Solutions To Homework 5

(1) Still More Height and Weight:

(a) The STATA printouts for the two simple regressions are shown below. The p-value for the simple regression of weight on age .0056 and the p-value for the simple regression of weight on height is .0203. Since both p-values are less than $\alpha = .05$ it appears that by themselves, both age and height have a significant linear relationship with weight. In simple linear regression both the t-test and the F test have null hypothesis $H_0 : \beta_1 = 0$ meaning there is not a significant relationship between X and Y and alternative hypothesis $H_A : \beta_1 \neq 0$ meaning there is a significant relationship between X and Y. Since a small p-value means we reject the null hypothesis, the small p-values indicate that the relationships between weight and the individual X variables ARE significant.

```
. regress Weight Height

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =    8.33
       Model |  2132.52386     1  2132.52386           Prob > F      =  0.0203
    Residual |  2047.07614     8  255.884518           R-squared     =  0.5102
-------------+------------------------------           Adj R-squared =  0.4490
       Total |      4179.6     9       464.4           Root MSE      =  15.996

------------------------------------------------------------------------------
      Weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Height |   2.160874    .748522     2.89   0.020     .4347795    3.886969
       _cons |  -17.80773   39.81467    -0.45   0.667    -109.6205    74.00507
------------------------------------------------------------------------------

. regress Weight Age

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =   14.07
       Model |      2664.5     1      2664.5           Prob > F      =  0.0056
    Residual |      1515.1     8    189.3875           R-squared     =  0.6375
-------------+------------------------------           Adj R-squared =  0.5922
       Total |      4179.6     9       464.4           Root MSE      =  13.762

------------------------------------------------------------------------------
      Weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         Age |       9.125   2.432768     3.75   0.006     3.515027    14.73497
       _cons |     -31.55   34.33565    -0.92   0.385    -110.7282    47.62816
------------------------------------------------------------------------------
```

(b) To see which model does a better job we look at R^2_{adj} (or R^2), RMSE and the F statistic or its corresponding p-value. Age has a higher R^2_{adj} meaning it explains more of the variability in weight than height does, it has a lower RMSE, meaning it leads to smaller prediction errors, and a bigger F/smaller p-value meaning that

age shows stronger evidence of a relationship with weight. Therefore, the model with age does a superior job.

(c) The STATA printout is given below. From it we see that the estimated regression equation is

$$\hat{Y} = -29.92 + 10.5Age - .396Height$$

To check whether the regression explains a significant amount of the variability in weight, we look at the overall F test. Our hypotheses are

$H_0 : \beta_1 = \beta_2 = 0$– i.e. neither age nor height helps to explain the variability in teenagers' weights. The model as a whole is NOT useful for predicting weight.
H_A : At least one of β_1, β_2 is not equal to 0 meaning that at least one of age and height IS useful for explaining the variation in weight. As a whole the model IS useful for predicting weight.

Our test statistic is $F = 6.23$ The p-value for this test is .0280 which is less than $\alpha = .05$. Therefore, we reject the null hypothesis and conclude that the regression including both age and height explains a significant amount of the variability in weight.

```
. regress Weight Age Height

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  2,     7) =    6.23
       Model |  2675.51176     2  1337.75588           Prob > F      =  0.0280
    Residual |  1504.08824     7  214.869748           R-squared     =  0.6401
-------------+------------------------------           Adj R-squared =  0.5373
       Total |      4179.6     9       464.4           Root MSE      =  14.658

------------------------------------------------------------------------------
      Weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         Age |   10.50048   6.605441     1.59   0.156    -5.118903    26.11987
      Height |  -.3958222   1.748475    -0.23   0.827    -4.530308    3.738664
       _cons |  -29.92317   37.27205    -0.80   0.448    -118.0576    58.21123
------------------------------------------------------------------------------
```

(d) To check whether age and height are significant in the multiple regression model we need to perform individual t tests. Our hypotheses for these tests are

$H_0 : \beta_1 = 0$, i.e. age does not explain any additional variability in weight beyond what is explained by height. It is not worth adding age to the model if height is already included.
$H_A : \beta_1 \neq 0$, i.e. age does explain some of the variation in weight beyond what is explained by height. It is worth adding age to the model even when height has been taken into account.

$H_0 : \beta_2 = 0$, i.e. The height of a teenager does not tell you anything more about their weight once you have taken their age into account. Height does not explain additional variability beyond that explained by age and is therefore not worth including in the model.
$H_A : \beta_2 \neq 0$, i.e. Height explains something about the person's weight beyond what is explained by their age. It is worth including height in the model even if age is already included in the model.

For the test of age, the t statistic is $t_{obs} = 1.59$ and the corresponding p-value is .1559. Since the p-value is greater than $\alpha = .05$ we fail to reject the null hypothesis that $\beta_1 = 0$. We conclude that age is not a

useful predictor **when height is in the model.** For the test of height, the test statistic is $t_{obs} = -.23$ and the corresponding p-value is .8274. This is a huge p-value. Once again we fail to reject the null hypothesis and conclude that height is not a useful predictor **when age is in the model.** Note the emphasis on the presence of the other variable in the model–this is a key part of the interpretation!

(e) In parts (a)-(d) we learned that age and height are both strongly related to weight and that the overall regression model is useful. However, neither of the individual variables appears useful in the multiple regression model. The key is the interpretation given in part (d)–neither of the variables is useful **when the other one is in the model.** This suggests that age and height **explain the same thing about weight.** In other words, we have a problem of **multicollinearity.** Our two predictor variables are strongly related to each other. One of the common side effects of multicollinearity is to make useful predictors appear insignificant. Another common side-effect of multicollinearity is to mess up the signs of the coefficients. Note that the coefficient of height is negative in the multiple regression model. This does not make sense. The taller a person is the more they should weigh. Intuitively it makes sense that age and height are related since you grow as you get older (at least up until the time you are a young adult and here we are looking at teenagers.)

(f) Our first check is to compute the correlation of age and height. The STATA printout is shown below. We get a whopping r = .92! This confirms that we have very strong multicollinearity.

```
. cor Age Height
(obs=10)

             |      Age   Height
-------------+------------------
         Age |   1.0000
      Height |   0.9198   1.0000
```

Since there are only two predictors the results from parts (a)-(e) plus the verification of the correlation would be sufficient proof of what has happened. However it is a good exercise to also get the variance inflation factors since they will be more informative in more complex problems. There are two ways to do this. One is to get STATA or SAS to give the VIFs as part of the regression printout. The output is as follows. Note that in STATA the VIFs are obtained with a post-estimation command, done after the regress command that for the MLR in part (c).

