**STA 437/1005, Fall 2008 — Assignment #1 Solutions.**

**Question 1:** Let $X$, $Y$, and $Z$ be independent random variables, all with the $N(1,1)$ distribution. Define the random variables $A$ and $B$ as follows:

$$\begin{aligned} A &= X + 2Y \\ B &= X + Y + Z \end{aligned}$$

Finally define $C$ as the random vector with $A$ and $B$ as components (ie, $C = [A \;\; B]'$).

a) What is the mean vector of $C$?

*See pages 75-76 of the text. We can find the expectations of $A$ and $B$ as*

$$E(A) = E(X + 2Y) = E(X) + 2E(Y) = 1 + 2 \times 1 = 3$$
$$E(B) = E(X + Y + Z) = E(X) + E(Y) + E(Z) = 1 + 1 + 1 = 3$$

*So the expectation of $C = [A \; B]'$ is $[3 \; 3]'$.*

b) What is the covariance matrix of $C$?

*See pages 75-76 of the text. We can combine the definitions of $A$, $B$, and $C$ to give that $C = W[X \; Y \; Z]'$, where the matrix $W$ is defined as*

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

*Since $X$, $Y$, and $Z$ are independent, and all have variance 1, the covariance matrix of $[X \; Y \; Z]'$ is the identity matrix, $I$. Using formula (2-45) from the text, we find that*

$$\text{Cov}(C) = WIW' = WW' = \begin{bmatrix} 1 & 2 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 5 & 3 \\ 3 & 3 \end{bmatrix}$$

c) What is the conditional distribution of $A$ given that $B = 1$?

*See pages 160-161 of the text. $C = [A \; B]'$ has a multivariate normal distribution (since $A$ and $B$ have normal distributions and are independent). According to Result 4.6 in the text, the conditional distribution of $A$ given a value for $B$ is also normal. The mean of this conditional distribution depends on the value of $B$ that is conditioned on:*

$$E(A \,|\, B = 1) = E(A) + \text{Cov}(A, B)[\text{Cov}(B)]^{-1}(1 - E(B)) = 3 + 3(3)^{-1}(1 - 3) = 1$$

*The conditional variance of $A$ does not depend on the particular value of $B$ that is conditioned on:*

$$\begin{aligned} \text{Var}(A \,|\, B) = \text{Cov}(A \,|\, B) &= \text{Cov}(A) - \text{Cov}(A, B)[\text{Cov}(B)]^{-1}\text{Cov}(B, A) \\ &= 5 - 3(3)^{-1}3 = 2 \end{aligned}$$

**Question 2:** Suppose $X$ is a random vector of length $p$ with covariance matrix $\Sigma_X$. Define $Y = QX$, where $Q$ is some $p \times p$ orthogonal matrix, and let $\Sigma_Y$ be the covariance matrix of $Y$.

a) Find a simple expression for $\Sigma_Y$.

*Referring again to page 76 of the text, we get that* $\Sigma_Y = Q\Sigma_X Q'$.

b) Suppose that $e$ is an eigenvector of $\Sigma_X$ with eigenvalue $\lambda$. Prove that $Qe$ is an eigenvector of $\Sigma_Y$, and find what eigenvalue is associated with it.

*To see that* $Qe$ *is an eigenvector of* $\Sigma_Y$*, we multiply it by* $\Sigma_Y$*, to get*

$$\Sigma_Y(Qe) \;=\; (Q\Sigma_X Q')(Qe) \;=\; Q\Sigma_X(Q'Q)e \;=\; Q\Sigma_X e \quad =\; Q(\lambda e) \;=\; \lambda(Qe)$$

*Here, we have used the fact that* $Q$ *is an orthogonal matrix, so that* $Q^{-1} = Q'$*, and the fact that* $e$ *is an eigenvector of* $\Sigma_X$ *with eigenvalue* $\lambda$*, so that* $\Sigma_X e = \lambda e$*. We see that* $Qe$ *is an eigenvector of* $\Sigma_Y$ *with eigenvalue* $\lambda$*.*

**Question 3:** Recall the spectral decomposition theorem: If $A$ is a $k \times k$ symmetric real matrix, it is possible to find a set of $k$ eigenvectors of $A$ that are orthogonal and have length one, and if $e_1, \ldots, e_k$ are any such set of eigenvectors, with eigenvalues $\lambda_1, \ldots, \lambda_k$, then $A = \lambda_1 e_1 e_1' + \cdots + \lambda_k e_k e_k'$.

