STAT 504 – Spring 2019 – Solutions to Assignment 1

PART I

1. The scatter plot shows that bmd is positively correlated with weight — women with higher weights tend to have higher bmd values. The plot does not reveal any obvious curvature in this trend. It may be a straight line relationship.

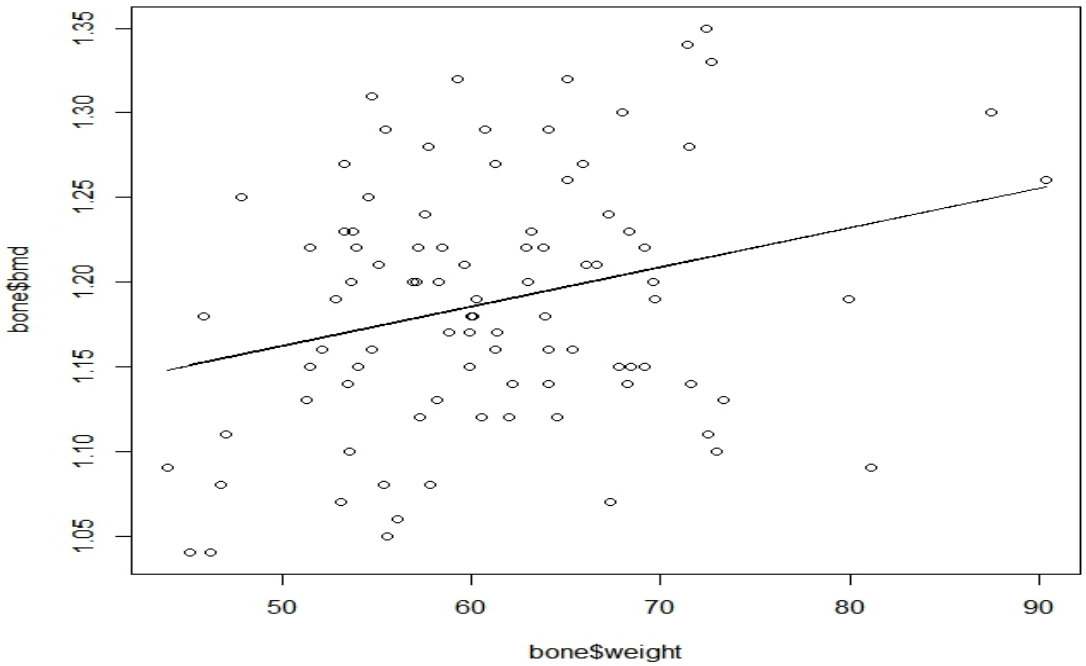
2. The least squares estimates of the coefficients yield

```
bmd =   1.04584 + 0.00233(weight)
        (.0522)   (0.0008)
```

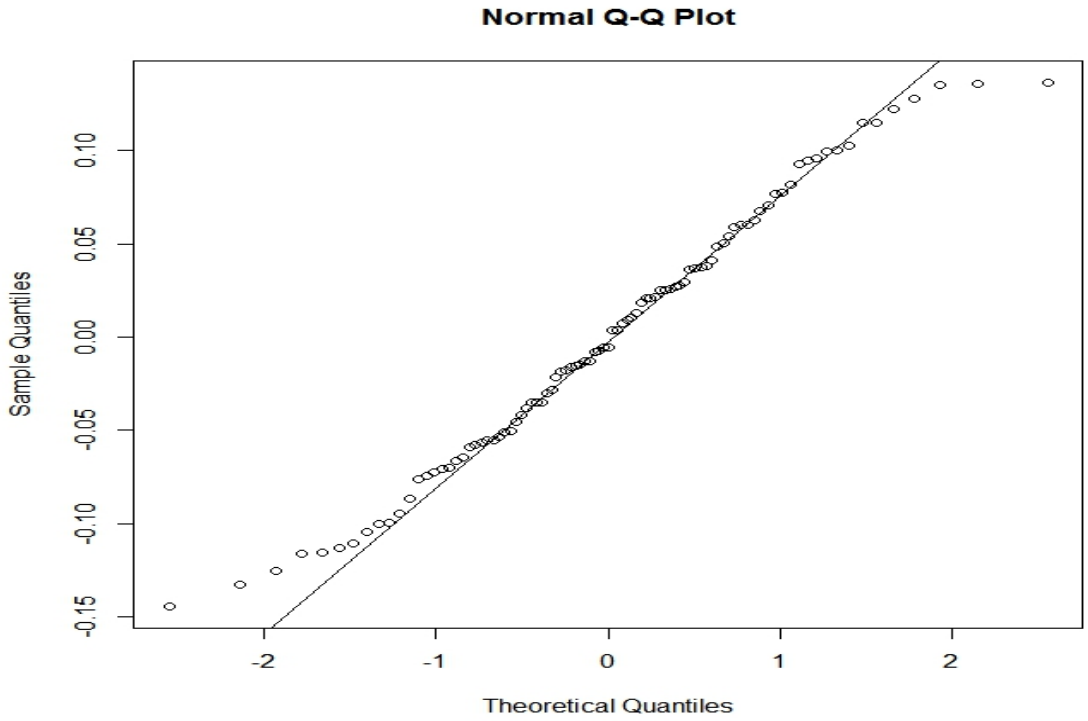
as the estimated regression line. Standard errors are given in parentheses underneath the estimated coefficients. The ANOVA table is

```
            Df Sum of Sq    Mean Sq  F Value     Pr(F)
weight       1 0.0384736 0.03847359 7.588347   0.0071
Residuals   91 0.4613780 0.00507009
```

