Solutions To Homework Assignment 3

Warmup Problems

(1) Birthweight and Smoking:

(a) The Bonferroni procedure says that when performing m hypothesis tests, to have a simultaneous significance level of α we should for each individual test use $\alpha^* = \alpha/m$. In our case since there are $k = 4$ groups and $k(k-1)/2 = 4 * 3/2 = 6$ possible mean comparisons so we use $\alpha^* = .05/6 = .00833$.

(b) The intervals calculated using the Bonferroni procedure would be wider. Intuitively this is because we are insisteing on being more sure about each individual comparison so we have to include more possible values in our intervals. Mathematically, the intervals are the same except that for the unadjusted ones we use $t_{.025,23} = 2.069$ while for the adjusted intervals we use $t_{.0083/2,23} = 2.886$ I got this latter value using the inverse ttail command in STATA. Don't forget that for the confidence intervals you need to divide the significance level by 2 to get the critical t value!

(c) To check our results correcting for multiple comparisons we we simply compare the original p-values to our new cutoff. If we let N stand for never smokers, F stand for former smokers, L stand for current low smokers and H stand for current high smoking mothers then the (two-sided) p-values from homework two tests were:

N vs F: .5359
N vs L: .0199
N vs H: .0037
F vs L: .1112
F vs H: .0314
L vs H: .5521

The only p-value less than .0083 is that for comparing the mean birthweight of babies of non-smoking mothers to that of high-smoking mothers. When we had not adjusted for multiple comparisons non-smoking mothers were different from current low-smoking mothers and former smokers were different from current high smokers but these differences become insignificant after doing the Bonferroni adjustment.

(2) Regression Basics:

(a) The simple linear model is $Y = \beta_0 + \beta_1 X + \epsilon$. It is our model for all elements of the population and says that we expect Y to be linearly related to X with the typical value of Y for a given X being $\beta_0 + \beta_1 X$ but allowing for individual variability, ϵ, about that typical value. The simple linear regression equation is $E(Y|X) = \mu_{Y|X} = \beta_0 + \beta_1 X$. It gives the **average** value of Y associated with a given X in the **population.** The estimated simple linear regression equation is $\hat{Y} = b_0 + b_1 X$. It gives our best **estimate** of the simple linear regression equation based on **sample** data.

(b) The least squares estimate is "good" in the sense that it chooses the line that goes **closest** (in terms of of squared vertical distance) to all the data points. A model is doing a good job if its predictions for Y are "close" to the actual values. If we further assume that the errors or individual variation is normally

distributed then the least squares estimate is also the maximum likelihood estimate. In other words it gives us the value of the slope and intercept most likely to have generated the data points we observed.

(c) There are many variables that might be useful in predicting a person's cholesterol level. Examples include dietary fat intake, numer of calories eaten per day, numer of hours exercised per day, whether there is a family history of heart disease or high cholesterol, etc. Most of the variables I listed will be positively correlated with cholesterol level. Eating high calorie foods with lots of fat will tend to raise your cholesterol and all these things are associated with being overweight. Having a family history of cholesterol or heart problems is also a risk factor. However healthy exercise would help to reduce your cholesterol level and would hence be negatively correlated. If we assume a linear relationship between these variables and cholesterol, one model might look like

$$Cholesterol = \beta_0 + \beta_1 Fat + \beta_2 Weight + \beta_3 Exercise$$

There are many other possibilities. In this problem it was not necessary to try to fill in numbers for the various parameters. In fact, if you don't have data, you can't really do that. Also note that the coefficients can be either positive or negative so you don't have to put + or - signs to indicate which way the relationship goes. You can just let the data tell you.

It is also not necessarily the case that all the relationships will be linear. For instance, there is a point beyond which more exercise will not help much in reducing your cholesterol. It probably doesn't matter whether I jog 100 miles per week or 105. Similarly there are probably limits on the effects of other variables. We will learn how to model such shapes in more detail later.

(3) Midterms Make You Sick (From Midterm 1, 2007): As noted on the assignment this is great practice for our midterm and you should try doing it yourself before reading the solutions!

(a) Plugging in the given numbers to the basic formulas gives

$$b_1 = \frac{SCP}{SSX} = \frac{-200}{40} = -5$$

and

$$b_0 = \bar{Y} - b_1\bar{X} = 25 - (-5)(5) = 25 + 25 = 50$$

Be careful of your signs!

(b) The intercept, b_0, is the average value of Y when $X = 0$. Here that means that b_0 is the average number of statisticitis cases (per 100 students) when there are 0 days until the exam–i.e. on the day of the exam istself! Here $b_0 = 50$ means that on average 50 out of every 100 students, or half the class, has statisticitis on the day of the exam. Ouch! Note that it is important to include the actual numeric values of b_0 and b_1 in your answers or you miss the real story!

The slope, b_1, gives the average change in Y for every one unit change in X. Here a 1 unit change in X is just 1 day so $b_1 = -5$ means that for every additional day away from the exam we are there are 5 **fewer** cases of statisticitis per 100 students–the more time there is left to study, the smaller the number of students who are sick. You could also reverse this and say for every day **closer** to the exam you get, on average the number of statisticitis cases goes up by 5 students per 100 or 5%. Note that you can incorporate your units into your explanations. You do not have to give them separately. If you wanted to, here b_0 would be in units of number of cases of statisticitis per 100 students, the same units as Y, and b_1 would be in number of cases per 100 students per day, the units of Y over the untis of X.

