STAT505 Assessment #1

1. For the following matrix, calculate the sample mean vector $\overline{\mathbf{x}}$, the sample covariance matrix $\mathbf{S}$, the sample correlation matrix $\mathbf{R}$, and the generalized sample variance $|\mathbf{S}|$.

$$\begin{bmatrix} X_1 & X_2 \\ 3 & 4 \\ 6 & -2 \\ 3 & 1 \end{bmatrix}$$

The sample mean is found by averaging the values in the first column:

$$\overline{x}_1 = \frac{3+6+3}{3} = 4.$$

Similarly, $\overline{x}_2 = 1$. Thus, $\overline{\mathbf{x}} = [4, 1]'$. The sample variance is

$$s_{11} = \frac{1}{2}\left((3-4)^2 + (6-4)^2 + (3-4)^2\right) = 3.$$

Similarly, $s_{22} = 9$. The covariance is

$$s_{12} = \frac{1}{2}\left((3-4)(4-1) + (6-4)(-2-1) + (3-4)(1-1)\right) = -4.5.$$

By symmetry, we have

$$\mathbf{S} = \begin{bmatrix} 3 & -4.5 \\ -4.5 & 9 \end{bmatrix}.$$

The generalized variance is $|\mathbf{S}| = 6.75$. Finally, computing each correlation by $r_{ij} = s_{ij}/\sqrt{s_{ii}s_{jj}}$, we have the correlation matrix:

$$\mathbf{R} = \begin{bmatrix} 1 & -0.866 \\ -0.866 & 1 \end{bmatrix}.$$

With the code below, SAS can compute these quantities.

```
data a;
input x1 x2 @@;
cards;
3 4 6 -2 3 1
;
proc corr data=a noprob; run;
proc iml;
  start genvar;
    one=j(nrow(x),1,1);
    ident=i(nrow(x));
    s=x‘*(ident-one*one‘/nrow(x))*x/(nrow(x)-1.0);
    genvar=det(s);
    print s genvar;
  finish;
  use a;
  read all var{x1 x2} into x;
  run genvar;
quit;
```

2. (*adapted from J&W Exercise 3.15*) Consider the data matrix

$$\begin{bmatrix} X_1 & X_2 & X_3 \\ 1 & 4 & 3 \\ 6 & 2 & 6 \\ 8 & 3 & 3 \end{bmatrix}.$$

We have $n = 3$ observations on $p = 3$ variables X_1, X_2, and X_3. Form the linear combinations

$$\mathbf{b}'\mathbf{X} = [1 \quad 1 \quad 1] \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = X_1 + X_2 + X_3$$

$$\mathbf{c}'\mathbf{X} = [1 \quad 2 \quad -3] \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = X_1 + 2X_2 - 3X_3$$

Find the sample means, variances, and covariance of $\mathbf{b}'\mathbf{X}$ and $\mathbf{c}'\mathbf{X}$

First, we can find the sample mean vector and sample covariance matrix using the steps from Problem 1 above. They are

$$\overline{\mathbf{x}} = \begin{bmatrix} 5 \\ 3 \\ 4 \end{bmatrix} \quad \text{and} \quad \mathbf{S} = \begin{bmatrix} 13.0 & -2.5 & 1.5 \\ -2.5 & 1.0 & -1.5 \\ 1.5 & -1.5 & 3.0 \end{bmatrix}.$$

Then, the sample mean vector for $\mathbf{b}'\mathbf{X}$ is

$$\mathbf{b}'\overline{\mathbf{x}} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 5 \\ 3 \\ 4 \end{bmatrix} = 5 + 3 + 4 = 12,$$

and the sample variance for $\mathbf{b}'\mathbf{X}$ is

$$\mathbf{b}'\mathbf{S}\mathbf{b} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 13.0 & -2.5 & 1.5 \\ -2.5 & 1.0 & -1.5 \\ 1.5 & -1.5 & 3.0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = 12.$$

Likewise, $\mathbf{c}'\overline{\mathbf{x}} = -1$, $\mathbf{c}'\mathbf{S}\mathbf{c} = 43$, and the covariance is $\mathbf{b}'\mathbf{S}\mathbf{c} = -3$. These calculations can be computed in SAS with this code.

```
data b;
input x1 x2 x3 @@;
bx = x1+x2+x3;
cx = x1+2*x2-3*x3;
cards;
1 4 3 6 2 6 8 3 3
;
proc corr data=b cov noprob;
var bx cx;
run;
```

3. Consider the data "corporations.dat" from 10 U.S. corporations. The variables are Sales, Profits, and Assets. All figures are in millions of dollars. These data may be input into SAS using the following code:

```
data corp;
infile 'v:\corporations.dat' delimiter='09'x;
input name :$20. sales profits assets;
run;
```

The "`delimiter`" command tells SAS the columns of data are delimited by tabs, and the "`:$20.`" symbols tell SAS the corporation names may have a length of up to 20 characters.

Consider the following linear combinations:

$$\text{Overhead} = \text{Sales} - \text{Profits}$$
$$\text{Assets after Sales} = \text{Assets} + \text{Sales}$$

(a) Find the sample mean of Overhead and the sample mean of Assets after Sales.

(b) Find the sample variance of Overhead and the sample variance of Assets after Sales.

(c) Find the sample covariance and correlation between Overhead and Assets after sales. Describe the relationship between them.

Using the approach above, the SAS proc corr procedure produces the output we want. The correlation (.94591) between overhead and assets after sales indicates a very strong positive linear relationship.

```
data corp;
  infile 'v:\corporations.dat' delimiter='09'x;
  input company :$20. sales profits assets;
  overhead = sales - profits;
  aas = assets + sales;
  run;
proc corr data=corp cov noprob;
  var overhead aas;
  run;
```

2 Variables:	overhead aas

Covariance Matrix, DF = 9		
	overhead	aas
overhead	950787934	2441106125
aas	2441106125	7004653354

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
overhead	10	59382	30835	593818	30003	122750
aas	10	143558	83694	1435575	58052	300271

Pearson Correlation Coefficients, N = 10		
	overhead	aas
overhead	1.00000	0.94591
aas	0.94591	1.00000