Use this theorem to prove that if $A$ is a symmetric matrix with eigenvectors $e_1, \ldots, e_k$ that are orthogonal and have length one, with non-zero eigenvalues $\lambda_1, \ldots, \lambda_k$, then $B = e_1 e_1'/\lambda_1 + \cdots + e_k e_k'/\lambda_k$ is the inverse of $A$. Note that although there may be several ways of proving this, for this question you should prove it by multiplying $A$ and $B$ and verifying that the result is the identity matrix, using the spectral decomposition theorem.

*From the spectral decomposition theorem,* $A = \lambda_1 e_1 e_1' + \cdots + \lambda_k e_k e_k'$*, so we can write* $AB$ *as*

$$\begin{aligned}
AB &= \left[\lambda_1 e_1 e_1' + \cdots + \lambda_k e_k e_k'\right] \cdot \left[e_1 e_1'/\lambda_1 + \cdots + e_k e_k' \lambda_k\right] \\
&= \sum_{i=1}^{k}\sum_{j=1}^{k} (\lambda_i e_i e_i')\,(e_j e_j'/\lambda_j) \\
&= \sum_{i=1}^{k} (\lambda_i e_i e_i')\,(e_i e_i'/\lambda_i) \;+\; \sum_{i \neq j} (\lambda_i e_i e_i')\,(e_j e_j'/\lambda_j) \\
&= \sum_{i=1}^{k} \lambda_i e_i (e_i' e_i) e_i'/\lambda_i \;+\; \sum_{i \neq j} \lambda_i e_i (e_i' e_j) e_j'/\lambda_j \\
&= e_1 e_1' + \cdots + e_k e_k'
\end{aligned}$$

*The product of sums in the first line above is equal to the sum of all the pairwise products, but the terms in this sum for* $i \neq j$ *are zero, since* $e_i' e_j = 0$ *when* $i \neq j$ *(since the eigenvectors are orthogonal. We can use the spectral decomposition theorem to see that the final result is equal to the identity matrix, because* $e_1, \ldots, e_k$ *are eigenvectors of the identiy matrix with eigenvalues of one (since* ***all*** *non-zero vectors are eigenvectors of the identity matrix with eigenvalues of one).*

**Question 4:** The effect (if any) of air pollution on mortality has been studied for many years, and has large implications for public policy. From the course web page,

`http://www.utstat.toronto.edu/~radford/sta437`

you can get a file containing daily data on weather, air pollution, and deaths in Toronto from 1992 to 1997. This data file contains 2192 lines, one per day, in time order, with each line containing the values of 10 variables. There is also a header line at the front with the names of the variables.

The variables are as follows:

| | |
|---|---|
| `year` | Year, from 1992 to 1997 |
| `month` | Month, from 1 (January) to 12 (December) |
| `day` | Day of the month, from 1 to 31 |
| `deaths` | Number of deaths in Toronto |
| `pressure` | Average air pressure, in kilopascals |
| `temperature` | Average temperature, in degrees Celcius |
| `humidity` | Average relative humidity, percent |
| `so2` | Average level of sulfer dioxide, ppb |
| `ozone` | Average level of ozone, ppb |
| `pm10` | Average level of 10 micron particulate matter, micrograms per cubic meter |

The `pm10` variable is observed on only some days, with the value for other days being set to `NA`.