(c) This piece is about knowing the relationships between the various regression quantitites. First, as indicated in the problem statement

$$SSR = b_1 * SCP = (-5)(-200) = 1000$$

The next step is to note that that the sums of squares sum: $SSE + SSR = SST$. Using the value of SST given in the problem statement plus the SSR we just computed gives

$$SSE = SST - SSR = 1250 - 1000 = 250$$

R^2 is the fraction of variability explained using time until the exam to predict number of statisticitis cases. We have

$$R^2 = \frac{SSR}{SST} = \frac{1000}{1250} = .80 = 80\%$$

Finally we need the root mean squared error which is

$$RMSE = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{250}{12-2}} = \sqrt{25} = 5$$

(d) The correlation squared is just R^2 so we can take the square root of the quantity we found in part (c). However we have to be careful about the sign. In this case there is a negative relationship between X and Y (negative slope) so we need to take the negative square root. Our correlation is

$$\hat{\rho} = r = -\sqrt{.8} = -.8944$$

This correlation is quite close to -1 so there is a strong negative relationship between time until the exam and number of statisticitis cases. This makes sense–the further you are from the exam the less worried students will be and hence the less likely they will be to get sick!

(e) This question just asks to look at R^2, the fraction of variability in statisticitis cases that is explained by time until the exam. Here our predictor variable explains 80% out of a possible 100% of the variability which is quite good but not absolutely stupendous. It would be slightly more accurate to look at R^2-adjusted which is

$$R^2_{adj} = 1 - \frac{SSE}{SST}\frac{n-1}{n-2} = 1 - (.2)(\frac{11}{10}) = .78$$

but the difference is fairly minor and I did not ask people to do the calculation on the exam.

(f) To see how good the predictions are we need to look at the errors we make and compare them to the values we are trying to predict. In other words we compare the root mean squared error to the Y values. Here $RMSE = 5$ which means that typically when we use time until the exam to predict number of statisticitis cases we are off by 5 cases (per 100 students). From the printout at the beginning of the problem we see that in this data set we had an average of $\bar{Y} = 25$ cases per 100 students on the days we measured so our typical error is roughly $5/25 = 20\%$. This is a fairly substantial error so I would say that the model is not making great predictions. However this is a bit of a judgment call depending on how important you think it is to get a precise estimate of the number of sick students.

(g) Our estimated regression equation is

$$\hat{Y} = 50 - 5X$$

We want to make a prediction for a week before the exam, namely X=7, and 20 days before the exam, which is X=20. This gives

$$\hat{Y}_7 = 50 - 5(7) = 15$$

and

$$\hat{Y}_{20} = 50 - 5(20) = -50$$

The first of these predictions is perfectly reasonable–a week before the exam we have 15 students sick per 100 or 15% of the class–a few students are starting to get worried. The second prediction is not reasonable. You can not have a negative number of students sick. What has happened is that we are predicting outside the range of our data. Note that our average X value in the data set was 5 which is much lower than 20. It really only makes sense to try to predict statisticitis close to the exam. If the exam is too far in the future students won't be worrying about it yet so the model we fit will not apply very well.

(4) Height and Weight:

Note: You do not need STATA for this problem, and in fact it would be basically impossible to use it since you do not have the original data!

(a) I would expect that height would have a strong positive relationship with weight, that exercise would have a negative relationship (the more you work out the more weight you lose) and that within this sample where people have mostly achieved their adult growth the relationship with age would be weak. Looking at the correlations (remember that covariance doesn't say much about the strength of the relationship because it is unit dependent) I would expect that height is predictor A, exercise is predictor B, and age is predictor C. Correlations that are high in absolute value indicate strong relationships while those near 0 indicate weak relationships. Here predictor B (exercise) has the strongest relationship with a correlation of -.75 (the negative sign is irrelevant) while age with a correlation of .1 is the weakest relationship.

(b) First, b_0 gives the weight when X=0, that is when a person is 0 inches tall (a not very practical situation unless you count the moment after conception)! For this to make sense b_0 must be in units of pounds to match Y. The intercept will always have the same units as the response variable. Here the value of the intercept suggests that a 0 inch tall person weighs 40 pounds which doesn't make much sense. We will see the reason for this in part (e). The slope, b_1, gives the increase in weight associated with each additional inch of height. Since b_1 gets multiplied by a number of inches to produce a weight value, it must be in units of pounds per inch. We see that for every additional inch of height a person's weight increases by, on average, 2 pounds. (Note that the relationship need not be completely causal!)

(c) The predicted weight level is obtained by plugging in X=60 since 5 feet corresponds to 60 inches. Be careful of your units!!!

$$\hat{Y} = 40 + 2(60) = 160$$

Thus we would expect such a person to be 160 pounds (which seems rather high! There may have been some overweight people in this study skewing the data.)

(d) Similarly, if we plug in X=0, we get a predicted weight of 40 pounds. This is just the intercept we interpreted in part (a). For an 18 inch baby we get a predicted weight of 76 pounds. This is obviously nonsense. The 0 inch person should weigh virtually nothing and a typical baby ways 6-8 pounds.