The ANOVA table shows that the coefficient for weight is significantly different from zero (p-value=.0071). Thus, higher weight appears to be associated with higher bmd values as expected. The plot of the residuals against weight reveals no obvious deviation from a straight line relationship. It also suggests that the homogeneity of error variances assumption is reasonable.



The normal probability plot of the residuals indicates that the distribution of the random errors has fewer extreme values than one would expect from a normal distribution, but this modest departure from a normal distribution would not have a noticeable adverse effect on F-tests or t-tests.



You have also accomplished some of these tasks by entering code into the R console window. Try using the code shown below.

```
#-------------------------------------------------------------#
# Read data from the file into a data frame.  Here we       #
# also add column names                                       #
#-------------------------------------------------------------#
bone <- read.table("c:/stat504/bone2.density.txt",
  col.names=c("id","group","bmd","weight","age","caffeine",
           "calcium", "growmilk", "teenmilk", "twenmilk" ) )

#----------------------#
# remove id column     #
#----------------------#
bone <- bone[,-1]

#-------------------------------------------------------------#
# (1) Plot bmd * weight                                       #
#-------------------------------------------------------------#
plot( bone$weight, bone$bmd )

#-------------------------------------------------------------#
# (2) Fit a regression model                                  #
#-------------------------------------------------------------#
bone.out1 <- lm( formula=bmd~weight, data=bone )

#--------------------------------------#
# Add the regression line to the plot  #
#--------------------------------------#
lines( bone$weight, fitted( bone.out1 ) )
```

```
#-----------------------------------------------#
# Report the estimates and their standard errors #
#-----------------------------------------------#
summary(bone.out1)
anova(bone.out1)
```

3. (a) Starting with a full model including all the explanatory variables the R stepwise search finds as the best model:

```
[Model 1]   bmd = 0.8687 + 0.0518 I(group2) + 0.0509 I(group3) + 0.0020(weight)
                  (0.0708)   (0.0163)          (0.0158)           (0.0008)

              + 0.0040(age) + 0.0001(twenmilk)
                (0.0016)       (0.0000+)
```

Note that I(group2) is a dummy variable that takes 1 if the observation is from Group2 (Walker) and 0 other wise. I(group3) is defined likewise. This parameterization for the dummy variables is achieved in R by choosing the 'treatment' contrast. If you do not specify it, R uses a different parameterization called the Helmert contrasts. You can see what contrasts were used with the `model.matrix()` function in R. See the program.

Several other models were proposed, typically adding higher order terms to the above model. Some people noticed a quadratic or curved trend in age. Note that the coefficient for age in the above model (model 1) is positive, as we would expect for women between the ages of 25 and 42. By adding an age^2 term to the model 1 we obtain

```
[Model 2]   bmd = 1.8955 + 0.0501 I(group2) + 0.0451 I(group3) + 0.0020(weight)
                  (0.4188)   (0.0159)          (0.0155)           (0.0007)

                    - 0.0576(age) + 0.0009(age^2) + 0.0001(twenmilk)
                      (0.0248)       (0.0004)          (0.0000+)
```

In model 2, there is a negative coefficient for *age* and a positive coefficient for age^2, which eventually provides an increasing trend in *bmd* as age increases. It would be inappropriate to extrapolate these trends to women over the age of 45, because *bmd* begins to decline after age 45.

