Problem 1

A multivariate data set contains 42 measurements on air-pollution variables in Los Angeles. The columns are x_1 = wind, x_2 = solar radiation, x_3 = CO, x_4 = NO, x_5 = NO_2, x_6 = O_3, and x_7 = HC.

1. Using notation consistent with our lessons, explain briefly what X_{ij} and $\overline{X}_j$ represent in terms of the variables here. Similarly, explain briefly what μ_j and σ_{jk} represent. What does it mean if $\sigma_{jk} = 0$?

   X_{ij} is ith observed measurement for the jth variable, $\overline{X}_j$ is the sample mean measurement for the jth variable, μ_j is the population mean of the jth variable, and σ_{jk} is the population covariance between variables j and k. If $\sigma_{jk} = 0$, then variables j and k are uncorrelated.

2. Explain briefly what it would mean if the correlation between solar radiation and CO is close to 1, but the partial correlation between solar radiation and CO is close to 0, given NO.

   If the correlation between solar radiation and CO is close to 1, they are strongly and positively linearly related. However, if their partial correlation is weak given NO, it means that this relationship can be explained by their common relationship with NO.

3. Consider only solar radiation and CO (carbon monoxide) for these parts.

   (a) Draw by hand or describe in words what you expect a QQ plot to look like if these variables have a multivariate normal distribution.

   A QQ plot shows the relationship between Mahalanobis distances and theoretical chi-square quantiles. If the variables are multivariate normal, then their Mahalanobis distances should agree with the chi-square quantiles, and the QQ plot should show a reasonably straight line

   (b) Repeat part (a) if these variables do not have a multivariate normal distribution.

   In the case where the variables are not multivariate normal, their Mahalanobis distances may not follow a chi-square distribution, and plotting them against the actual chi-square quantiles will usually reveal curvature and outliers.

   (c) With sample mean vector $[74.0, 5.0]'$ and sample covariance matrix

$$\begin{bmatrix} 300.0 & 4.0 \\ 4.0 & 1.5 \end{bmatrix}$$

   compute a 95% confidence interval for the population mean of solar radiation and a 95% confidence interval for the population mean of CO. Show your work for this.

   With 95% confidence and a sample size of 42, the t multiplier is 2.02. For solar radiation, the confidence interval is

$$74.0 \pm 2.02\sqrt{300.0/42} = (68.60, 79.40)$$

   For CO, the interval is

$$5.0 \pm 2.02\sqrt{1.5/42} = (4.62, 5.38)$$

   (d) Explain why one may *not* say that "with 95% confidence, both of the intervals above in (c) cover their respective parameters simultaneously."

   In repeated sampling, each of the formulas above will capture its respective population mean 95% of the time. However, there would be samples where only one interval captures its parameter. So, the percentage of the time they *both* capture the parameters would be lower than 95%.

Problem 2

Given $\mathbf{X} = [X_1, X_2, X_3]' \sim N(\mu_{\mathbf{X}}, \mathbf{\Sigma_X})$ with

$$\mu_{\mathbf{X}} = \begin{bmatrix} -3 \\ 1 \\ 4 \end{bmatrix} \quad \text{and} \quad \mathbf{\Sigma_X} = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix},$$

what is the conditional distribution of $X_1+X_2+X_3$ given that $2X_1-X_2-X_3 = c$ and $X_3-X_2 = d$? Define $\mathbf{Y} = \mathbf{AX}$, where

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 2 & -1 & -1 \\ 0 & -1 & 1 \end{bmatrix}$$

so that $\mathbf{Y} \sim N(\mu_{\mathbf{Y}}, \mathbf{\Sigma_Y})$ with

$$\mu_{\mathbf{Y}} = \mathbf{A}\mu_{\mathbf{X}} = \begin{bmatrix} 2 \\ -11 \\ 3 \end{bmatrix} \quad \text{and} \quad \mathbf{\Sigma_Y} = \mathbf{A\Sigma_X A}' = \begin{bmatrix} 4 & -7 & -1 \\ -7 & 19 & 7 \\ -1 & 7 & 7 \end{bmatrix},$$

Now, we can apply our result for conditional normal distributions to say that $Y_1|Y_2 = c, Y_3 = d$ is normally distributed with mean μ and variance σ^2, where

$$\mu = 2 + \begin{bmatrix} -7 & -1 \end{bmatrix} \begin{bmatrix} 19 & 7 \\ 7 & 7 \end{bmatrix}^{-1} \begin{bmatrix} c+11 \\ d-3 \end{bmatrix} = -\frac{32}{7} - \frac{1}{2}c + \frac{5}{14}d.$$

and

$$\sigma^2 = 4 - \begin{bmatrix} -7 & -1 \end{bmatrix} \begin{bmatrix} 19 & 7 \\ 7 & 7 \end{bmatrix}^{-1} \begin{bmatrix} -7 \\ -1 \end{bmatrix} = \frac{6}{7}.$$

```
proc iml;
* parameters for X;
mu_X = {-3,1,4};
sigma_X = {1 -2 0,-2 5 0,0 0 2};
A = {1 1 1,2 -1 -1,0 -1 1};
* parameters for Y = AX;
mu_Y = A*mu_X;
sigma_Y = A*sigma_X*t(A);
* partitioned parameters;
mu_Y1 = mu_Y[1];
mu_Y2 = mu_Y[2:3];
sigma_Y11 = sigma_Y[1,1];
sigma_Y12 = sigma_Y[1,2:3];
sigma_Y21 = t(sigma_Y12);
sigma_Y22 = sigma_Y[2:3,2:3];
* conditional parameters;
* mean is linear function of y2 with coefficients;
beta1 = sigma_Y12*inv(sigma_Y22);
beta0 = mu_Y1-beta1*mu_Y2;
beta = beta0||beta1;
var = sigma_Y11-sigma_Y12*inv(sigma_Y22)*sigma_Y21;
print beta var;
quit;
```

Problem 3

Consider the SAS code

1. How would this code need to be altered to produce 90% prediction ellipses for a bivariate normal distribution with mean vector and covariance matrix

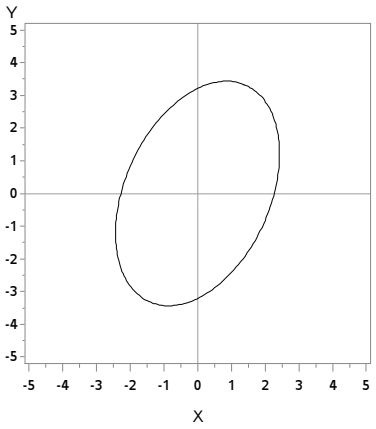
$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 2.0 & -.5 \\ -.5 & 1.0 \end{bmatrix}$$

   In the IML procedure, "sigma" would be altered as follows, and the line beginning with "z" would be altered to reflect the change in confidence:

```
sigma={2.0000 -0.5000, -0.5000 1.0000};
z=z*d*e'*sqrt(4.605);
```

2. Include both the original and modified ellipses, or explain in words what qualities you would see with these plots. Comment on their differences.

   In addition to having smaller area because of the reduction in confidence, the modified ellipse is angled with a negative slope because of the negative covariance, and the modified variance is larger in the horizontal direction.



unmodified ellipse

modified ellipse

Note: although this image shows the modified ellipse mean at $(2, 3)$*, it should still be at* $(0, 0)$.