(e) The fact that the model gives nonsensical results in part (d) suggests that it breaks down for very low values of X. Often a linear model only holds in a particular range. For instance, the relationship between height and weight may be very different before and after the puberty growth spurt. We would not expect this particular model to hold all the way down to X=0. Also, we probably had no data with value of X near

0 since we never record weight for people who are near 0 inches tall. If this data set only included adults we would be even worse off predicting for small heights. Remember that it is dangerous to predict outside the range of your data!

(5) The Tranquilizing Effect of Homework?:

(a) The STATA scatterplot is given in the accompanying graphics file. The Y values get higher as the X values get higher meaning that higher cortisol levels are associated with higher sedation (and presumably a stronger neurological effect of the drug.)

(b) The STATA and SAS printouts of the correlation and covariance are given below. The correlation, which is not unit dependent, gives the more reliable indication of the strength of the relationship. Here, the correlation is .96 which is very high indeed, indicating a very strong and positive relationship between cortisol and sedation. This is consistent with what we say from the scatterplot. Note that STATA gives the correlation of each variable with itself (which is of course 1) and the covariance of each variable with itself (which is just the variance of that variable) on the main diagonal of its printout. SAS would have done this too if I had just used the "var" subcommand but because I split it into a "var" piece and a "with" piece it gave me only the correlation/covariacne between the two variables. Make sure you know which number to look at on your printouts!

```
IN STATA:

. corr sedation cortisol
(obs=7)

             | sedation cortisol
-------------+------------------
    sedation |   1.0000
    cortisol |   0.9615   1.0000


. corr sedation cortisol, cov
(obs=7)

             | sedation cortisol
-------------+------------------
    sedation |  168.238
    cortisol |  23.2976  3.48952

IN SAS:

proc corr cov data = work.hw3;
var sedation;
with cortisol;
run;

                                        The SAS System

                                    The CORR Procedure

                                1 With Variables:    cortisol
```

1 Variables: sedation

Covariance Matrix, DF = 6

		sedation
cortisol	cortisol	23.29761905

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
cortisol	7	9.35714	1.86803	65.50000	6.60000	11.80000	cortisol
sedation	7	54.28571	12.97066	380.00000	32.00000	66.00000	sedation

Pearson Correlation Coefficients, N = 7
Prob > |r| under H0: Rho=0

	sedation
cortisol	0.96154
cortisol	0.0005

(c) The printouts of the simple linear regression in both STATA and SAS are shown below. We see that $b_0 = -8.19$ which means that subjects with no blood cortisol would have a negative sedation scores. Obviously this doesn't make real world sense but 0 does not seem to be a physically plausible value for cortisol level so we shouldn't worry about this too much. In fact the intercept is not significantly different from 0 meaning it could be that 0 cortisol corresponds to 0 sedation. The estimated slope $b_1 = 6.68$ points/μg/dL which means that that a 1 μg/dL increase in blood cortisol is associated with an increase of 6.68 points on the sedation scale on average. To really interpret this it would be helpful to know the range of both normal cortisol values and the sedation scale used!!

IN STATA:

```
. reg sedation cortisol

      Source |       SS       df       MS              Number of obs =       7
-------------+------------------------------           F(  1,     5) =   61.27
       Model |  933.271821     1  933.271821           Prob > F      =  0.0005
    Residual |  76.1567506     5  15.2313501           R-squared     =  0.9246
-------------+------------------------------           Adj R-squared =  0.9095
       Total |  1009.42857     6  168.238095           Root MSE      =  3.9027

------------------------------------------------------------------------------
    sedation |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cortisol |   6.676446   .8529243     7.83   0.001     4.483935    8.868958
```

```
       _cons |  -8.186748   8.116109    -1.01   0.359    -29.04987    12.67637
------------------------------------------------------------------------------
```

IN SAS:

```
proc reg data = work.hw3;
model sedation = cortisol;
run;
```

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: sedation sedation

Number of Observations Read	124
Number of Observations Used	7
Number of Observations with Missing Values	117

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	933.27184	933.27184	61.27	0.0005
Error	5	76.15673	15.23135		
Corrected Total	6	1009.42857			

Root MSE	3.90274	R-Square	0.9246
Dependent Mean	54.28571	Adj R-Sq	0.9095
Coeff Var	7.18925		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-8.18675	8.11611	-1.01	0.3594
cortisol	cortisol	1	6.67645	0.85292	7.83	0.0005

(d) There is a significant relationship between cortisol and sedation if $\beta_1 \neq 0$. We can test this formally using an overall F test (which asks if the model as a whole is useful.) In this problem we have

$H_0 : \beta_1 = 0$–there is no (linear) relationship between cortisol level and sedation. The model is not useful for explaining variation in sedation.
$H_A : \beta_1 \neq 0$–there is a relationship between blood cortisol and sedation; the model is useful overall or is a significant improvement over not using cortisol level to help predict sedation.

Our test statistic is $F_{obs} = MSR/MSE = 61.27$, the ratio of explained to unexplained variability. I got the

F statistic from the STATA/SAS printouts above. The corresponding p-value is $P(F_{1,5} \geq 61.27) = .0005$. Since this p-value is much smaller than $\alpha = .05$ we reject the null hypothesis and conclude that there is a significant relationship between blood cortisol level and sedation.

