1. In the handout on 1-way ANOVA, we considered the following problem:

An economist wants to test whether mean housing prices are the same regardless of which of 3 air-pollution levels typically prevails. A random sample of house purchases in 3 areas yields the price data below.

MEAN HOUSING PRICES (THOUSANDS OF DOLLARS):

	Pollution Level		
Observation	Low	Mod	High
1	120	61	40
2	68	59	55
3	40	110	73
4	95	75	45
5	83	80	64
Σ	406	385	277

You will now consider an alternative approach to this problem. Let Y = Mean housing price. Let X1 = 1 if pollution is low, 0 if pollution is moderate, -1 if pollution is high. Let X2 = 0 if pollution is low, 1 if pollution is moderate, and -1 if pollution is high. The data for the 15 observations can then be written as follows:

X1	X2	Y
1	0	120
1	0	68
1	0	40
1	0	95
1	0	83
0	1	61
0	1	59
0	1	110
0	1	75
0	1	80
-1	-1	40
-1	-1	55
-1	-1	73
-1	-1	45
-1	-1	64

An old but still trustworthy version of SPSS produced the following:

```
              * * * *  M U L T I P L E  R E G R E S S I O N  * * * *

Listwise Deletion of Missing Data

            Mean  Std Dev  Variance  Label

Y         71.200   23.752   564.171  House Price
X1          .000     .845      .714  Low Pollution
X2          .000     .845      .714  Moderate Pollution

N of Cases =    15

Correlation, Covariance, 1-tailed Sig, Cross-Product:

                  Y         X1         X2

Y             1.000       .459       .384
            564.171      9.214      7.714
                .          .043       .079
            7898.400    129.000    108.000

X1              .459      1.000       .500
               9.214       .714       .357
                .043       .          .029
             129.000     10.000      5.000

X2              .384       .500      1.000
               7.714       .357       .714
                .079       .029       .
             108.000      5.000     10.000

Equation Number 1    Dependent Variable..   Y   House Price

R Square             (1)
Standard Error       (2)

Analysis of Variance
                    DF      Sum of Squares      Mean Square
Regression           2             (3)              (4)
Residual            12             (5)              (6)

F =       (7)           Signif F =  .1887

------------------------------ Variables in the Equation -----------------------------

Variable              B        SE B     95% Confdnce Intrvl B       Beta         T  Sig T

X2             (8)          8.152709        (9)      23.563227    .206376      (10)  .4904
X1             (11)         (12)         -7.763227    (13)         .355820     1.227  .2435
(Constant)     (14)         5.764836    58.639502    83.760498                 (15)   .0000
```

a) Fill in the missing entries (1) - (15). The information that has been left in, and possibly the results from our earlier discussion of this problem, can serve as double-checks on your answer.

b) What does this tell us about the relationship between level of pollution and housing prices?

c) Comment on how your results compare with our earlier analysis of this problem. What do you think the differences or similarities can be attributed to?

d) After you've done this by hand, you can double-check your answers using Hw09.sps – PROVIDED you can fix some syntax errors in the program and fill in the missing regression card.

2. An educator believes that the more hours per weekday a student's best friend spends on homework (ZHWORK), the more hours the student will tend to spend on homework (XHWORK). She also thinks that XHWORK will be affected by the student's socio-economic status (measured on a scale called XBBSESRW) and by whether or not the student is in the academic track (XTRKACAD - which is coded 1 if the student is in the academic track, 0 otherwise.) Using an old mainframe version of SPSS^X, she obtains the following results:

```
            MEAN  STD DEV  LABEL

 XHWORK    3.968    2.913  TIME ON HOMEWORK PER WEEK
 ZHWORK    3.975    2.930  TIME ON HOMEWORK PER WEEK
 XTRKACAD   .321     .467  X IN ACADEMIC TRACK
 XBBSESRW  -.071     .686  SES COMPOSITE SCALE SCORE

 CORRELATION:

              XHWORK     ZHWORK   XTRKACAD   XBBSESRW

 XHWORK        1.000       .326       .303       .179
 ZHWORK         .326      1.000       .222       .152
 XTRKACAD       .303       .222      1.000       .285
 XBBSESRW       .179       .152       .285      1.000

 STANDARD ERROR      2.65806

 ANALYSIS OF VARIANCE
                    DF      SUM OF SQUARES      MEAN SQUARE
 REGRESSION          3         13219.80246       4406.60082
 RESIDUAL         9299         65699.89382          7.06526

 F =     623.69935       SIGNIF F =  .0000

------------------- VARIABLES IN THE EQUATION ------------------
 VARIABLE              B        SE B       BETA         T  SIG T

 XBBSESRW        .320998     .042126    .075555     7.620  .0000
 ZHWORK          .263356     .009690    .264956    27.180  .0000
 XTRKACAD       1.390122     .062694    .222876    22.173  .0000
 (CONSTANT)     2.496854     .049167               50.783  .0000
```

a) Compute the following:

