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.

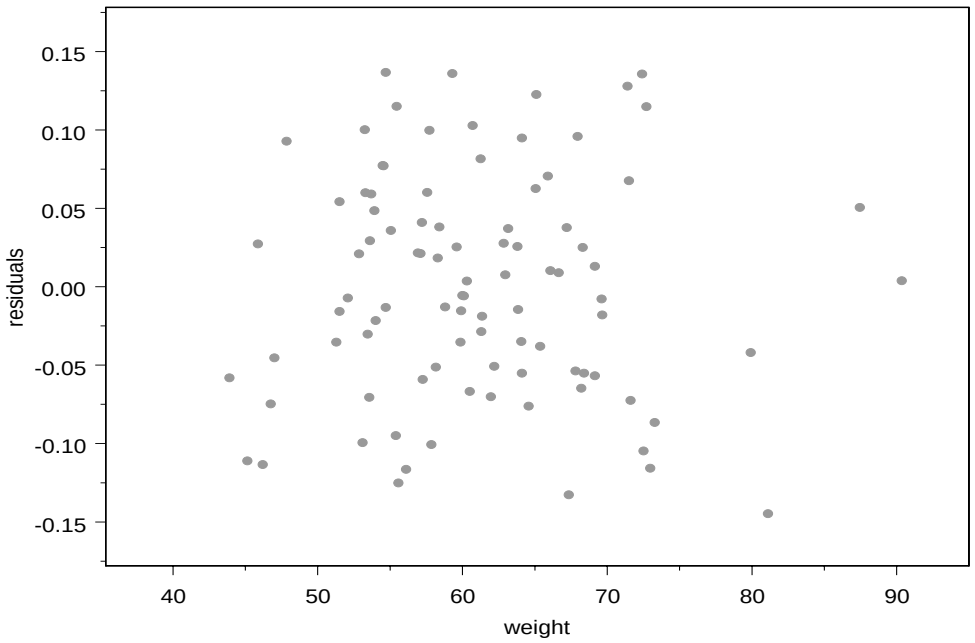
2. The least squares estimates of the coefficients yield

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

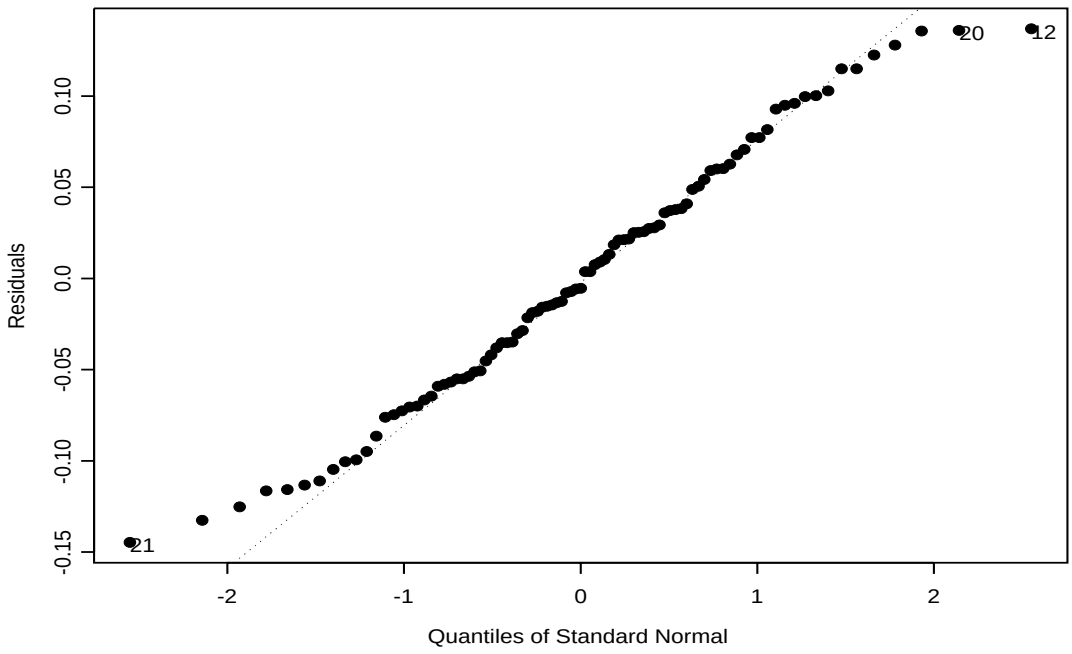
as the estimated regression line. Standard errors are given in parentheses underneath the estimated coefficients. The ANOVA table is

```
            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.



Your were expected to use the pull down menus in the Windows version of S-Plus to do the computations and create graphs. You could have also accomplished some of these tasks by entering code into the command 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("/your directory/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 S-Plus 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 S-Plus by choosing the 'treatment' contrast. If you do not specify it, S-Plus uses a different parameterization called the Helmert contrasts. You can see what contrasts were used with the `model.matrix()` function in S-plus. 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.