(e) If we want to know whether the model makes good predictions we need to compare the root mean squared error with the values we are trying to predict. here $RMSE = 3.90$ which says on average we make an error of just under 4 points when we use cortisol level to predict sedation score. The sedation scores we are trying to predict range from 32 to 66 with an average of 54.3. In percentage terms this is an error of roughly $3.9/32 = .122 = 12.2\%$ to $3.9/66 = .059 = 5.9\%$ with an average error of $3.9/54.3 = .0718 = 7.18\%$. Overall this doesn't seem like a horrible error unless we need to be really precise in predicting the sedation scores. However it is not as impressive as the R^2 value of 92.5% seemed. Remember that we learn different things from the various measures of model performance. My favorite 3 are the F test, the RMSE and the R^2 value. Together they give you a very good idea of how the model is working overall.

Problems To Turn In

(a) No solution needed :)

(b) The Tukey procedure is specifically designed for doing all pairwise comparisons of means after an ANOVA while the Scheffe procedure is designed to simultaneously cover ALL possible linear combinations (of which there are of course infinitely many). Because the Scheffe procedure includes all of the Tukey comparisons and many more it has to be more conservative so uses a critical value that produces wider confidence intervals and bigger p-values. Both of these procedures try to improve on Bonferroni by taking advantage of the fact that all the comparisons involve the same set of means being compared to each other in different ways so that you are not really doing a set of completely independent tests. You would use Tukey if you only cared about pairs of means and Scheffe if you cared about comparing more complicated sugroups of your data set.

(c) Most of the procedures in the handout take the point of view that you want to overall have only an α chance of falsely rejecting any of your null hypotheses. As a result you have to be much more conservative on each of the individual tests. In contrast, the omnibus test and false discovery rate procedure are more tilted towards making sure you correctly reject the null when it is false, possibly at the expense of making some mistakes, and therefore are not as conservative about the individual tests. With the omnibus approach you don't correct α at all but rather see how many times you reject and comapre this to the number of false positives you would expect given the number of tests you did. For example, suppose $\alpha = .05$ and you do 20 tests. You would expect if the null were true in all 20 cases that you would get 1 rejection anyway just by chance ($.05 = 1/20$). So if you rejected in, say, 5 of the 20 tests, you would believe that 4 out of 5 of your significant results were real. The false discovery rate procedure lets you directly control what fraction of your "significant" results are likely to be false. In both cases you will make some mistakes and not be 100(1-α)% sure of all your results simultaneously but you will both have a good idea of how many mistakes you are making and you will be much more likely to detect the situations where the alternative is true. False discovery rates and similar procedures have become particularly popular in genetics where people are testing the relationships of thousands of genes to a particular disase outcome. The Bonferroni procedure or its relatives are totally impractical here because the resulting significance level gets so small. However if you could start with a list of 10,000 genes and reduce it to a candidate set of 20 genes of which maybe only 5 were right you would be very happy because you could then test those 20 in other ways in the lab rather than having to examine all 10,000 in detail.

(d) Since there are 5 different locations there are "5-choose 2" or 10 possible pairwise comparisons. Therefore instead of using $\alpha = .05$ we need to use $\alpha^* = .05/10 = .005$ as the significance level for our individual confidence intervals. We also have a total of n = 125 data points and k = 5 groups so the degrees of freedom for our confidence intervals are n-k = 120. Where originally we would have used $t_{\alpha/2,n-k} = t_{.025,120} = 1.98$

we now need $t_{\alpha^*/2,n-k} = t_{.0025,120} = 2.86$. I got the latter value from STATA using the following command:

```
. display invttail(120, .0025)
2.8598648
```

The mean and standard error estimates remain the same whether we are doing a Bonferroni correction or not. The only thing that changes is the value of "t" we are using. Therefore to get the percentage increase in width of the confidence interval we just need to take the ratio of the two critical t-values. Here we have $2.86/1.98 = 1.44$ so the intervals will be 1.44 times as wide or a whopping 44% wider.

(e) The STATA printout with the Bonferroni adjusted p-values is as follows:

```
. oneway particle location, tabulate bonferroni

           |         Summary of particle
   location |        Mean   Std. Dev.       Freq.
------------+------------------------------------
          A |        14.2   7.0945989          25
          C |       19.08   6.9277221          25
          P |       37.88   6.7473452          25
          W |       36.76   5.0767444          25
          Y |    41.43006   8.8187441          25
------------+------------------------------------
      Total |   29.870012   13.050315         125

artlett's test for equal variances:  chi2(4) =   7.0030  Prob>chi2 = 0.136

                    Comparison of particle by location
                              (Bonferroni)
Row Mean-|
Col Mean |          A          C          P          W
---------+--------------------------------------------
       C |       4.88
         |      0.156
         |
       P |      23.68       18.8
         |      0.000      0.000
         |
       W |      22.56      17.68      -1.12
         |      0.000      0.000      1.000
         |
       Y |    27.2301    22.3501    3.55006    4.67006
         |      0.000      0.000      0.769      0.205
```

(e.i) Based on this printout air quality in the class room with the air conditioner (A) is not significantly different from that in the cafeteria (C) and the air quality in the parking lot (P), window-open classroom (W) and yard (Y) are all not different from each other. We get this simply by noting that all the corresponding p-values in the above table are above $\alpha = .05$. Thus it seems that we really have two groups of pollution levels. From the table of means above, the air-conditioned classroom and the cafeteria are better (lower pollution) that the parking lot, yard and window open classroom. Our best guess (based on the sample mean estimates) is that the order (from lowest to highest pollution) is A, C, W, P, Y but we can only be

95% sure that the (A,C) pair is significantly better than the (W,P,R) triplet.

