Test	Multivariate Skewness & Kurtosis	Test Statistic Value	p-value
sshist	89	Shapiro-Wilk	.	0.988	0.57814
verbal	89	Shapiro-Wilk	.	0.654	0.00000
science	89	Shapiro-Wilk	.	0.986	0.43859
	89	**Mardia Skewness**	**42.7109**	**666.000**	**0.00000**
	89	**Mardia Kurtosis**	**63.2857**	**41.584**	**0.00000**

Also, consider the two chi-square q-q plots. Much like in the univariate analogy we seek the line to be a linear representation of the data's distribution following a hypothetical normal curve. Figure 2 shows the multivariate q-q plot for the dataset with the univariate outlier, Figure 3 for the data with the observation removed.

Figure 2. Chi-square Q-Q plot for multivariate non-normal dataset

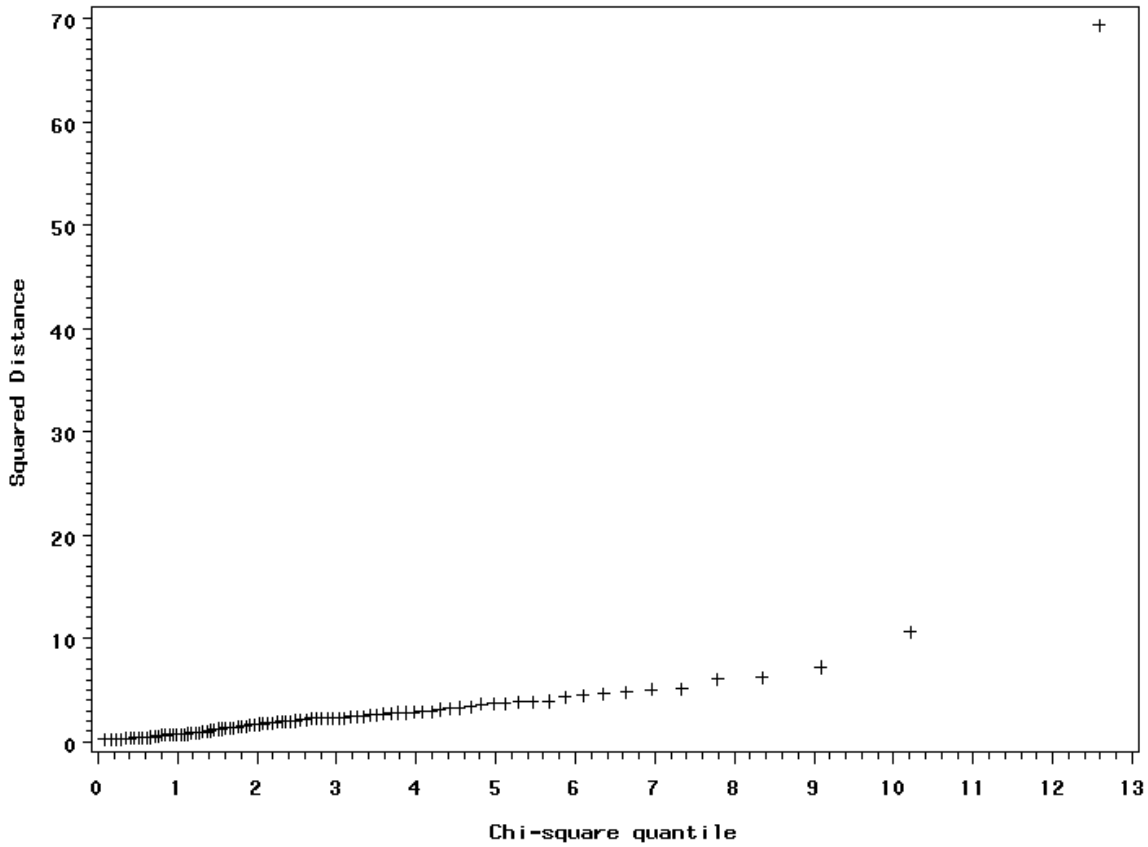
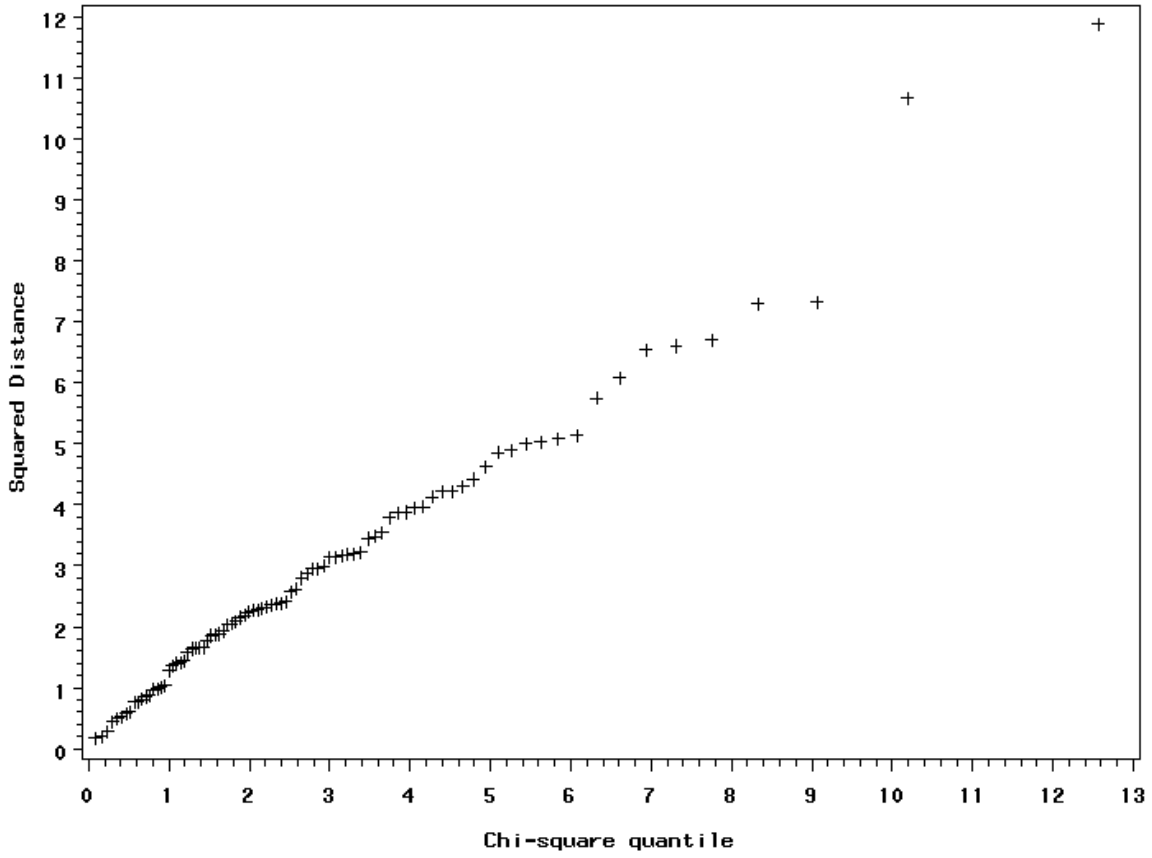