IN STATA:

```
. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
       ageq1 |      6.50    0.153894
      height |      6.50    0.153894
-------------+----------------------
    Mean VIF |      6.50
```

IN SAS:

```
                                The REG Procedure
                                  Model: MODEL1
                            Dependent Variable: weight
```

Number of Observations Read	124
Number of Observations Used	10
Number of Observations with Missing Values	114

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2675.51179	1337.75589	6.23	0.0280
Error	7	1504.08821	214.86974		
Corrected Total	9	4179.60000			

Root MSE	14.65844	R-Square	0.6401
Dependent Mean	96.20000	Adj R-Sq	0.5373
Coeff Var	15.23746		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-29.92317	37.27206	-0.80	0.4485	0
ageQ1	1	10.50048	6.60544	1.59	0.1559	6.49798
height	1	-0.39582	1.74847	-0.23	0.8274	6.49798

Alternatively one can get the VIFs directly using the formula

$$VIF = \frac{1}{1 - R_j^2}$$

where R_j^2 is the R-squared value obtained by regressing predictor X_j on the remaining predictors. Here that means we need to regress age on height and height on age. The corresponding simple linear regression printouts are shown below:

```
. regress ageq1 height

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =   43.98
       Model |  27.0753918     1  27.0753918           Prob > F      =  0.0002
    Residual |  4.92460817     8  .615576022           R-squared     =  0.8461
-------------+------------------------------           Adj R-squared =  0.8269
       Total |          32     9  3.55555556           Root MSE      =  .78459

------------------------------------------------------------------------------
       ageq1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
      height |   .2434837   .0367133     6.63   0.000     .1588227    .3281447
       _cons |   1.153799   1.952819     0.59   0.571     -3.34941    5.657007
------------------------------------------------------------------------------

. regress height ageq1

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =   43.98
       Model |  386.420032     1  386.420032           Prob > F      =  0.0002
    Residual |  70.2840151     8  8.78550188           R-squared     =  0.8461
-------------+------------------------------           Adj R-squared =  0.8269
       Total |  456.704047     9  50.7448941           Root MSE      =   2.964

------------------------------------------------------------------------------
      height |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ageq1 |       3.475   .5239723     6.63   0.000     2.266718    4.683282
       _cons |   4.109997   7.395252     0.56   0.594    -12.94348    21.16348
------------------------------------------------------------------------------
```

From this we get the variance inflation factor for age is

$$VIF_{age} = \frac{1}{1 - .8461} = 6.498$$

and the variance inflation factor for height is

$$VIF_{height} = \frac{1}{1 - .8461} = 6.498$$

Note that in this case the two variance inflation factors are the same. This is always true when your MLR has only two variables because the R^2 value for one of the X's on the other is just the square of the correlation and the correlation is the same whichever variable you take first! Both these VIFs are quite high–certainly enough to cause concern. Your book suggests that a VIF over 4 is bad and one over 10 is really bad but this is a very rough rule of thumb and needs to be combined with other evidence you see of multicollinearity problems in your model. It is possible to have multicollinearity (a relationship among the X variables) but for it not to cause a serious enough problem to need to remove one of the variables. Here the problem is severe enough that we need to do something.

(g) To fix our multicollinearity problem, one of the two variables should be removed. We certainly don't need both variables in the model since they are telling us the same thing. Since the p-value for age alone was smaller (.0056) than that for height alone (.0203) and we saw in several other ways in part (b) that age was the better individual predictor I would choose to keep age in the model. Note also that the simple linear regression model including just age has the highest R^2_{adj} of the three models which is additional evidence that it is the best choice. The value of R^2_{adj} actually went down when we added height to the model which indicates that we were overfitting–adding height did not help explain additional variability. It may seem clinically that height should be a better predictor than age–this is simply an artifact of the sample generated when I created the data–we would have to do a real-life study to be sure! Maybe next year I will fix the numbers....

(h) Mediation occurs when a predictor variable, X, affects an outcome variable, Y, **indirectly** through its effect on an intermediate variable, M which is thought to more directly cause Y. Mediation is conceptually about a causal path: X causes M which in turn causes Y. It is not defined statistically (as we have seen statistical methods rarely show causation!) but statistics can be used to test whether data are consistent

with a proposed mediation model. Specifically, if X causes Y solely through it's effect on M (full mediation) then X should be correlated with both M and Y, M should be correlated with Y, and when you regress Y on both M and X, X should no longer be significant but M should. If X causes Y partly through it's effect on M and partly through a direct effect on Y (partial mediation) then in the multiple regression of Y on M and X, X will still be significant but not as strongly significant as when M is not accounted for. In this example, with weight as the outcome, we have two possible mediation models. One would be that height mediates the effect of age on weight and the other would be that age mediates the effect of height on weight. The first path makes sense conceptually. The idea would be that increasing age (during the childhood/teenage years) causes you to weigh more but the reason it does so is that as you get older you grow (get taller) and correspondingly you get heavier. If we draw the mediation diagram the other way, we would be imagining that being taller causes you to weigh more but basically it causes you to weight more because it makes you older. Height is certainly correlated with age but being taller doesn't make you get older, so this doesn't make sense as a mediation model. While the first potential mediation model may seem reasonable, we don't have great support for it in our data. While it is certainly true that age is correlated with both height and weight and that height is correlated with weight from our various analyses, when we fit the multiple regression model, neither height nor age is significant so we do not have evidence that height provides any explanatory power beyond what is explained by age. By most definitions this would not count as mediation though some people have argued that you should only need M to be significantly related to Y on its own, not after adjusting for X. In any case we can not draw any firm conclusions about the causal pathyway in this example. There is nothing in the data that directly contradicts the age to height to weight theory but there isn't anything that strongly supports that directionality either.

(2) Harry Potter and the Sorceror's Statistic:

(a) An ANCOVA is basically a situation where you want to do an ANOVA (comparison of group means) but feel you need to adjust for other factors to make the comparisons fair. Here probably the most logical ANOVA structure would be a comparison of the relative profitability of Action, Romance and Comedy films, adjusting for differences across these genres in production and marketing factors. The only other potential grouping variable here is the "Big Name Star" variable. You could conceivably think of this as a t-test for whether having a star matters, adjusting for types of film and promotional status as well. Either way the adjustments probably are important to make the groups comparable. Action films for instance may typically cost a lot more to make (all those car chases and explosions) and so if you didn't adjust for the cost of production it could be hard to tell whether it was the type of film or the amount spent on it that increased the box office–maybe a romance film that had just as much spent on it as a typical action film would do as well. Similarly, different types of films might be more or less likely to have a big name star in them and since a star is (almost by definition) supposed to draw more people to the movie this could also take away from your ability to discern differences due to genre. This is not a clear-cut thing. In fact you could argue that the things that make people like action films better are precisely the things that cost money to make so that you can't separate the effect of genre from the effect of cost. All things to think about.....

(b) After adjusting for the other factors in this model, ction/adventure films are the most profitable, followed by romance films, and comedy films are the least profitable. We deduce this by looking at the coefficients of the Action and Comedy variables, X_5 and X_6. Note that $X_5 = 1, X_6 = 0$ corresponds to an action film, $X_5 = X_6 = 0$ is a romance film, and $X_5 = 0, X_6 = 1$ is a comedy film. Thus romance films serve as a reference category. Since $b_5 = 4.588$ we see that, all other things being equal, an action film makes \$4.588 million more than a romance film, while $b_6 = -5.1441$ means than on average a comedy makes \$5.1441 million less than an otherwise equivalent romance film. Note that these amounts only give the **difference** in box office sales between the different types of films. It is not accurate to say that a comedy film creates a loss or that a romance film brings in no money. To know the actual box office sales of a film you need to know the values of $X_1 - X_4$ as well.