The Type III Sum of Squares for these two models are found as follows. Notice that all the coefficients appear to be significant at the $\alpha = .05$ level of significance. Also, note that residual plots reveal no serious departures from model assumptions.

```
Anova Table (Type III tests)

Response: bmd
             Sum Sq Df  F value    Pr(>F)
(Intercept) 0.58336  1 150.5946 < 2.2e-16 ***
group       0.05377  2   6.9409  0.001597 **
weight      0.02784  1   7.1873  0.008783 **
age         0.02308  1   5.9586  0.016672 *
twenmilk    0.04224  1  10.9054  0.001392 **
Residuals   0.33701 87

Anova Table (Type III tests)

Response: bmd
              Sum Sq Df F value    Pr(>F)
(Intercept) 0.074882  1 20.4812 1.920e-05 ***
group       0.045747  2  6.2561  0.002906 **
weight      0.026948  1  7.3707  0.008011 **
age         0.019698  1  5.3876  0.022647 *
I(age^2)    0.022583  1  6.1767  0.014880 *
twenmilk    0.040148  1 10.9808  0.001348 **
Residuals   0.314429 86
```

(b) The R^2 and AIC for the above two models are computed in the table below. Model2 gives smaller AIC and higher R^2, and thus it is preferred to Model 1.

Model	R^2	AIC
Model 1	0.3257762	-246.7558
Model 2	0.3709557	-250.9325

The AIC values in this table are computed directly from the formula :

-2(log-likelihood) + 2(number of parameters),

using the mse of the more complex model (Model2) as a variance estimate.

The R `AIC(object)` function can give AIC values more easily but uses variance estimates from each model (so, it may not be appropriate to compare two models with these values).

Many students gave AIC values that came from the stepwise selection procedure in R. Those AIC values are based on a monotone transformation of the formula shown above and a variance estimate from the largest possible model in the search. That formula is

$$(\text{Sum of squared residuals}) + 2(\text{number of parameters})(\text{error variance})$$

One can compare two models with respect to the AIC values provided by the stepwise selection procedure in R. Models with smaller AIC values are preferred in that they have reduced bias without unnecessarily increasing variances of estimated means.

(c) Examine a scatter plot matrix or correlation matrix.

```
           bmd weight    age caffeine calcium growmilk teenmilk twenmilk
     bmd 1.000  0.277  0.226    0.141   0.179    0.158    0.153    0.309
  weight 0.277  1.000  0.021    0.174  -0.017    0.079    0.073   -0.001
     age 0.226  0.021  1.000    0.207   0.010   -0.028   -0.122    0.025
caffeine 0.141  0.174  0.207    1.000  -0.172    0.151    0.101    0.036
 calcium 0.179 -0.017  0.010   -0.172   1.000    0.268    0.277    0.472
growmilk 0.158  0.079 -0.028    0.151   0.268    1.000    0.846    0.606
teenmilk 0.153  0.073 -0.122    0.101   0.277    0.846    1.000    0.638
twenmilk 0.309 -0.001  0.025    0.036   0.472    0.606    0.638    1.000
```

[Associations among the explanatory variables]

Strong associations among the explanatory variables include: teenmilk and growmilk (0.846), teenmilk and twenmilk (0.638), growmilk and twenmilk (0.605), calcium and twenmilk (0.472). Also, group and age are possibly associated (more younger dancers, more older walkers).

[Associations between bmd and the non-exercise explanatory variables]

Explanatory variables possibly associated with bmd are: twenmilk (0.309), weight(0.277), age(0.226) and group (women involed in a program of aerobic walking or dancing tend to have higher levels of bmd than women who do not exercise).

(d) Since the coefficients for the group indicators (for Group2 and Group3) appear to be significantly positive, we can conclude that even after adjusting for effects of age, weight, caffeine consumption, calcium consumption, and milk consumption, there appear to significant benefits of moderate exercisein increasing *bmd* for women between the ages of 25 and 42.