```
Type III Sum of Squares
          Df Sum of Sq    Mean Sq  F Value      Pr(F)
    group  2 0.0537739 0.02688695  6.94090 0.00159673
   weight  1 0.0278415 0.02784147  7.18731 0.00878257
      age  1 0.0230819 0.02308191  5.95862 0.01667191
 twenmilk  1 0.0422441 0.04224409 10.90536 0.00139242
Residuals 87 0.3370118 0.00387370

Type III Sum of Squares
          Df Sum of Sq    Mean Sq  F Value      Pr(F)
    group  2 0.0457468 0.02287341  6.25615 0.00290632
   weight  1 0.0269484 0.02694837  7.37070 0.00801066
      age  1 0.0196978 0.01969777  5.38757 0.02264659
 I(age^2)  1 0.0225830 0.02258303  6.17673 0.01488006
 twenmilk  1 0.0401476 0.04014761 10.98085 0.00134846
Residuals 86 0.3144288 0.00365615
```

(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 :

$$\text{-2(log-likelihood)} + 2\text{(number of parameters)},$$

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

The S-plus `AIC(object)` function (that is available in S-plus Version 6) 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 S-Plus. As discussed in e-mail sent to each of you, 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

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

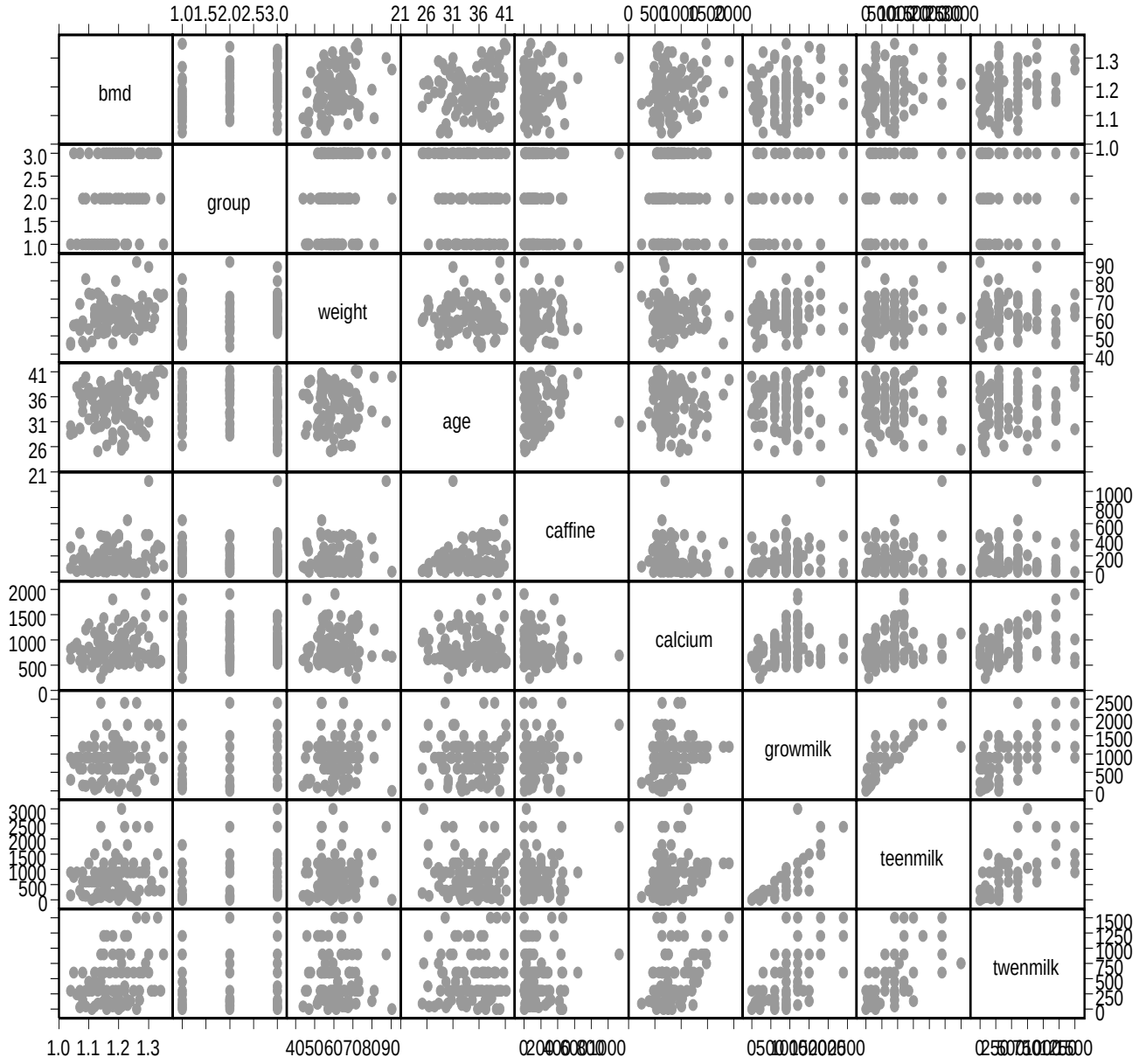
One can compare two models with respect to the AIC values provided by the stepwise selection procedure in S-Plus. Models with smaller AIC values are preferred in that they have reduced bias without unnecessarily increasing variances of estimated means.

(c) 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 S-plus 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.