(c) I think there is a significant difference between comedy and action films, all else equal. We saw in part (b) that action films made more than romance films and comedies made less than romance films. From the p-values in the model both of those differences are highly significant. Since the average difference between action and comedy films is even greater I would expect it to be significant too. The only way this is likely to get messed up is if I had many more romance films than either of the other types so I had more certainty about the comparison of romances to each of the other genres than for the direct comparison of action to comedy. We could also look at the condidence intervals for the Action and Comedy intervals and see if they overlapped (implying the same relative profitability versus romance films). However as we saw on an earlier assignment this sort of indirect comparison of confidence intervals is not as precise as a direct test. In STATA or SAS we could use the "test" command to formally check whether there was a difference in the coefficients of the Action and Comedy variables, since $\beta_5 - \beta_6$ is our estimate of the difference in box office sales between these two genres.

(d) Multicollinearity occurs when two or more of your predictor variables (the X's) are highly correlated with each other. Looking at the correlation table on the printout we see that the variables theaters, ads, and production costs all have very high correlations with one another (above .88 with 1 being the highest possible). Thus we certainly have evidence of multicollinearity. (Note that it is irrelevant that theaters, ads, and costs are highly correlated with box-office sales. Correlations between the X variables and Y do not show multicollinearity–in fact we want the X's to be highly correlated with Y so they will be useful predictors!) To decide what variables to include in our model we must look at the p-values for the individual t-tests. We see that all of our predictors are significant using $\alpha = .05$ (given the presence of the other variables) except for theaters and ads. Therefore I suspect that the best thing to do would be to drop these two variables and keep the rest. This makes sense since theater and ads were two of the three variables that were creating the multicollinearity problem. You could also use the fact that of theaters, ads, and costs, costs was the most highly correlated with box office sales and hence should be the one to be kept. However, it is really better to use the p-values since the star and film type variables may affect the relationships among the other variables. Also, technically, you should really only remove one variable from the model at a time because the current interpretations and p-values are dependent on the presence of ALL the other variables. Once you remove one variable, all the others can change. I would remove ads first since it has the highest p-value and refit the model. If all the variables were then significant I would stop. Otherwise I would remove the variable with the next highest p-value and so on. You could also try removing theaters first to make sure the highr p-value on ads wasn't a fluke of the multicollinearity. Ultimately the model that has all significant predictors and the highest R^2, lowerst $RMSE$, etc. is the best chocie. Given the current printout and the correlations among the variables however, one can be pretty confident that the final result would be a model dropping just theaters and ads (and in fact this is correct.)

(3) Interaction Basics:

(a) For a person who does exercise regularly, $X_2 = 1$. Plugging this into our regression equation gives

$$\hat{Y} = 80 + 1.2X_1 - 15(1) - .2X_1(1) = 65 + X_1$$

For a person who does not exercise regularly, $X_2 = 0$ and we get

$$\hat{Y} = 80 + 1.2X_1 - 15(0) - .2X_1(0) = 80 + 1.2X_1$$

(b) When you have an interaction term in a model interpretations are a little tricky. The intercept is not so bad. It is the average value of Y when all the X's are 0. Here that amounts to the average heart rate of a person who does not exercise regularly when they have run for 0 minutes–i.e. their resting heartrate. Our standard interpretation of $b_1 = 1.2$ would be that for every extra minute exercised heart rate goes up 1.2 beats per minute on average, assuming all the other variables are held fixed. But you can't assume the interaction term is held fixed here because it includes X_1! What we can say is that this interpretation is correct for people who do NOT excerise regularly since for them the interaction term disappears. Similarly,

we would like to say the coefficient $b_2 = -15$ means that on average people who exercise regularly have a heart beat that is 15 beats per minutes lower on average than that of a person who doesn't exercise regularly, all else equal. But again, the interaction term includes X_2 so we can not hold it fixed. Rather what we can say is that the INTERCEPT of the equation for people who exercise regularly is 15 beats per minute lower than that for people who don't excercise regularly. In other words, these people have a resting heart rate ($X_1 = 0$ minutes exercised) that is 15 beats per minute lower. When you plug in $X_1 = 0$ the interaction piece goes away. In general when you have an interaction term you have to specify the value of the second variable for which you are interpreting the coefficient of the first variable. Finally the coefficient of the interaction term tells us that for every extra minute exercised the heart rate of a person who exercises regularly goes up .2 beats per minute LESS than that of a person who does not exercise regularly. The interaction coefficient is giving us the DIFFERENCE in slope between the two groups. In fact, the easiest way to interpret these coefficients is to interpret the equations in part (a) separately for the two groups and not the differences between the intercepts and slopes. We see that people who exercise regularly have a resting heart rate 15 beats per minute lower and also that their heart rates do not increase as much for every minute exercised. This makes perfect sense.

(c) If the interaction term were not significant, our model would in effect be

$$\hat{Y} = 80 + 1.2X_1 - 15X_2$$

In other words, both people who exercised regularly and those who did not would have the same increase in heart rate per minute exercised–i.e. the same SLOPE–but people who exercised regularly would have a heart rate that was 15 beats per minute lower on average at any given amount of exercise time–i.e. the two groups have different INTERCEPTS. Visually this means the two groups would have parallel regression lines where is in parts (a) and (b) we were allowing lines that were not parallel.

(4) Don't Drink and Derive: This problem is good practice for how interactions might be incorporated in a more complicated model.

(a) Different kinds of drinks have different amounts of alcohol. For instance, 4 ounces of beer contains much less alcohol than 4 ounces of wine which in turn has far less alcohol than 4 ounces of hard liquor. Thus we would expect the effect of a certain amout of an alcoholic beverage to depend on which type of beverage it was. This is exactly what an interaction term is designed to do.

(b) Because there are three alcohol types (beer, wine and hard liquor) we will need three separate "slopes" for the relationship between BAC and amount drunk. The three groups are represented by two indicator variables, in this case Beer and Liquor, so we will need to add interaction terms for these indicators. The third slope will be rerpesented by the "amount" variable itself. Thus we need to add $X_{Beer*Amount}$ which will equal X_1, Amount, when the person drinks beer and 0 otherwise, and $X_{Liquor*Amount}$ which will equal X_1, Amount when the person drinks hard liquor, and 0 otherwise. The coefficient of the original Amount variable will tell us the slope relating amount drunk to BAC level for wine drinkers (all else equal). The coefficient for X_8, the Beer-Amount interaction will tell us how much greater or smaller the slope of the BAC-Amount relationship is for beer drinkers compared to wine-drinkers. Presumably, since beer has less alcohol content per unit volume than wine, this slope will be less steep (BAC content will go up more slowly) and hence b_8 will be negative. Similarly, the coefficient of X_9, the Amount-Liquor interaction will tell us how much greater the slope for the BAC-Amount relationship is for hard liquor drinkers than for wine drinkers, all else equal. Since hard liquor has the highest percentage of alcohol per unit volume we expect b_9 to be positive.

(c) Rather than going through the complicated procedure developed above we could simply create an "Amount" variable that was not the size of the drink in ounces, but rather the total alcohol content of the drink, which would be higher for a 5 ounce glass of wine than for a 5 ounce glass of beer, and so on. It is really the amount of actual alcohol, not the number of ounces of liquid in the glass that is important here!