We could conduct a multiple group mean comparison to confirm this conclusion. The following is the output from R for the first model. This indicates that essentially there is no difference in average *bmd* levels between the exercise groups (walker and dancer groups) while the non-exercise group has a significantly lower average *bmd* level than either of the exercise groups.

```
        Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = bmd ~ group + weight + age + twenmilk, data = bone)

Estimated Quantile = 2.384
95% family-wise confidence level
```

```
Linear Hypotheses:
          Estimate   lwr        upr
2 - 1 == 0  0.0518026  0.0128909  0.0907143
3 - 1 == 0  0.0509175  0.0132801  0.0885549
3 - 2 == 0 -0.0008851 -0.0401153  0.0383451

> summary(mc.bone1, test=univariate())

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = bmd ~ group + weight + age + twenmilk, data = bone)

Linear Hypotheses:
             Estimate Std. Error t value Pr(>|t|)
2 - 1 == 0  0.0518026  0.0163219   3.174  0.00208 **
3 - 1 == 0  0.0509175  0.0157874   3.225  0.00177 **
3 - 2 == 0 -0.0008851  0.0164555  -0.054  0.95723
```

You should not forget that this is an observational study. There are other dietary, environmental, behavioral, and genetic factors that could affect *bmd* and were not measured in this study. If some of these un-monitored factors vary across exercise groups, they could account for the difference in mean *bmd* levels across exercise groups. You cannot truly establish a cause and effect relationship from an observational study. Nevertheless, until future information becomes available, it would seem wise to advise women that a regular program of moderate exercise can help to increase *bmd*.

R code for performing the calculations and creating graphs for problem 3 is shown below. It is assumed that you have used the code previously listed to create a data frame called "bone".

```
#-------------------------------------------------------------#
# Change the group variable to a factor and specify           #
#  the treatment contrasts                                    #
#-------------------------------------------------------------#
bone$group <- as.factor(bone$group)

# Change the parameterization                                 #
#    (Otherwise, Helmert contrasts are used by default)       #
options(contrasts=c('contr.treatment','contr.ploy'))

# First fit a full model with every single explanatory term   #
bone.lm.full <- lm( formula= bmd ~ ., data=bone )

# Search for a best model using backward selection            #
bone.best <- step(bone.lm.full)

# Present the formula of the selected model                   #
formula(bone.best)

# Fit another model adding a quadratic term for age           #
bone.best2 <- lm( formula = bmd ~ group + weight + age +
                     I(age^2) + twenmilk,  data=bone )

#      Report the estimates and the SS                        #
#      For the claulation of different type of SS,            #
#      package "car" should be called                         #
model.matrix(bone.best)        # shows the X matrix used      #
summary( bone.best )           # shows estimated coefficients #
anova(bone.best)               # shows ANOVA with Type I   SS #
Anova(bone.best,type=c("III")) # shows ANOVA with Type III SS #
```

```
model.matrix(bone.best2)
summary( bone.best2 )
anova(bone.best2)
Anova(bone.best2,type=c("III"))

# Residual Plots and Model Assessment       #
# attach the MASS library (to use studres())#
# split the graphical window                #

library( MASS )
par( mfrow=c(2,2) )

plot( fitted(bone.best), studres(bone.best) )
abline( h=0 )
qqnorm(studres(bone.best), main="Studentized Residuals (Model 1)")
qqline(studres(bone.best))

plot( fitted(bone.best2), studres(bone.best2) )
abline( h=0 )
qqnorm(studres(bone.best2), main="Studentized Residuals (Model 2)")
qqline(studres(bone.best2))

#-----------------------------------------------------------#
#  (b) Report R^2 and AIC of the selected model             #
#-----------------------------------------------------------#
bone.summary1 <- summary( bone.best )
bone.summary2 <- summary( bone.best2 )

# lists the names of attributes of the summary object #
names(bone.summary1)

bone.summary1$r.squared # R^2 values                              #
bone.summary2$r.squared
AIC(bone.best)          # gives AIC based on the original formula #
AIC(bone.best2)
                        # Compute AIC manually using sig^2 from models#
sse1 <- deviance(bone.best)
sse2 <- deviance(bone.best2)
mse2 <- deviance(bone.best2) / bone.best2$df.residual
n <- nrow(bone)

loglik1 <- -.5*n*(log(2*pi)+log(mse2)) -.5*(sse1/mse2)
loglik2 <- -.5*n*(log(2*pi)+log(mse2)) -.5*(sse2/mse2)

aic1 <- -2 * loglik1 + 2*6
aic2 <- -2 * loglik2 + 2*7
c(aic1, aic2)

#-----------------------------------------------------------#
#  Draw a Trellis scatter plot matrix                       #
#-----------------------------------------------------------#

points.lines <- function(x, y){
    points(x, y)
    }
pairs(bone[,-1], panel=points.lines)
```

```
#  Correlation Matrix                          #
round( cor( bone[,-1] ), 3 )

#----------------------------------------------#
#  Pairwise comparison of group means          #
#  For the multiple comparison                 #
#  package "multcomp" should be called         #
#----------------------------------------------#

aov.bone1 <- aov(bmd ~ group + weight + age + twenmilk, data=bone)
aov.bone2 <- aov( bmd ~ group + weight + age +I(age^2) + twenmilk,data=bone)

mc.bone1 <- glht(aov.bone1,linfct=mcp(group="Tukey"))
mc.bone2 <- glht(aov.bone2,linfct=mcp(group="Tukey"))

confint(mc.bone1)
summary(mc.bone1, test=univariate())

confint(mc.bone2)
summary(mc.bone2, test=univariate())
```