(e.ii) The p-values for the pairwise comparisons of means from the air quality problem in homework 2 were as follows. Note that A = the air-conditioned classroom, C = the cafeteria, P - the parking lot, W = the window-open classroom and Y = the yard:

A vs C: .0156
A vs P: .0000
A vs W: .0000
A vs Y: .0000
C vs P: .0000
C vs W: .0000
C vs Y: .0000
P vs W: .5745
P vs Y: .0769
W vs Y: .0205

According to these results the air-conditioned classroom had significantly better air-quality (p-value = .0156) than the cafeteria and the window-open classroom had significantly better air-quality than the yard (p-value = .0205). Thus we lost two significant results when we corrected for the multuple comparisons.

(e.iii) The reason that STATA adjusts the p-values is that it doesn't know at what (overall) significance you want to perform your test. In general, STATA gives p-values and leaves you to decide whether or not they are significant rather than giving you the α value. The Bonferroni procedure we learned in class says to take the original p-values for your m tests and to reject the null hypothesis if $p < \alpha^* = \alpha/m$. This is equivalent to rejecting if $m * p < \alpha$. In other words if we multiply the original p-values by the number of tests we can simply compare to our overall desired p-value. Thus STATA gives the p-values multiplied by the number of tests and then you can compare them to any desired overall significance level.

e.iv) Optional Bonus: To get the original p-values from the Bonferroni adjusted p-values all you have to do is divide by the number of tests. This works fine in most cases. However if the original p-value was large enough that when multiplied by the number of tests the result was greater than 1 then STATA just reports 1 as the adjusted p-value since you can't have a probability greater than 1. This means that if the Bonferroni adjusted p-value is 1 you may not recover the exact original p-value when you do the division. However this is unlikely to be a problem as a test where the adjusted p-value is 1 (as unsignificant as it is possible to be) was not likely to be significant on the original scale either!

(f) The Holm procedure says to rank the p-values in decreasing order and then do the Bonferroni procedure for the tests consecutively counting only how many tests you have left to do. If you start with m comparisons then for the first test you will use α/m, for the second test you will use $\alpha/(m-1)$ and so on. In this problem six of the p-values are 0 to four decimal places which is all the computer packages give. However we can still rank them by their F statistics or, since in this example the group sizes are all the same, by the magnitude of the mean differences. The bigger the F statistic or mean difference, the smaller the p-value. Thus our tests in order are A vs Y, A vs P, A vs W, C vs Y, C vs P, C vs W, A vs C, W vs Y, P vs Y and P vs W. With an overall $\alpha = .05$ and $m = 10$ tests we need to compare these p-values to respectively .05/10 = .005, .05/9 = .0056, .05/8 = .00625, .05/7 = .0071, .05/6 = .0083, .05/5 = .01, .05/4 = .0125, .05/3 = .0167, .05/2 = .025 and .05. The first 6 tests are significant since .0000 is less than all of the α^* values. The 7th test has a p-value of .0156 and has to be compared to the value $\alpha^* = .0125$. This is very close but this test is not quite significant and neither are any of the subsequent ones. Thus we actually get the same 6 significant results as we got with the Bonferroni procedure. However we note that the Bonferroni procedure used $\alpha^* = .005$ for all the tests and therefore was not even close to rejecting

for the 7th test (air conditioned room vs cafeteria.) If we had been using a slightly larger overall α the Holm procedure would have detected this result as significant whereas the Bonferroni procedure would not have. In general the Holm procedure is less conservative but it doesn't always result in a different conclusion.

(g) Optional Bonus: The SAS printouts for the Tukey and Bonferroni procedures are shown below. In this case they produce the same answer in terms of which groups are significantly different. However if you look at the number labeled "least significant difference" you will see that the value is greater for the Bonferroni procedure (5.69) than for the Tukey procedure (5.51). That number is the smallest difference in means for which the procedure would conclude two groups are significantly different. Here the Tukey procedure is slightly less conservative than the Bonferroni procedure but not all that much. That difference grows as the number of groups being compared gets bigger. How the Hom procedure fits into this is less obvious. The Bonferroni and Tukey procedures use the same significance cutoff for all the comaprisons while the Holm procedure uses a less and less stringent criterion as you move through the tests. Depending on how many significant or close to significant tests there are this tradeoff could go either way. Here all three tests give the same conclusions.

```
The commands in SAS for obtaining the Bonferroni and Tukey comparisons are

proc anova data = tmp1.hw3;
class location;
model particle = location;
means location/bon;
run;

proc anova data = tmp1.hw3;
class location;
model particle = location;
means location/tukey;
run;

The corresponding output is

Bonferroni:

                                The ANOVA Procedure

                        Bonferroni (Dunn) t Tests for particle

 NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher
                            Type II error rate than REGWQ.


                        Alpha                             0.05
                        Error Degrees of Freedom           120
                        Error Mean Square              49.47938
                        Critical Value of t             2.85986
                        Minimum Significant Difference   5.6899


                Means with the same letter are not significantly different.
```