(5) Short and Sweet:

(a) We have four groups in the model, MC, FC, MN and FN. If we want FC to be the reference group then we need three indicators, one for each of the other groups. For instance we could define $X_{MC} = 1$ if the person is a man using the current insulin control techniques and 0 otherwise. The other two indicators X_{MN}, X_{FN} would be defined anagously and the regression equation would be

$$\hat{Y} = b_0 + b_{MC}X_{MC} + b_{MN}X_{MN} + b_{FN}X_{FN}$$

We know that for an ANOVA set up treated as a regression, the intercept is just the mean of the reference group so $b_0 = 120$. The coefficients for each of the indicators gives the difference in means between that group and the reference group so we have $b_{MC} = 125-120 = 5, b_{MN} = 100-120 = -20$ and $b_{FN} = 105-120 = -15$. The resulting estimated regression equation is

$$\hat{Y} = 120 + 5X_{MC} - 20X_{MN} - 15X_{FN}$$

(b) If we use men on the new insulin control method as the reference group then their mean is the intercept so $b_0 = 100$, To get the other coefficients we compare the MC, FC and FN groups to the MN group. We get $b_{MC} = 125 - 100 = 25, b_{FC} = 120 - 100 = 20$ and $b_{FN} = 105 - 100 = 5$ with an estimated regression equation of

$$\hat{Y} = 100 + 25X_{MC} + 20X_{FC} + 5X_{FN}$$

(c) The ANOVA table for a regression is the same as the corresponding ANOVA ANOVA table with SSR = SSB, SSE = SSW, SST = SST, etc. To answer the first part of the question we need to compute R^2 and R^2_{adj}, the two estimates of the percentage of variability explained. We get

$$R^2 = \frac{SSR}{SST} = \frac{SSB}{SST} = \frac{6400}{22600} = .2831 = 28.31\%$$

$$R^2_{adj} = 1 - \frac{SSW/(n-G)}{SST/(n-1)} = 1 - \frac{16200/60}{22600/63} = .2473 = 24.73\%$$

Either way it looks like the ANOVA only explains about a quarter of the variability in blood sugar levels which is not very good, but it is hardly surprising since there are many factors besides gender and treatment that would play a role includung weight, exercise, genetics, disease severity, etc.

To tell whether the predictions are good we need to compare RMSE to the Y values we are trying to predict. In an ANOVA MSE = MSW, so RMSE = $\sqrt{MSW} = \sqrt{270} = 16.43$. This means typically we make an error of about 16.43 mg/dL in predicting blood sugar levels. Since the levels themselves have a grand mean of about 110, with group means randing from 100-125 this seems like a pretty big error. (16.43/110 is about a 15% error). Again this result is not very susprising.

(d) The overall F test for the ANOVA which is equivalent to the overall F test for the regression, is highly significant with a p-value of .00016 so we conclude that there is evidence of a difference in means between some of the groups. This implies that there is some impact of either gender or method or both though we can't tell from the F test which effects are present. For that we need the pairwise comparisons of group means (ANOVA speak) or tests about the regression coefficients for the indicator variables (regression speak).

(e) We definitely have enough information to fill in the parameter estimates table. We showed in parts (a) and (b) how to get the estimates of the regression coefficients. The estimates of the standard errors can be obtained from the ANOVA formula for the standard error of a linear combination

$$\hat{LC} = \sqrt{MSW \sum \frac{c_j^2}{n_j}}$$

The intercept is just a linear combination corresponding to the mean of a single group (one constant = 1, the rest = 0) while the coefficients for the indicators correspond to pairwise differences in means (one constant = 1, one constant = -1, the rest = 0). Similarly the t statistics for all the coefficients are just the t-statistics for tests that the particular linear combination they correspond to. From the t-statistics one can compute the p-values. Similarly, the formulas for the confidence interval of a linear combination gives the confidence intervals for the regression coefficients.

(f) Dr. Sweet defined the gender indicator as $X_G = 1$ for males and 0 for females and the treatment indicator as $X_T = 1$ for people using her new device and 0 for those who got treatment as usual. An interaction variable is just a product so $X_{GT} = X_G * X_T$.

(g) An interaction deals with the **joint** effect of two predictor variables on the outcome, Y. It does NOT imply a relationship between the predictor variables themselves. Specifically, here, an interaction between gender and treatment would mean that the effectiveness of the new device depends on gender (i.e. it works differently for men and women) or alternatively that the difference in blood sugar levels for men and women is different depending on whether or not they are using the new technique.

(h) The p-value for the interaction term, Gender*Treat, in the above model is .247, much larger than $\alpha = .05$, which means there is no significant evidence of an interaction in this model. Practically speaking this says that there is no difference in benefit from the new device for men and women. A lot of people who took the exam on which this problem originally appeared talked about different "slopes" for men and women or for the two treatment groups but that doesn't really make sense in this context since both the variables involved in the interaction are indicators.

(i) The interaction looks at whether the effect of the device is different for men versus women. The effect of the device for women is $\mu_{FN} - \mu_{FC}$ and the effect for men is $\mu_{MN} - \mu_{MC}$. Thus the linear combination of interest is

$$LC = (\mu_{MN} - \mu_{MC}) - (\mu_{FN} - \mu_{FC}) = \mu_{MN} + \mu_{FC} - \mu_{MC} - \mu_{FN}$$

This is a standard linear combination and can be analyzed using the techniques we learned for ANOVA. Specifically our best estimate is

$$\hat{LC} = 100 + 120 - 125 - 105 = -10$$

which is the same as the coefficient of the interaction. Note that to get the sign right I had to look at the difference for men minus the difference for women since it was the men who were coded as 1 in the model of part (h)! Similarly the standard error is

$$se(\hat{LC}) = \sqrt{MSW \sum \frac{c_j^2}{n_j}} = \sqrt{270(\frac{(1)^2}{20} + \frac{(1)^2}{12} + \frac{(-1)^2}{12} + \frac{(1)^2}{20})} = \sqrt{72} = 8.49$$

Turn-In Problems

(6) Relationships of the HAART:

(a) The multiple regression printout is shown below. Note that since we want the non-HAART group to serve as the reference we do NOT use its indicator in the model!

```
. regress vload aonly bonly combo

      Source |       SS       df       MS              Number of obs =     124
-------------+------------------------------           F(  3,   120) =   26.94
       Model |  32.2306727     3  10.7435576           Prob > F      =  0.0000
    Residual |  47.8529587   120  .398774656           R-squared     =  0.4025
-------------+------------------------------           Adj R-squared =  0.3875
       Total |  80.0836314   123  .651086434           Root MSE      =  .63149

------------------------------------------------------------------------------
       vload |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       aonly |  -.4067836   .1603976    -2.54   0.012    -.7243598   -.0892075
       bonly |  -.6832356   .1603976    -4.26   0.000    -1.000812   -.3656595
       combo |  -1.398368   .1603976    -8.72   0.000    -1.715944   -1.080792
       _cons |   4.615798   .1134182    40.70   0.000     4.391238    4.840359
------------------------------------------------------------------------------
```

(b) When you fit an ANOVA as a regression model the intercept gives you the average value of Y for the reference group and the coefficients of the indicator variables give you the difference in means between those groups and the reference group. From the above printout we see that $\bar{Y}_N = b_0 = 4.62$ for the non-HAART group. To get the mean for any of the othe groups we just have to add it's coefficient to the mean of the reference group, e.g.

$$\bar{Y}_A = \bar{Y}_N + (\bar{Y}_A - \bar{Y}_N) = b_0 + b_A = 4.62 + (-.41) = 4.21$$