Figure 3. Chi-square Q-Q plot for multivariate normal dataset



5) Conduct univariate one-sample t-tests. What are your conclusions?

Three univariate one-sample t-tests were computed to compare hypothesized values against observed means.

T-Tests ($\bar{x} = 527.82$ $\mu_0 = 520$)

Variable	DF	t Value	Pr > |t|
sshist	87	0.96	0.3413

As can be seen from the above *t*-test the difference between the observed and hypothesized means for the *sshist* variable were not statistically significant (*p* > .05). This would imply that the observed variable mean does not vary significantly from what was expected.

T-Tests ($\bar{x} = 54.807$ $\mu_0 = 55$)

Variable	DF	t Value	Pr > |t|
verbal	87	-0.16	0.8720

Similarly, no significant mean difference was observed for the *verbal* variable, $t(87) = -.16$, $p > .05$.

T-Tests ($\bar{x} = 25.034$ $\mu_0 = 22$)

Variable	DF	t Value	Pr > |t|
science	87	5.86	<.0001

Last, the *science* variable's difference between observed and hypothesized means was statistically significant, $t(87) = 5.86$, $p < .001$. This would suggest that of the three variables only the *science* variable varied significantly from what would have been expected. However, since the three variables constitute a collection of scores on a single series of assessments it would be advisable to analyze the three scores collectively in a multivariate approach.

6) Compute the corresponding Hoteling's T^2.

Following from your reading and from class lecture there are two ways you could investigate the multivariate mean differences by using either the traditional T^2 computation (with the appropriate F transformation) or using the book's shorthand computation. The traditional (Johnson & Wichern, 2002) Hotelling's T^2 can be expressed as:

$$T^2 = N \times (\overline{X} - \overline{\mathrm{M}})^T \times S^{-1} \times (\overline{X} - \overline{\mathrm{M}}),$$

which can be expressed in a *F* statistic form by the following transformation:

$$F = \frac{n-p}{p \times (n-1)} \times T^2,$$

with a critical value equal to the *F* statistic with *p* and *n-p* degrees of freedom. Therefore when transforming the T^2 value we can compare it against the critical *F* in order to infer a statistically significant difference between the mean vectors of the observed and hypothesized populations. In the example discussed the *proc iml* derived statistics (see attached syntax) were $T^2 = 44.0294$, $F = 14.3391$ with a critical cut off value of 2.7119. Further, the Raykov and Maercoulides (2008) shorthand formula can be used,

$$T^{2*} = (\overline{X} - \overline{\mathrm{M}})^T \times S^{-1} \times (\overline{X} - \overline{\mathrm{M}}),$$

with a critical cut off value that has been scaled appropriately as:

$$T^{2*}_{CRIT} = \frac{p \times (n-1)}{n \times (n-p)} \times F_{(p,\ n-p,\ \alpha)}\ .$$