Bon Grouping	Mean	N	location
A	41.430	25	Y
A			
A	37.880	25	P
A			
A	36.760	25	W
B	19.080	25	C
B			
B	14.200	25	A

Tukey:

The ANOVA Procedure

Tukey's Studentized Range (HSD) Test for particle

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	120
Error Mean Square	49.47938
Critical Value of Studentized Range	3.91694
Minimum Significant Difference	5.5105

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	location
A	41.430	25	Y
A			
A	37.880	25	P
A			
A	36.760	25	W
B	19.080	25	C
B			
B	14.200	25	A

(7) A Healthy Homework Exercise:

(a) We know that

$$b_1 = \frac{SCP}{SSX} = \frac{-900}{600} = -1.5$$

and

$$_0 = \bar{Y} - b_1\bar{X} = 150 - (-1.5)(4) = 156$$

The most common mistake here is to forget to account for the negative sign on b_1 when doing the calculation for b_0. Our estimated regression equation is $\hat{Y} = 156 - 1.5X$.

(b) Your answers need to be specific both as to the context of the problem and the actual numbers involved. Do not just say that b_0 is the average value of Y when X is 0. Say what X and Y are and what the implications are in context. Here $b_0 = 156$ tells us that the average weight of people who do not exercise (X=0) is 156 pounds. This seems plausible–people who do not exercise may tend to be a bit overweight and this number seems a bit on the high side for a sample that should include both men and women. The units of b_1 are the same as the units of Y, namely pounds. You did not need to say that separately if you incorporated it in your explanation as I did above. The value $b_1 = -1.5$ means that on average each additional hour of exercise per week is associated with weight being **LOWER** by 1.5 pounds. The units are pounds per hour per week. Note that this does NOT mean that for every hour you exercise you lose 1.5 pounds! It says that if I take, for instance, people who exercise 4 hours per week and compare them to people who exercise 5 hours per week, on average the people who exercise 5 hours per week will weigh 1.5 pounds less than the people who only exercise 4 hours per week. Even if you view this relationship as causal (which it may not be!) you have to do the extra hour of exercise every week to maintain the 1.5 pounds lower of weight. This value also seems fairly reasonable to me–in general exercise makes you fitter so you might expect people who exercise more to weigh less than similar people who don't exercise as much. However the difference probably won't bee too big since (a) there are many other factors that affect weight besides just exercise and (b) some kinds of exercise (like weight-lifting) may actually increase muscle mass and so not in net decrease your weight. We accepted any reasonable discussion of whether you thought this value made sense.

(c) It is important here to be careful with your units. We are told the person exercises 10 hours per day but X is supposed to be measured in hours per week. 10 hours per day equates to X = 70 hours per week so our predicted value is

$$\hat{Y} = 156 - 1.5(70) = 51$$

We predict such a person weights 51 pounds. Obviously this is not a plausible weight for an adult or indeed anyone but a very young child and a very young child wouldn't exercise 10 hours per day. In fact, 10 hours a day is not a very realistic amount of exercise period–it is almost certainly outside the range of our data–no one in the class data bases claimed to excerise nearly this much!–so the model does not give reliable predictions.

(d) According to the model we get 1.5 pounds lower weight for each additional hour per week of exercise, so if we assume there is a **causal** relationship–i.e. that increasing our exercise will actually lower our weight–then to go down 6 pounds we need to increase our exerise by $6/1.5 = 4$ hours per week from what we are currently doing. This seems like quite a lot of extra exercise for a fairly small weight change! Also of course it may not be the exercise (or only the exercise) that is resulting in people who exercise more to weigh less. For instance, people who regularly exercise may also be more careful about what they eat and maintaining a healthy lifestyle in other ways which may be the actual cause of their lower weight although it doesn't seem implausible here to attribute some degree of causality to the exercise.

(e) Using the information in the problem statement we first get

$$Cov(X,Y) = \frac{SCP}{n-1} = \frac{-900}{101-1} = -9$$

Now the correlation is just the covariance divided by the two standard deviations. We are givien $Var(Y) = 900$ so $SD(Y) = \sqrt{Var(Y)} = \sqrt{900} = 30$. We also know that

$$Var(X) = \frac{SSX}{n-1} = \frac{600}{101-1} = 6$$

So $SD(X) = \sqrt{6}$. Putting this all together we have

$$Corr(X,Y) = \frac{-9}{30\sqrt{6}} = -.122$$

To determine the strength of a relationship we look at the correlation since it is unit free, unlike the covariance whose magnitude is hard to assess. Correlations near 1 or -1 imply strong relationships while correlations near 0 imply weak relationships. here our correlation of -.122 is weak. It is a negative relationship which isn't too surprising–we would expect exercise to increase fitness and on the whole reduce weight. However there are many other factors which affect weight such as genetics, gender, diet, etc., so it is not surprising that amount of exercise is not a super strong predictor. Common mistakes on this problem include equating SCP with the covariance, equating SSX with the variance of X, and forgetting to take square roots. This usually leads to a correlation greater than 1 in absolute value which should tell you something is wrong!! There or other ways to arrive at the answer such as noting that $b_1 = Cov(X,Y)/Var(X)$.