Similarly $\bar{Y}_B = 4.62 - .68 = 3.94$ and $\bar{Y}_{AB} - 4.62 - 1.40 = 3.22$. Using this information it is easy to construct the regression model using any other group as the reference. For instance, if we use HAART A as the reference group we have $b_0 = \bar{Y}_A = 4.21$. To get the coefficients for the other indicators we take the difference between them and the HAART A group. $b_N = \bar{Y}_N - \bar{Y}_A = 4.62 - 4.21 = .41$. Note that this is just the reverse of the coefficient for group A in the model with non-HAART as the reference. Similarly, $b_B = 3.94 - 4.21 = -.27$ and $b_{AB} = 3.22 - 4.21 = -.99$. The resulting estimated regression equation is

$$\hat{Y} = 4.21 + .41X_A - .27X_B - .99X_{AB}$$

We can verify this by fitting the regression in STATA with the HAART A only indicator left out since it is to serve as the reference group. The printout confirms our calculations above up to rounding:

```
. regress vload n bonly combo

      Source |       SS       df       MS              Number of obs =     124
-------------+------------------------------           F(  3,   120) =   26.94
       Model |  32.2306727     3  10.7435576           Prob > F      =  0.0000
    Residual |  47.8529587   120  .398774656           R-squared     =  0.4025
-------------+------------------------------           Adj R-squared =  0.3875
       Total |  80.0836314   123  .651086434           Root MSE      =  .63149

------------------------------------------------------------------------------
       vload |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
           n |   .4067836   .1603976     2.54   0.012     .0892075    .7243598
       bonly |   -.276452   .1603976    -1.72   0.087    -.5940281    .0411242
       combo |  -.9915844   .1603976    -6.18   0.000     -1.30916   -.6740082
       _cons |   4.209015   .1134182    37.11   0.000     3.984455    4.433575
------------------------------------------------------------------------------
```

(c) The overall F test for an ANOVA is equivalent to the overall F test for a regression. The primary difference is in the way the mathematical hypotheses are stated. For an ANOVA your null hypothesis is that all the means are equal. In a regression your hypothesis is that the differences between the reference group and the other groups are all 0. Using the model of part (a) with the no-HAART group as the reference we would write:

$H_0 : \beta_A = \beta_B = \beta_{AB} = 0$–none of the HAART groups has a different average viral load than the non-HAART group–i.e. viral load is not related to which treatment you receive.
H_A :–at least one of the $\beta's$ is not equal to 0. There is a difference in average viral load between at least two of the treatment groups.

The overall F test is highly significant with F = 26.94 and a p-value of .0000 to as many decimal places as STATA gives us so we conclude that there is evidence of a difference in means between some of the groups. This implies that there are differences between the treatment groups but we can't tell from the F test specifically what these differences are. For that we need the pairwise comparisons of group means (ANOVA speak) or tests about the regression coefficients for the indicator variables (regression speak) which we will do in part (e).

(d) To answer the first part of the question we need to look at R^2_{adj} which estimates the percentage of variability in viral load explained by the treatment regimens. From either of our two regression printouts above we see that $R^2_{adj} = .3875 = 38.75\%$ meaning that treatment regimen explains just over a third of the variability in viral load. This is not particularly good, but it is hardly surprising since there are many factors besides treatment that would play a role in a person's viral load including severity of illness, at what stage in the illness the measurement was taken, age, concommitant health factors, etc.

To tell whether the predictions are good we need to compare RMSE to the Y values we are trying to predict. From our printouts, RMSE = .63. This is our average prediction error on the scale of log 10 viral load. This number is not particularly intuitive. However we note that the means of the four treatment groups ranged from roughly 3.2 to 4.2 on the same scale. Thus our RMSE represents an error of roughly 15-20% (.63/4.2 to .63/3.2) which is quite a lot. Again this result is not very susprising. It is rare that an ANOVA without other covariates added does a spectacular job of explaining the variability in outcome.

(e) Now we want to test for pairwise differences in mean between the treatment groups. In the regression setting we have to phrase these tests in terms of the regression coefficients, β. Note that in an ANOVA setting the β's give the difference in means between each of the other groups and the reference groups. Thus we can getthe comparisons with the reference group just by doing the t-tests for the coefficients of the indicators. As an example I will write out the details of the test comparing the HAART A group to the non-HAART group based on the model in part (a):

$H_0 : \beta_A = 0$–there is no difference in mean viral load between the HAART A and non-HAART group.
$H_A : \beta_A \neq 0$–there is a difference in viral load between the HAART A and non-HAART groups.

The test statistic is $t_{obs} = -2.54$ and the corresponding p-value (two-sided) is .012 so we reject the null hypothesis at $\alpha = .05$ and conclude there is a difference in average viral load between the HAART A and non-HAART groups. The p-values for the comparisons of HAART B and the combination therapy to non-HAART are 0 so these treatment groups also have significantly different mean viral loads from our reference

groups. To do the other three comparisons we need a slightly different form of test which we can not get directly from the printout. For instance suppose we wanted to compare the HAART B group to the combination therapy. Really what we want is to test whether $\mu_B = \mu_{AB}$. As we noted in part (a), $\mu_B = \beta_0 + \beta_B$ and $\mu_{AB} = \beta_0 + \beta_{AB}$–we get the group means by adding the coefficients of the indicators to the reference group. Thus to test $\mu_B = \mu_{AB}$ amounts to testing $\beta_0 + \beta_B = \beta_0 + \beta_{AB}$ or equivalently $\beta_B = \beta_{AB}$. Intuitively this should make sense. We are testing whether the differences between each of the B and AB groups and the reference group are the same. If they are then those two groups would have to have the same mean as each other. We write the hypotheses as:

$H_0 : \beta_B = \beta_{AB}$–there is no difference in average viral load between the people on HAART B and those on the combination therapy (these groups are equally different from the non-HAART group).
$H_A : \beta_B \neq \beta_{AB}$–there is a difference in mean viral load between those on HAART B and those on the combination of HAART A and B.

The test statistic is

$$t_{obs} = \frac{(b_B - b_{AB}) - 0}{s.e.(b_B - b_{AB})}$$

I never actually showed you how to calculate the standard error for a combination of coefficients in a regression but we could work it out using the ANOVA formulas for the linear combination of the difference in means which is what this test statistic represents. If we do the test in STATA it will give us an F statistic which is just the square of the t-statistic. The commands and output are as follows:

```
. test bonly = combo

 ( 1)  bonly - combo = 0

       F(  1,   120) =   19.88
            Prob > F =    0.0000

. test aonly = combo

 ( 1)  aonly - combo = 0

       F(  1,   120) =   38.22
            Prob > F =    0.0000

. test bonly = aonly

 ( 1) - aonly + bonly = 0

       F(  1,   120) =    2.97
            Prob > F =    0.0874
```

We see that the difference between each of the individual therapies and the combination therapy are highly significant with p-values of 0 out to as many decimal places as STATA gives us. However the difference between the HAART A and HAART B groups is not quite significant–the p-value is .0874. Thus in that one case we can not reject the null hypothesis of no difference. Overall it seems that each of the HAART treatments is significantly different from the non-HAART treatment (and in fact they are better since the mean

viral loads in all these groups are lower) and that the combination therapy is better than each of the other therapies. However we can not be sure one of the individual therapies is better than the other. Note that we need to be very careful about this global statement though since we did the tests WITHOUT adjusting for multiple comparisons! If we did adjust, say using Bonferroni as was done on a previous assignment in the ANOVA framework, we actually could not be sure that the HAART A therapy was better than no HAART simultaneously with all our other conclusions.