1. The sample size
2. R^2 (try calculating this at least 2 different ways and make sure your results are consistent)
3. The 95% confidence interval for XTRKACAD
4. The covariance of XBBSESRW and XHWORK (i.e., $s_{XBBSESRW,XHWORK}$)
5. The cross-product of XBBSESRW and XTRKACAD (i.e. $XP_{XBBSESRW,XTRKACAD}$, or, using our other notation, $SP_{XBBSESRW,XTRKACAD}$
6. $M_{XTRKACAD,\ ZHWORK}$ (i.e. Σ (XTRKACAD * ZHWORK))

b) Briefly discuss the implications of these findings. Answer such questions as, How much time per weekday does the average student spend on homework? What percentage of the students are in the academic track? Does it matter to students how much their friends study? If so, how much? Do students in the academic track tend to study more than students not in the academic track (once over variables are taken into account)? What is the most "important" determinant of how much a student studies? How much of the variation in time spent studying can these 3 variables account for, and how much variation must be due to other factors not yet considered?

c) To the best of your ability, discuss any theoretical or statistical problems you see with the above model specification. In particular, what problems might there be with using ZHWORK as an independent variable when XHWORK is the dependent variable? (I don't really expect you to get this, but see what you can come up with. This serves as a lead-in to a much broader discussion that occurs in later statistics classes.)

3. Construct ANOVA tables based on the following information. Also, report the value of R^2 if it is not already given in the problem.

a) Dependent variable: Occupational prestige
Independent variables: Education, IQ, Father's occupational prestige
n = 100, R^2 = .3, MSE = 10

b) Number of independent variables: 10.
n = 50, F = 5, SSR = 80.

c) Number of independent variables: 5.
n = 100, R^2 = .4, s_y = 1.

4. (Optional; this will be good extra practice if the calculations required in problem #2 are not enough for you to feel comfortable with the computational techniques.) Using the raw data presented in problem 1 (i.e. the data for X1, X2, and Y), compute the M, XP, s, and r matrices. (Once you have computed one quantity, you are free to use it when computing anything else, e.g. once you have the M matrix you don't have to keep on using the raw data.) Since SPSS reports everything except the M matrix, it ought to be pretty easy to double-check your answers.

5. Consider again the following problem (adapted from Hays, 4th edition, p. 607): In a study of the origins of gender stereotyping of young girls, a random sample of 35 intact families was taken, in which there was an oldest (or only) girl in the 9th grade. The father answered a questionnaire about his interest in sports and received a score X1. The mother answered a similar questionnaire and received a score X2. The physical education instructor of each girl rated her on general athletic ability, and this was used as variable X3. The dependent variable Y was the girl's own score on a questionnaire on interest in sports. We will use all the variables this time. Hw09.sps has everything you need, except you will have to add the regression command.

a. Regress Y on X1, X2, and X3.

b. Get the plot of Y on X1, X2, X3, and $\hat{Y}$. To do this, the last parameter on the regression card should be

/SCATTERPLOT (Y, X1) (Y, X2) (Y, X3) (Y, *PRED) .

c. Which variable is most strongly correlated with Y? Who seems to have the stronger influence on the daughter, the father or the mother?

d. What does R^2 equal? What is the standard error of the estimate?

e. If X1 = X2 = 30 and X3 = 10, what is the predicted value for Y?

f. For each of the following, if α = .05, should we reject or not reject the null hypothesis?

(i) H_0: $\beta_1 = 1$ H_A: $\beta_1 <> 1$	(ii) H_0: $\beta_2 = 0$ H_A: $\beta_2 <> 0$	(iii) H_0: $\beta_3 = 1$ H_A: $\beta_3 <> 1$

6. Use Stata to confirm your answers to two or more of the following.

a. Problem 1a. You can use the file hw09-1.dta.

b. Problems 2a2 and 2a3. You can use the file hw09-2.dta. Or, you can create a pseudo-replication of the data, which is what I did. See the handout on using Stata for OLS regression. Your numbers will differ slightly because of rounding error.

c. Problem 4. You can do this even if you did not work the problem by hand. Again you can use hw09-1.dta. See the handout on using Stata with Multiple Regression & Matrices.

d. Problem 5a. You can use the file hw09-5.dta.