(f) The proportion of variability explained by the regression is simply

$$R^2 = \frac{SSR}{SST} = \frac{1350}{90000} = .015 = 1.5\%$$

Alternatively we could get this by squaring the correlation $(-.122)^2 = .015$. Either way we see that amount exercised explains a very small percentage of the variability in weight. As noted above there are many other factors that contribute to a person's weight so it is not surprising that the percentage is low but maybe surprising that it is this low.

(g) To tell whether the model is making good predictions we need to compare the root mean squared error to the Y values we are trying to predict. Here $RMSE = \sqrt{895.45} = 29.92$ meaning we are typically off by nearly 30 pounds when we use amount exercised to predict weight. The people in our sample weight an average of 150 pounds so this is an error of roughly 20%. This is not terribly good. I would be mad if someone thought I weight 30 pounds more and confused if they thought I weight 30 pounds less than I actually did!

(h) To test whether or not there is a significant relationship is equivalent to asking whether the slope of our regression line is 0 since then X does not effect Y. Thus our hypotheses are

$H_0 : \beta_1 = 0$–Knowing how much a person exercises does not help us to predict their weight.
$H_A : \beta_1 \neq 0$–There is a relationship between how much a person exercises and their weight.

As our test statistic we use F (which measures overall whether the model explains a lot fot he variability in Y). From our ANOVA table $F = 1.51$ and our p-value is

$$P(F_{1,n-2} \geq F_{obs}) = P(F_{1,99} \geq 1.51) = .2221$$

When I computed the ANOVA table I got the exact p-value from STATA:

```
. display Ftail(1,99, 1.51)
.22205204
```

On an exam if I didn't give you the p-value the best you could do would be to compare your observed F to the $\alpha = .05$ critical value on an F table. Here $1.51 < F_{1,99,.05} = 3.94$ so it is not large enough to reject the null hypothesis. We have insufficient evidence to show there is a relationship between amount exercised and

weight which is not too surprising given our earlier results.

(i) Optional Bonus: To find the ANOVA table from the data given at the beginning of the problem we first note that $Var(Y) = SSTY/(n-1)$ so $SST = Var(Y) * (n-1) = 900 * (101-1) = 90000$. Next we recall (see the warm-up problems) that $SSR = b_1 * SCP = (-1.5)(-900) = 1350$. We get the final sum of squares by noting that $SSE = SST - SSR = 90,000 - 1350 = 88650$. The degrees of freedom in a simple linear regressionare 1 for SSR, $n-2 = 99$ for the error and $n-1 = 100$ for the total. To get the mean squares we just divide the sums of squares by their respective degrees of freedom. $MSR = SSR/1 = SSR = 1350$. $MSE = SSE/(n-2) = 88650/99 = 895.45$. Finally, the F statistic is $F = MSR/MSE = 1350/895.45 = 1.51$. Putting these all together in our ANOVA table gives

Source	SS	df	MS	F
Regression	1350	1	1350	1.51
Error	88650	99	895.45	
Total	90000	100		

(8) Parenteral Nutrition:

(a) The scatterplots are shown in the accompanying graphics file. Something of an upward trend is visible in all of these graphs but to me the points look like they most closely follow a straight line in the dietary calcium plot. The gfr plot looks good for the first half but then spreads out and in the others the points are also more diffuse looking but this is something of a judgement call.

(b) STATA and SAS printouts of the correlations and covariances are given below. We judge the strength of the relationship based on the correlation. Here urinary calcium has the weakest correlation with gfr at r = .41, closely followed by urinary sodium at r = .49. The correlation with protein is r = .63 and the strongest relationship is with dietary calcium at r = .76. This roughly confirms my reading of the scatterplots but in fact all these correlations are reasonably high. All the variables show positive relationships with urinary calcium.

```
IN STATA:

Correlations:
. cor urinarycalcium dietarycalcium gfr urinarysodium protein
(obs=27)

             | uri~cium dietar~m      gfr uri~dium  protein
-------------+---------------------------------------------
urinarycal~m |   1.0000
dietarycal~m |   0.7584   1.0000
         gfr |   0.4103   0.1615   1.0000
urinarysod~m |   0.4934   0.1640   0.6155   1.0000
     protein |   0.6343   0.8823   0.2122   0.1827   1.0000

Covariances:
. cor urinarycalcium dietarycalcium gfr urinarysodium protein, cov
(obs=27)

             | uri~cium dietar~m      gfr uri~dium  protein
-------------+---------------------------------------------
urinarycal~m |  3736.15
dietarycal~m |  7448.47    25820
```

```
         gfr |  442.846  458.261  311.872
urinarysod~m |  1163.63  1016.99  419.389  1488.59
     protein |  1156.53  4228.78  111.778  210.211  889.704
```

IN SAS:

```
proc corr cov data = tmp1.hw3;
var urinarycalcium;
with dietarycalcium gfr urinarysodium protein;
run;
```

The CORR Procedure

4 With Variables: dietarycalcium gfr urinarysodium protein
1 Variables: urinarycalcium

Covariance Matrix, DF = 26

	urinarycalcium
dietarycalcium	7448.474359
gfr	442.846154
urinarysodium	1163.628205
protein	1156.525641

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
dietarycalcium	27	173.03704	160.68614	4672	0	554.00000
gfr	27	50.22222	17.65989	1356	16.00000	91.00000
urinarysodium	27	36.14815	38.58228	976.00000	0	134.00000
protein	27	36.62963	29.82790	989.00000	0	101.00000
urinarycalcium	27	74.66667	61.12409	2016	1.00000	220.00000