(f) Professor Helpful is interested in a range of values that will include the viral loads of most of the **individual** subjects taking the combination therapy. The will therefore be better off using a prediction interval than a confidence interval. The confidence interval will have a high probability of including the AVERAGE viral load of people on this treatment but that does not mean that it will include all the individual subjects. Actually, as Adam noted in an e-mail to the class, even the prediction interval may well not include 95% of the individuals. Basically, using a 95% PI means that if we take a data set and calculate the regression and get a PI for a particular individual then there is a 95% chance the procedure will capture the viral load of the particular individual we asked about. However if we got an unlucky sample we wouldn't just be missing that one individual, we would probably be missing LOTS of individuals. Another way of saying this is that a 95% PI doesn't have an 100% chance of including 95% of individuals. Any particular interval has a high chance of including any particular individual but that isn't the same thing. However a PI is still the best choice Professor Helpful can make.

(g) The people who are taking HAART A are the people in the A only and combination groups. The people not taking HAART A are the people in the B only and non-HAART groups. Higher viral loads are worse. Thus what we really want to prove (as our alternative hypothesis) is that

$$\frac{mu_B + \mu_N}{2} - \frac{mu_A + \mu_A B}{2} > 0$$

Using the notation we developed in part (a) we can write each of the group means in terms of the regression coefficients: $\mu_N = \beta_0, \mu_A = \beta_0 + \beta_A$, and so on. Substituting these values in we get

$$\frac{\beta_0 + \beta_B + \beta_0}{2} - \frac{\beta_0 + \beta_A + \beta_0 + \beta_{AB}}{2} > 0$$

Note that each mean has a β_0 piece so that there terms cancel out and we are left with

$$\frac{\beta_B - \beta_A - \beta_{AB}}{2} > 0$$

The factor of 1/2 is of course not important to the actual hypothesis test (we could multiply the whole expression by 2) but it would be important if we wanted to estimate the difference in means between the HAART A takers and non-takers. Our best estimate of this difference, using the regression coefficients from part (a) is

$$\frac{b_B - b_A - b_{AB}}{2} = \frac{-.68 - (-.41) - (-1.4)}{2} = .57$$

Our hypotheses in terms of the regression coefficients are

$H_0 : \beta_B \leq \beta_A + \beta_{AB}$–those taking HAART A are not better off than those not taking HAART A.
$H_A : \beta_B > \beta_A + \beta_{AB}$–those taking HAART A have a lower viral load on average than those not taking HAART A.

As before we can get STATA to perform an F test about this combination of the regression coefficients. STATA's test will be two-sided so we will in principle need to divide the p-value in half to get the one-sided test we want. In fact (see below) the test is so highly significant that the p-value is essentially 0. The best we can conclude is that the two-sided p-value is less than .00005 so the 1-sided p-value must be less than

.000025 but whatever it is, it is certainly small enough to reject the null hypothesis and to conclude that those taking HAART A are better off (have lower viral loads) than those not taking HAART A. We can do the same test for those taking HAART B. We simply flip the roles of HAART A and B in our expression. This difference is even more significant as we can tell from the much higher F value. Our best estimate of the difference in means between the hAART B takers and not takers is, on the scale of log 10 viral load:

$$\frac{b_A - b_B - b_{AB}}{2} = \frac{-.41 - (-.68) - (-1.4)}{2} = .84$$

```
. test bonly = aonly+combo

 ( 1) - aonly + bonly - combo = 0

       F(  1,   120) =   24.46
            Prob > F =    0.0000
. test aonly = bonly + combo

 ( 1)  aonly - bonly - combo = 0

       F(  1,   120) =   54.51
            Prob > F =    0.0000
```

(h) The indicator for "takes HAART A" should be 1 if the person takes HAART A (i.e. is in the HAART A or combo group) and 0 otherwise (i.e. if the person is in the non-HAART or HAART B group). The simplest way to get this is just to add the original indicators for HAART A only group and the combination group. Similarly to get the "takes HAART B" indicator we can add the indicators for the B only and combination groups. To get the interaction between "takes A" and "takes B" we could just multiply our two new indicators. However, the result will be the same as our old indicator for the combination group because that corresponds exactly to those who take both medications. The STATA commands are

```
. gen takesA = aonly+com
. gen takesB = bonly + combo
```

(i) The resulting regression model fit in terms of these new variables plus the interaction is

```
. regress vload takesA takesB combo

      Source |       SS       df       MS              Number of obs =     124
-------------+------------------------------           F(  3,   120) =   26.94
       Model |  32.2306727     3  10.7435576           Prob > F      =  0.0000
    Residual |  47.8529587   120  .398774656           R-squared     =  0.4025
-------------+------------------------------           Adj R-squared =  0.3875
       Total |  80.0836314   123  .651086434           Root MSE      =  .63149

------------------------------------------------------------------------------
       vload |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      takesA |  -.4067836   .1603976    -2.54   0.012    -.7243598   -.0892075
      takesB |  -.6832356   .1603976    -4.26   0.000    -1.000812   -.3656595
       combo |  -.3083487   .2268365    -1.36   0.177    -.7574692    .1407717
```

```
       _cons |   4.615798   .1134182    40.70   0.000     4.391238    4.840359
------------------------------------------------------------------------------
```

(j) We are asked if there is a significant interaction between taking HAART A and taking HAART B. To determine this we have to look at the t-test for the combination term which is the interaction variable in this model. It's p-value is .177 which is not less than our significance level of $\alpha = .05$ so we do not have sufficient evidence to show there is an interaction. What this means is that the effect of HAART B does not depend on whether the person is taking HAART A and vice versa. Another way to say this is that the difference in viral loads between the A and AB groups is the same as the difference between the N and B groups and so on. We call such a situation an "additive" effect. To get the effect of treatments A and B together you simply add up their individual effects. If there had been a significant interaction it would have suggested that the effect of HAART B did depend on whether the person was taking HAART A and vice versa. This could occur either if the medications were synergistic (i.e. each one enhanced the effect of the other) or if they were antisynergistic (i.e. once you were taking one of the treatments adding the other wouldn't have as much effect as it would in isolation).

(k) Optional Bonus: I only give a sketch of the solution here to reserve the problem for future years. If you did the problem and have questions about your answer feel free to ask. Basically the idea is the following. A CI for Y is supposed to be an interval for the average value of Y at a given set of X values. Here the "set of X values" is simply an indication of what group the person belongs to. Thus we are just asking for CIs for the average log 10 viral load in each group. We can use our usual confidence intervals for a sample mean but using RMSE which is the pooled estimate of the standard deviation across all the groups rather than the standard deviation of points within the single group for which we are calculating the CI. To get the prediction interval we just need to add in an extra factor of RMSE to account for the variation of individuals about the group mean. (See Homework 4, Problem 6(h) for the way the two standard errors are related.) From this point calcuating the PI for part (f) is easy since we have already computed the group means and the RMSE. For the final part of the bonus problem you need to note that the X values are just 1's and 0's. If you let n_1 and n_2 be the number of 1's and 0's it is just a matter of algebra to crank through the formulas for $\bar{X}$ and SSX and see that you end up with the usual two-sample t-test formula.