PART II

1.(a) Since A is an orthogonal matrix, we have $A^TA = I$. From the properties of determinants, we have $|A| = |A^T|$ and $|A^TA| = |A^T||A|$. It follows that $|A|^2 = |A||A| = |A^T||A| = |A^TA| = |I| = 1$. Hence, $|A| = 1$ or $|A| = -1$.

(b) Since an idempotent matrix satisfies $AA = A$, we have $|A| = |AA| = |A||A| = |A|^2$. Consequently, $|A| = 1$ or $|A| = 0$.

2. Because B is a $k \times p$ matrix of rank k, the rows of B are a set of linearly independent vectors. This implies that $\mathbf{w} = B^T\mathbf{y} \neq \mathbf{0}$ for any $\mathbf{y} \neq \mathbf{0}$. Then, since A is positive definite, $\mathbf{y}^TBAB^T\mathbf{y} = \mathbf{w}^TA\mathbf{w} > 0$ for any $\mathbf{y} \neq \mathbf{0}$. Therefore, BAB^T is positive definite.

3. First show that if $A = P^TP$ for some nonsingular matrix P, then A is positive definite and symmetric.

Suppose $A = P^TP$ for some nonsingular matrix P. Then for any $\mathbf{y} \neq \mathbf{0}$, we have $\mathbf{y}^TA\mathbf{y} = \mathbf{y}^TP^TP\mathbf{y} = (P\mathbf{y})^TP\mathbf{y}$. Since P is nonsingular and $\mathbf{y} \neq \mathbf{0}$, it follows that $P\mathbf{y} \neq \mathbf{0}$. Consequently, $\mathbf{y}^TA\mathbf{y} = (P\mathbf{y})^TP\mathbf{y} > 0$ for any $\mathbf{y} \neq \mathbf{0}$. Therefore, A is positive definite. It follows immediately from $A = P^TP$ that A is symmetric.

Now show if A is a symmetric and positive definite matrix, then $A = P^TP$ for some nonsingular matrix P.

Suppose A is a symmetric, positive definite matrix. Then, we can use the spectral decomposition of A to obtain $A = UDU^T$, where

$$D = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_P \end{bmatrix},$$

is a diagonal matrix with the eigenvalues of A on the diagonal and the $i-th$ column of U is the eigenvector corresponding to λ_i. Since A is positive definite, all of the eigenvalues are positive and we can define

$$D = D^{1/2}(D^{1/2})^T, \quad \text{where} \quad D^{1/2} = \begin{bmatrix} \sqrt{\lambda_1} & 0 & \dots & 0 \\ 0 & \sqrt{\lambda_2} & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{\lambda_P} \end{bmatrix}.$$

Let $P^T = UD^{1/2}$, then we have $A = P^TP$. Furthermore, since A is positive definite $\mathbf{y}^TA\mathbf{y} > 0$ for any $\mathbf{y} \neq \mathbf{0}$. Then, $\mathbf{y}^TA\mathbf{y} = \mathbf{y}^TP^TP\mathbf{y} > 0$ which implies that $P\mathbf{y} \neq \mathbf{0}$ for all $\mathbf{y} \neq \mathbf{0}$. Consequently, P is nonsingular.