In interpreting this data, it is important to note that the weather itself is known to have an effect on mortality, and that the weather also has an effect on the level of pollutants (`so2`, `ozone`, and `pm10`).

Read this data into R, look at it, and report any conclusions you may find about how these variables are related, whether they have normal distributions, and how they might be transformed to have distributions closer to normal. In your report, include a small number of plots or other R output that justifies your conclusions.

In your report, you should also discuss to what extent this data can be regarded as a random sample from a distribution that is of interest regarding the question of whether air pollution has an effect on mortality.

For this assignment, you need not perform any formal statistical tests. You should just make informal assessments based on plots and sample statistics, and using your common sense knowledge.

*This answer refers to a PDF file of plots produced using R, which is available from the course web page, along with the R program that produced them.*

*Whether this data can be regarded as a random sample from a distribution of interest can be largely discussed without even looking at the data. Our common sense knowledge tells us that a hot day is often followed by another hot day, and similarly for other weather variables, so we would not expect daily temperatures (and hence also things that may depend on temperature, such as pollution and deaths) to be independent. This lack of independence will certainly be an issue for any formal statistical tests (not done for this assignment). We also know that the weather is different in different months, so we don't expect a single distribution for the weather (and hence other variables too). We might also expect that pollution levels have changed over the years. We might be able to ignore these changes in distributions if the relationship of deaths to the other*

*variables stays the same, but since that is also not clear, we would really need to do an analysis that accounts for seasonal variation and trends.*

*The first page of plots has boxplots for each variable other than year/month/day versus year and versus month. These plots confirm that there are considerable differences in temperature, humidity, and ozone for different months, and also noticable differences by month in the other variables — deaths, pressure (variance rather than mean), so2, and pm10. Trends over the years are less pronounced, but there seems to have been some decrease in pm10 (at least with regard to extreme values) and some increase in ozone.*

*The second page of plots has scatterplots of all pairs of variables other than year/month/day. These scatterplots show that deaths are negatively correlated with temperature, and that ozone is positively correlated with temperature. This could cause confusion when assessing whether ozone causes increased deaths, and indeed the correlation of deaths with ozone is -0.11, which is opposite to what one might expect, but this may be misleading due to the correlations of deaths and ozone with temperature.*

*These scatterplots also show that many pairs of variables don't have bivariate normal distributions. We can also look at the univariate distributions for each variable, using histograms and normal QQ plots, as is done on the third and fourth page of plots. We can observe the following:*

- *Deaths has a right-skewed distribution. This is as expected if the count of deaths each day to has a Poisson distribution. Theory says that taking the square root should produce a more normal distribution (see page 192 in the text). This should be considered, though one also needs to consider whether this transformation might cause other difficulties (eg, in interpretation, or with regard to linearity of the relationship of death to other variables).*
- *Pressure seems to have close to a normal distribution.*
- *Temperature has a bimodal distribution. It seems pointless to try to make this more normal by a transformation (a complex transformation would be needed, which would probably cause other problems).*
- *Humidity has a left-skewed distribution. There's really no reason to expect it to be normally-distributed, and it may be pointless to try to make it so by a transformation.*
- *SO2 has a right-skewed distribution. Since some values for so2 are zero, a log tranformation would be problematic. Taking the square root might help make the distribution more normal, but again might cause other problems.*
- *Ozone is also right-skewed, and again a square root (or log) transformation might make it more normal.*
- *PM10 is right-skewed, and again a square root (or log) transformation might make it more normal.*

*The fifth page of plots show histograms and normal QQ plots for the square root of deaths, the square root of so2, the square root of ozone, and the log of pm10, which show that these variables are closer to having normal distributions after these transformations.*

*The sixth page shows scatterplots of all pairs of variables except year/month/day after these transformation. Some relationships seem a bit clearer, such as the negative relationship of deaths to ozone, and some bivariate distributions seem closer to normal (eg, sqrt ozone versus log pm10). Some bivariate distributions are clearly not normal, however (eg, temperture versus sqrt ozone).*