```
     Estimate Std.Error Lower Bound Upper Bound
1-2 -0.051800    0.0163     -0.0907     -0.0129 ****
1-3 -0.050900    0.0158     -0.0886     -0.0133 ****
2-3  0.000885    0.0165     -0.0384      0.0401
```

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*.

S-Plus 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 +
                    age^2 + twenmilk,  data=bone )

#      Report the estimates and the SS                      #
model.matrix(bone.best)       # shows the X matrix used       #
summary( bone.best )          # shows estimated coefficients  #
anova(bone.best)              # shows ANOVA with Type I   SS  #
ssType3.lm(bone.best)         # shows ANOVA with Type III SS  #

model.matrix(bone.best2)
summary( bone.best2 )
anova(bone.best2)
ssType3.lm(bone.best2)

                     # Residual Plots and Model Assessment       #
library( MASS )      # attach the MASS library (to use studres())#
par( mfrow=c(2,2) )  # split the graphical window                #

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.summary)

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)
    lines(loess.smooth(x, y, 0.90))
}

par(din=c(7,7), pch=18, mkh=.15, cex=1.2, lwd=3)
pairs(bone, panel=points.lines)

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

  #-----------------------------------------------#
  #  Pairwise comparison of group means           #
  #-----------------------------------------------#
mc.bone <- multicomp( bone.best )
mc.bone <- multicomp( bone.best2 )
mc.bone
mc.bone2
plot(mc.bone)
```