(7) If Memory Serves...

(a) Dr. Brain is asking whether the relationship between age and memory score differs for patients and healthy subjects. Another way of saying this is that the age-memory relationship (or slope) depends on whether or not you are a patient which is precisely the definition of an interaction. To create theage by group interaction we just take the product of the two variables. The STATA command for this and the resulting multiple regression printout are shown below.

```
. gen agegroup = age*group
(64 missing values generated)

. regress memory age group agegroup

      Source |       SS       df       MS              Number of obs =      60
-------------+------------------------------           F(  3,    56) =   92.03
       Model |  13480.6019     3  4493.53398           Prob > F      =  0.0000
    Residual |  2734.31516    56  48.8270565           R-squared     =  0.8314
-------------+------------------------------           Adj R-squared =  0.8223
       Total |  16214.9171    59  274.829104           Root MSE      =  6.9876
```

```
------------------------------------------------------------------------------
      memory |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.2545832   .0700042    -3.64   0.001    -.3948186   -.1143479
       group |  -9.534811   5.662514    -1.68   0.098    -20.87819    1.808568
    agegroup |  -.2560998   .1049149    -2.44   0.018    -.4662695     -.04593
       _cons |   104.1128   3.366299    30.93   0.000     97.36933    110.8563
------------------------------------------------------------------------------
```

(b) To check whetehr there is a significant interaction all we have to do is perform a t-test for the agegroup variable in the model from part (a). Our hypotheses are

$H_0 : \beta_3 = 0$–the interaction term, agegroup, does not explain additional variability in memory score beyond what is explained by age and group alone.
$H_A : \beta_3 \neq 0$–the interaction term does add explanatory power to the model. The relationship between memory and age is different for patients than for healthy subjects.

Our test statistic is $t_{obs} = -2.44$ and the correpsonding p-value is .018 which is less than our significance level of $\alpha = .05$ so we reject the null hypothesis and conclude that there is an interaction between age and diagnostic status. In particular since the interaction is negative we conclude that memory score deteriorates more rapidly in patients (group = 1) than it does in healthy subjects (group = 0).

(c) For a healthy subject group = 0 so from the printout the estimated regression equation is

$$\hat{Y} = 104.113 - .254age$$

For a patient group = 1 and the estimated regression equation is

$$\hat{Y} = 104.113 - .254age - 9.535 - .256age = 94.578 - .510age$$

The easiest way to plot the lines is to find the intercepts when age = 0 and at some fairly high age, say 80, and to connect the dots. At age 0 healthy subjects would have an average memory score of 104.113 and patients a score of 94.578 (assuming babies can take the test–you can think of this as a measure of innate memory ability!) At age 80 healthy subjects would have a memory score of $\hat{Y} = 104.113 - .254(80) = 83.793$ and patients would have an average score of $\hat{Y} = 94.578 - .51(80) = 53.778$. The graph is included in a separate file.

(d) We can't interpret the coefficients of group and age separately in this model because it is not possible to change one of the variables and hold all the others fixed. The interaction term changes whenever one of age and group changes. However we can interpret the coefficients jointly by looking at what happens WITHIN each of the diagnostic groups. Specifically, we see that $b_0 = 104.113$ is the average memory score of a person who is 0 years old (an infant) and who does not have the disease. $b_1 = -.254$ means that FOR HEALTHY PEOPLE (group = groupage = 0), each extra year of age is associated on average with a quarter of a point lower memory score. The coefficients of the group and interaction variable tell us how the patients DIFFER from the healthy people. Specifically, $b_2 = -9.535$ says that on average 0 year old infants with the disease will have memory scores nine and a half points lower than their healthy counterparts and $b_3 = -.256$ means that the average memory score of patients tends to go down a quarter of a point more per year than that of health subjects. Of course this relationship is not necessarily causal and it does not mean that every subject's memory score drops at the indicated rate as they age. Rather we are looking cross sectionally at the average relationship between memory score and age.

(e) One would expect that as people get older their memory deteriorates, even if they are healthy. What the interaction in part (b) tells us is that among patients memory deteriorates EVEN MORE than we would

expect from the results of normal aging. This part of the memory deterioration with age must be ILLNESS RELATED. Thus if we fit a model in patients only and adjust for age we will be "adjusting out" not only the normal aging effect but also some of the effects of illness. Since Dr. Brain is interested in assessing the effects of illness he will be getting rid of exactly some of the variability he wants to study which is bad. Moreover we are told that he is specifically interested in the effects of illness duration which will be highly correlated with age (the longer your illness lasts, the older you will get). Thus the "illness-related" piece of the age effect may be exactly the illness duration effect he wants to study. If he adjusts for the full effects of age instead of just the normal effects of aging in this group he will not be able to see the effects of illness duration.

(f) Sandra's idea is a good one (as her name should suggest) and is actually something I do regularly in my psychiatry studies. Basically what she is doing is subtracting out only the effects of normal aging. What is left is a memory score that includes any illness effects that are related to age. We do not believe these effects are really DUE to aging because we know what effects aging should have by looking at the healthy subjects. We will therefore not include age in our subsequent models for the patients but rather will concentrate on variables like duration of illness that are the more likely causes of the illness effects but are correlated with age and would have been confounded if we had to include age in the models. The printout for the regression in the healthy subjects is shown below along with the commands for creating the new variable and checking its values.

```
Regression for healthy subjects:

. regress memory age if group==0

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  1,    28) =   11.88
       Model |  645.760524     1  645.760524           Prob > F      =  0.0018
    Residual |  1522.30777    28  54.3681346           R-squared     =  0.2979
-------------+------------------------------           Adj R-squared =  0.2728
       Total |  2168.06829    29  74.7609756           Root MSE      =  7.3735

------------------------------------------------------------------------------
      memory |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.2545832   .0738697    -3.45   0.002    -.4058984    -.103268
       _cons |   104.1128   3.552177    29.31   0.000     96.83653    111.3891
------------------------------------------------------------------------------

Creating predicted values: (Note: I created the predicted values for both
patients and healthy subjects because doing so was harmless but you could
also have used the ‘‘if group==1’’ command to do patients only)

gen mempred= 104.1128 - .2545832*age

Creating adjusted values:

gen memadjusted = mempred - memory

Comparing adjusted values to my predefined variable memorynormed: We can
see that basically these values are all 0 up to rounding.  I must have
rounded the coefficients slightly differently the two times I did this!
```

```
gen memtest = memorynormed - memadjusted
(94 missing values generated)

. summarize memtest

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     memtest |        30    -.001336    .0000531  -.0014496  -.0012684
```