Pearson Correlation Coefficients, N = 27
Prob > |r| under H0: Rho=0

	urinarycalcium
dietarycalcium	0.75836 <.0001
gfr	0.41025 0.0335
urinarysodium	0.49342 0.0089

```
                    protein                   0.63434
                                               0.0004
```

(c) We need to fit four simple linear regresison models, all with urinary calcium as the response and each of the other variables as a predictor. The printouts are shown below:

IN STATA:

```
. regress urinarycalcium dietarycalcium

      Source |       SS       df       MS              Number of obs =      27
-------------+------------------------------           F(  1,    25) =   33.84
       Model |  55866.4585     1  55866.4585           Prob > F      =  0.0000
    Residual |  41273.5415    25  1650.94166           R-squared     =  0.5751
-------------+------------------------------           Adj R-squared =  0.5581
       Total |       97140    26  3736.15385           Root MSE      =  40.632

------------------------------------------------------------------------------
urinarycal~m |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
dietarycal~m |   .2884765   .0495908     5.82   0.000     .1863424    .3906106
       _cons |   24.74954   11.60949     2.13   0.043     .8393576    48.65973

. regress urinarycalcium gfr

      Source |       SS       df       MS              Number of obs =      27
-------------+------------------------------           F(  1,    25) =    5.06
       Model |  16349.4445     1  16349.4445           Prob > F      =  0.0335
    Residual |  80790.5555    25  3231.62222           R-squared     =  0.1683
-------------+------------------------------           Adj R-squared =  0.1350
       Total |       97140    26  3736.15385           Root MSE      =  56.847

------------------------------------------------------------------------------
urinarycal~m |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gfr |   1.419962   .6312997     2.25   0.034     .1197762    2.720148
       _cons |    3.35301   33.53974     0.10   0.921    -65.72337    72.42939
------------------------------------------------------------------------------

. regress urinarycalcium urinarysodium

      Source |       SS       df       MS              Number of obs =      27
-------------+------------------------------           F(  1,    25) =    8.05
       Model |  23649.7184     1  23649.7184           Prob > F      =  0.0089
    Residual |  73490.2816    25  2939.61126           R-squared     =  0.2435
-------------+------------------------------           Adj R-squared =  0.2132
       Total |       97140    26  3736.15385           Root MSE      =  54.218

------------------------------------------------------------------------------
```

```
urinarycal~m |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
urinarysod~m |   .7816969   .2755944     2.84   0.009     .2140997    1.349294
       _cons |   46.40977   14.42638     3.22   0.004     16.69809    76.12146
------------------------------------------------------------------------------
```

```
. regress urinarycalcium protein

      Source |       SS       df       MS              Number of obs =      27
-------------+------------------------------           F(  1,    25) =   16.83
       Model |  39087.5528     1  39087.5528           Prob > F      =  0.0004
    Residual |  58052.4472    25  2322.09789           R-squared     =  0.4024
-------------+------------------------------           Adj R-squared =  0.3785
       Total |       97140    26  3736.15385           Root MSE      =  48.188

------------------------------------------------------------------------------
urinarycal~m |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     protein |     1.2999   .3168333     4.10   0.000     .6473693     1.95243
       _cons |   27.05182   14.85567     1.82   0.081    -3.544009    57.64765
------------------------------------------------------------------------------
```

IN SAS:

To save space I won't repeat them all but here is a sample:

The REG Procedure
Model: MODEL1
Dependent Variable: urinarycalcium

Number of Observations Read	124
Number of Observations Used	27
Number of Observations with Missing Values	97

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	55866	55866	33.84	<.0001
Error	25	41274	1650.94166		
Corrected Total	26	97140			

Root MSE	40.63178	R-Square	0.5751
Dependent Mean	74.66667	Adj R-Sq	0.5581
Coeff Var	54.41756		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	24.74954	11.60949	2.13	0.0430
dietarycalcium	1	0.28848	0.04959	5.82	<.0001

(d) SST represents the total variability in Y and is computed without regard to any of the X variables so it does not change depending on what our predictors are as long as we are using the same outcome variable. The degrees of freedom are the same because each of our four models has 1 X variable and the same number, n = 27, of data points.

(e) We see that all four of the variables have a significant relationship with urinary calcium since the p-values for the F tests are all less than .05. However it is relatively close for GFR which has a p-value of .0335.

(f) The percentage of variability explained is given by R^2 or R^2_{adj} with values close to 1 or 100% being good. Here the percentages explained range from 16.8% (or adjusted 13.5%) for GFR to 57.5% (55.8% adjusted) for dietary calcium which is the best predictor by this measure by quite a bit. The next best is around 40%. The quality of the predictions is determined by comparing the root mean squared error to the average Y values. Here we have the same average Y in each model so the relative accuracy of the predictions can just be seen by ordering the RMSE values. Dietary calcium has the lowest RMSE at 40.63. However the mean value is only 74.67 so these errors are in fact still very large in percentage terms. Finally, in terms of the tests from part (e), although all the tests are significant, dieary calcium has the largest F value and the smallest p-value so it is the strongest predictor. In a set of simple linear regressions with the same Y values but different X's all these measures will give the same ranking of the variables.