4. First create a matrix in R.

```
> V<-matrix(c(5.0,4.0,3.2,4.0,5.0,4.0,3.2,4.0,5.0),ncol=3)
> V
     [,1] [,2] [,3]
[1,]  5.0    4  3.2
[2,]  4.0    5  4.0
[3,]  3.2    4  5.0
```

(a) Evaluate eigenvalues and eigenvectors

```
> eigen(V)
$values:
[1] 12.4787754  1.8000000  0.7212246
```

```
$vectors
           [,1]          [,2]       [,3]
[1,] -0.5639516  7.071068e-01  0.4265661
[2,] -0.6032555  1.881091e-16 -0.7975480
[3,] -0.5639516 -7.071068e-01  0.4265661
```

(b) Evaluate the trace of V:

```
> sum(diag(V))
[1] 15
```

(c) Evaluate the determinant of V

```
> prod(eigen(V)$values)
[1] 16.2
```

(d) Evaluate the inverse of V

```
> solve(V)
              [,1]       [,2]       [,3]
[1,]  5.555556e-01 -0.4444444  0.0000000
[2,] -4.444444e-01  0.9111111 -0.4444444
[3,] -1.776285e-16 -0.4444444  0.5555556
```

5. (a)

$$\text{If } B = \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2} & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_n} \end{bmatrix} \text{ then } B\Sigma B \text{ is a correlation matrix.}$$

(b) Here is one possible way to construct a function to evaluate correlation matrices.

```
> corr.matrix <- function(x){
+      a<-diag( diag(x)^(-0.5) )
+      r<-a%*%x%*%a
+      return(r) }
```

(c) Compute a correlation matrix from the V matrix in problem 6.

```
> corr.matrix(V)
     [,1] [,2] [,3]
[1,] 1.00  0.8 0.64
[2,] 0.80  1.0 0.80
[3,] 0.64  0.8 1.00
```

6. (a) Compute determinants of A and B:

```
> A<-matrix(c(4,4.001,4.001,4.002),ncol=2)
> B<-matrix(c(4,4.001,4.001,4.002001),ncol=2)
```

```
> prod(eigen(A)$values)
[1] -1e-006

> prod(eigen(B)$values)
[1] 3e-006
```

(b) This is an example of a situation the solve( ) function fails to find the inverse of nearly singular matrices. It was put on this assignment to help you gain a healthy respect for round off errors and the inability of any computer to store numbers with infinite precision. You could easily compute the inverses of A and B without using a computer. If you want to use R for Windows, you can use the ginv function, which can compute an Moore-Penrose Generalized Inverse of a matrix. ginv function requires loading the **MASS** package. Note the difference in the inverses, even though A and B are nearly the same.

```
> ginv(A)
          [,1]     [,2]
[1,] -4002000  4001000
[2,]  4001000 -4000000

> ginv(B)
          [,1]     [,2]
[1,]  1334000 -1333667
[2,] -1333667  1333333
```

7. (a)

$$R^{-1} = \begin{bmatrix} \frac{1}{1-b^2} & \frac{-b}{1-b^2} \\ \frac{-b}{1-b^2} & \frac{1}{1-b^2} \end{bmatrix}$$

(b)

$$R^{-1} = \begin{bmatrix} \frac{1}{1-b^2} & \frac{-b}{1-b^2} & 0 \\ \frac{-b}{1-b^2} & \frac{1+b^2}{1-b^2} & \frac{-b}{1-b^2} \\ 0 & \frac{-b}{1-b^2} & \frac{1}{1-b^2} \end{bmatrix}$$

(c) By numerically obtaining the inverses of 4×4 and 5×5 versions of R, it is readily seen that the pattern suggested by the results in parts (a) and (b) holds for any order. For an $n \times n$ matrix of this type, the inverse of R is a tri-diagonal matrix of the form

$$R^{-1} = \begin{cases} \frac{1}{1-b^2} & i = j = 1 \text{ or } i = j = n, \\ \frac{1+b^2}{1-b^2} & i = j = 2, \ldots, n-1, \\ \frac{-b}{1-b^2} & |i-j| = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Consequently, the inverse of V is a tri-diagonal matrix of the form

$$V^{-1} = \begin{cases} \frac{1}{\sigma^2(1-b^2)} & i = j = 1 \text{ or } i = j = n, \\ \frac{1+b^2}{\sigma^2(1-b^2)} & i = j = 2, \ldots, n-1, \\ \frac{-b}{\sigma^2(1-b^2)} & |i-j| = 1, \\ 0 & \text{otherwise.} \end{cases}$$