Subsequently, the Raykov and Marcoulides estimate (here called *T-squared star*) was .5003 with a critical value of .0946. It follows that the two statistics ought to evaluate the relationship equivalently. In fact, both methods will produce the same ratio of observed statistic to critical value (allowing small deviations due to rounding error).

$$T^2\ Ratio = \frac{F}{F_{CRIT}} = \frac{14.3391}{2.7119} = 5.2875$$

$$T^{2*}\ Ratio = \frac{T^{2*}}{T^{2*}_{CRIT}} = \frac{.5003}{.0946} = 5.2885$$

7) What is your interpretation regarding the data based on your analyses in questions 5 and 6?

When investigating single univariate one-sample t-test it is not entirely clear whether there is any interrelationship between any other variables in the dataset since the focus is singular on the specific value of the hypothesized population for one variable only. The multivariate generalization allows for a singular assessment of all three tests not only providing information regarding the unique constellation of all variable means, but also decreasing *Type I Error* rates while generally (especially when multivariate normality is met) increasing power.

In the above example only the *science* variable based one-samples *t*-test was significant leading the researcher to fail to reject the corresponding null hypotheses for the other two variables. However, given the multivariate assessment of mean differences it was shown that (using either the traditional or shorthand form) the null hypothesis of equal means would be rejected at the $p < .05$ level suggesting a statistically significant difference in means for the observed sample.

8) Transform the dataset you have computed the Hoteling's T^2 on as follows: Divide *sshist* by 100 (call it *hist*), multiply verbal by 1.5 and add 10 (call it *verb*), and subtract 5 from *science* and divide it by 5 (call it *scien*). What are the new measures of central tendency?

```
The MEANS Procedure

Variable     N          Mean       Std Dev       Minimum       Maximum
hist        88        5.2782        0.7665        3.4800        7.3300
verb        88       92.2102       16.8249       47.5000      122.5000
scien       88        4.0068        0.9715        1.8000        6.2000
```

It is also important to remember that if variables are being linearly transformed their corresponding hypothesized population means need to be adjusted accordingly. Consequently, the corresponding hypothesized new mean values for the *hist, verb* and *scien* variables are 5.2, 92.5 and 3.4 respectively.

9) Rerun the Hoteling's T^2 on your newly transformed dataset. ***(Remember you must also transform the corresponding hypothesized means accordingly).***

The resulting T^2 and T^2–star values are identical to those computed in question 6. Both the shorthand form and the traditional T^2 formula should have produced the same statistics.

10) What conclusions can you draw regarding the multivariate generalization of the t-test?

From the above simulation it should be apparent that the multivariate generalization of the *t*-test, much like its univariate synonym, is invariant to linear transformations. This means that even though both the variables and the corresponding means were changes so as to appear on the surface have a different relational quality, the underlying relationship is unaffected by the new scale the variable has been transformed to. This is both important and practical when considering the transformation of one variable into another scale (say Fahrenheit degrees into Celsius, or Dollars into Euros) whilst still being interested in the multivariate comparison of the individual variable means against some population estimates.

```
/* ---------------------------------------------------------------------------------------- */
/*       Lab Assignment No.2: Multivariate Data Analsyis (SAS Syntax)        */
/*                                  Due February 18, 2009                  ------      */
/* ---------------------------------------------------------------------------------------- */

%let path='C:\ YOUR PATH HERE \';
libname Multi &path;


/* QUESTION 1 */

/* Scatterplot Matrix */
proc insight data=Multi.t5_2;
  scatter sshist verbal science * sshist verbal science;
run;
quit;

/* Producing and saving out the correlation matrix */
proc corr data=Multi.t5_2 cov  csscp outp=Multi.t5_2_cor;
title 'Correlations';
run;


/* QUESTION 2 */

/* Analyzing univariate statistics and normality */
proc univariate data=Multi.t5_2 normaltest plots;
title "Univariate Analysis of T5_2 Dataset";
run;

/* Deleting outlier observation in verbal variable */
data Multi.t5_2b;
          set Multi.t5_2;
          if verbal < 190;
run;

/* Rerunning univariate statistics */
proc univariate data=Multi.t5_2b normaltest plots;
title "Univariate Analysis of T5_2b Dataset";
run;

proc insight data=Multi.t5_2b;
  scatter sshist verbal science * sshist verbal science;
run;
quit;
```

```
/* QUESTION 3 */

/* Producing Leverage and Mahalonobis Values */
proc reg data=Multi.t5_2b;
        model sshist=verbal science / influence;
        output out=hii H=Leverage;
run;
quit;

proc iml;
title 'Mahalonobis Values';

        use hii;
        read all var {Leverage} into hii[colname=name];
        N = nrow(hii);
        md = (N-1)*(hii-(1/N));
        cases = (1:N)`;
        values = cases || hii || md;

        /* Critical Values */
        hiicrit1 = (2*3)/N;
        hiicrit2 = 2*((3+1)/N);
        hiicrit3 = ((2*(gaminv(.99,1.5)))/(N-1))+(1/N);
        mdcrit_05 = 2*(gaminv(.95,1.5));
        mdcrit_01 = 2*(gaminv(.99,1.5));

        print hiicrit1;
        print hiicrit2;
        print hiicrit3;
        print mdcrit_05;
        print mdcrit_01;

        create Multi.vals from values [colname={'ID' 'Leverage' 'MD'}];
        append from values;

quit;

/* Central distribution of computed values */

proc means data=Multi.vals mean median std stderr skew kurt max min clm;
run;

/* Plot the two values */

proc gplot data=Multi.vals;
        title 'Multivariate Outliers';
```

```
        plot Leverage*id / vref=.0909;
        plot MD*id / vref=11.3449;
run;
quit;


/* QUESTION 4 */

/* ------------------------------------------------- */
/* Run the mulnorm macro specification */
/* ------------------------------------------------- */

/* Compute the Mardia estimates */
%multnorm(data=Multi.t5_2b, var = sshist verbal science)

/* Compare this to the dataset with the one univariate outlier */
%multnorm(data=Multi.t5_2, var = sshist verbal science)


/* QUESTION 5 */

/* One sample t-tests for the three variables */
proc ttest data=Multi.t5_2b
        h0 = 520;
        var sshist;
run;
proc ttest data=Multi.t5_2b
        h0 = 55;
        var verbal;
run;
proc ttest data=Multi.t5_2b
        h0 = 22;
        var science;
run;


/* QUESTION 6 */

/* Hotelling's T-squared */
/* Saving out Variance / Covariance Matrix */
proc corr data=Multi.t5_2b nocorr cov outp=covariance;
run;
data Multi.covar;
        set covariance;
        put sshist verbal science;
        if _type_='COV';
```

```
        drop _type_ _name_;
run;

/* Computing Hotelling's T-squared */
proc iml;
        use Multi.covar;
        read all into S [colname=name];
        use Multi.t5_2b;
        read all into t5 [colname=name];

        N = nrow(t5);
        vecone = j(N,1);
        mu = {520, 55, 22};
        t5m = ((vecone`*t5)/N)`;
        diff = t5m-mu;

        tsqr_str = diff`*(inv(S))*diff;
        tcrit_str = ( (3*87)/(88*85) )*finv(.95, 3, 85);

        print 'Shorthand form';
        print tsqr_str;
        print tcrit_str;

        tsqr = N*(diff`*(inv(S))*diff);
        tintof = ((88-3)/(3*(88-1)))*tsqr;
        tcrit = finv(.95, 3, 85);

        print 'T-squared in class form';
        print tsqr;
        print tintof;
        print tcrit;

quit;

/* QUESTION 8 */

data Multi.t5_new;
        set Multi.t5_2b;
        hist = sshist/100;
        verb = (verbal*1.5)+10;
        scien = (science-5)/5;
        drop sshist verbal science;
run;

proc means data=Multi.t5_new;
run;
```

```
/* QUESTION 9 */

proc corr data=Multi.t5_new nocorr cov outp=covariance;
run;
data Multi.newcovar;
        set covariance;
        put hist verb scien;
        if _type_='COV';
        drop _type_ _name_;
run;

proc iml;
        use Multi.newcovar;
        read all into S [colname=name];
        use Multi.t5_new;
        read all into t5 [colname=name];

        N = nrow(t5);
        vecone = j(N,1);
        mu = {5.2, 92.5, 3.4};
        t5m = ((vecone`*t5)/N)`;

        print 'New Variable Means';
        print t5m;

        diff = t5m-mu;

        tsqr_str = diff`*(inv(S))*diff;
        tcrit_str = ( (3*87)/(88*85) )*finv(.95, 3, 85);

        print '2nd Shorthand form';
        print tsqr_str;
        print tcrit_str;

        tsqr = N*(diff`*(inv(S))*diff);
        tintof = ((88-3)/(3*(88-1)))*tsqr;
        tcrit = finv(.95, 3, 85);

        print '2nd T-squared in class form';
        print tsqr;
        print tintof;
        print tcrit;

quit;
```