(g) The simple linear regression of memory normed on age in the patients (group=1) is shown below. The relationship is significant based on the p-value of for the F test or the t-test (which are equivalent in an SLR). Note that I didn't have to specify group=1 in the regression statement because memorynormed was defined only for the patients. I did this when I created the variable to avoid mistakes. This is hardly surprising. In part (d) we found a very significant interaction between age and group telling us the memory scores deteriorated much faster with age in the patients than in the healthy subjects. The memory normed variable represents the difference in age effect between what was observed in a patient and what we would have expected if they were healthy–namely exactly the interaction effect. Thus we would expect the relationship to be significant since the interaction was significant earlier. Note that I set up the differences so that memory normed is telling us how much WORSE the memory score of the patients are from what we would have expected from normal aging. This why my coefficient is now positive rather than negative.

```
. regress memorynormed age

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  1,    28) =   12.12
       Model |   524.43902     1   524.43902           Prob > F      =  0.0017
    Residual |  1212.00719    28  43.2859712           R-squared     =  0.3020
-------------+------------------------------           Adj R-squared =  0.2771
       Total |  1736.44621    29  59.8774556           Root MSE      =  6.5792

------------------------------------------------------------------------------
memorynormed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .256103   .0735768     3.48   0.002     .1053878    .4068182
       _cons |   9.533255   4.287113     2.22   0.034     .7515015    18.31501
------------------------------------------------------------------------------
```

(h) The multiple regression of memorynormed on age and illness duration is shown below. To test whether age is a significant predictor in this model we use the t-test. Our hypotheses are

$H_0 : \beta_1 = 0$–age does not explain any variability in memory score beyond what is explained by illness duration and is therefore not worth adding to the model.
$H_A : \beta_1 \neq 0$–age explains additional variability in memory score beyond what is explained by illness duration and is worth including in the model.

Our test statistic is $t_{obs} = -1.50$ and the corresponding p-value is .145 which means we fail to reject. We do NOT have sufficient evidence to show that age tells us anything that illness duration does not.

```
. regress memorynormed age illnessdur

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  2,    27) =  371.03
       Model |  1675.48271     2  837.741355           Prob > F      =  0.0000
    Residual |   60.963503    27  2.25790752           R-squared     =  0.9649
-------------+------------------------------           Adj R-squared =  0.9623
       Total |  1736.44621    29  59.8774556           Root MSE      =  1.5026

------------------------------------------------------------------------------
memorynormed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0316543   .0210906    -1.50   0.145    -.0749287    .0116201
  illnessdur |   .9138978   .0404767    22.58   0.000     .8308466    .9969491
       _cons |  -.6855976   1.078682    -0.64   0.530    -2.898871    1.527676
------------------------------------------------------------------------------
```

(i) In part (g) age was BY ITSELF a significant predictor. In part (h) age was not significant AFTER taking into account illness duration. This suggests that illness duration explains the same things about (adjusted) memory score in patients that age does. (In fact illness duration presumably explains some additional variability since it is significant and age is not.) Another way to say this is that there is multicollinearity between age and illness duration. There are a number of different ways to check this theory. First we note that the sign on age is negative in the multiple regression, suggesting that memory improves as the person gets older which does not make sense (remember that memory normed is set up so that it says how much worse memory is than expected for a person's age so a negative coefficient on age actually corresponds to a "good" outcome.) One of the common effects of multicollinearty is to change or make non-sensical the direction of an observed relationship. Second we can calculate the correlation between age and illness duration. This calculation is shown below. The correlation is moderately strong at r = .6 though it is not outrageously high. This correlation is hardly unexpected. The longer a person has had a disease the older they must be. Finally, we can check the variance inflation factors. The VIF basically tells us how much bigger the standard errors of the coefficients are because of the relationships among the X variables. We can get these directly in STATA by using the estat vif command (printout shown below) or manually by fitting thre regression of age on illness duration and vice versa and getting the **. These calculations are also shown below. Either way we get VIFs of 1.58 for both variables. In a regression with two variables the VIFs will always be the same because the two variables must be equally correlated with each other. In a regression with 3 or more predictors this need not be the case. The VIFs of 1.58 are actually not that large by the rough rule of thumb given in our text. This is consistent with the correlation between age and illness duration being strong but not unduly so. In this case we do still want to take out one of the variables, age, because it is not telling us anything useful beyond illness duration. However the multicollinearity problem is not so strong that it has messed up the results for the illness duration variable. Illness duration was strongly significant and must be telling us things that age is not. This should not be surprising since we already have regressed out the effects we thought were really due to aging and are expecting any additional age effects to be due to illness progression which is correlated with age.

```
. corr age illnessdur
(obs=30)

             |      age illnes~r
-------------+------------------
         age |   1.0000
```

```
  illnessdur |   0.6043   1.0000

. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
         age |      1.58    0.634836
  illnessdur |      1.58    0.634836
-------------+----------------------
    Mean VIF |      1.58

. regress illnessdur age

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  1,    28) =   16.11
       Model |  792.725772     1  792.725772           Prob > F      =  0.0004
    Residual |  1378.15007    28  49.2196455           R-squared     =  0.3652
-------------+------------------------------           Adj R-squared =  0.3425
       Total |  2170.87585    29  74.8577878           Root MSE      =  7.0157

------------------------------------------------------------------------------
  illnessdur |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .3148681   .0784579     4.01   0.000     .1541545    .4755818
       _cons |   11.18161    4.57152     2.45   0.021      1.81728    20.54595
------------------------------------------------------------------------------

. regress age illnessdur

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  1,    28) =   16.11
       Model |  2919.80289     1  2919.80289           Prob > F      =  0.0004
    Residual |  5076.06377    28  181.287992           R-squared     =  0.3652
-------------+------------------------------           Adj R-squared =  0.3425
       Total |  7995.86667    29   275.71954           Root MSE      =  13.464

------------------------------------------------------------------------------
         age |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  illnessdur |   1.159736   .2889795     4.01   0.000     .5677887    1.751684
       _cons |   22.54077   8.676188     2.60   0.015     4.768402    40.31313
------------------------------------------------------------------------------
```

The formula for the variance inflation factor associated with variable X_j is obtained using the R^2 value given by regressing X_j on all the other predictors. With only two predictors this is just the square of the correlation between the predictors and will produce the same result for both. We have

$$VIF = \frac{1}{1 - R_j^2} = \frac{1}{1 - .3652} = 1.58$$

This is exactly what we obtained from STATA.

(j) The mediation diagram is a triangle with a direct arrow from your original predictor X to your outcome Y and an indirect path from X to Y through the potential mediatior, M (i.e. arrows from X to M and M to Y.) To check whether the mediation hypothesis is reasonable we need to know X is related to both Y and M, that M is related to Y. If these relationships don't hold there can't be mediation because either there is no relationship between X and Y for M to explain, or M is not related to one of X and Y and therefore can't be the reason for the relationship between them. For full mediation we need the relationship between X and Y to go away when M is included in the regression model and for partial mediation we need for the relationsip between X and Y to get weaker when M is added to the model. Either way the final step is a multiple regression of Y on both X and M. The printout below shows the four regression models. We see from the p-values that all the simple linear regressions are significant at $\alpha = .05$ and that in the multiple regression activation is significant while illness duration has become borderline not significant with a p-value of .08. We could call this full mediation. However given how close it is and the relatively small sample sizes there is a possibility that this is only partial mediation and that more data would give us a clearer picture of what is going on.

```
. regress memorynormed illnessdur

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  1,    28) =  708.12
       Model |  1670.39653     1  1670.39653           Prob > F      =  0.0000
    Residual |  66.0496837    28  2.35891727           R-squared     =  0.9620
-------------+------------------------------           Adj R-squared =  0.9606
       Total |  1736.44621    29  59.8774556           Root MSE      =  1.5359

------------------------------------------------------------------------------
memorynormed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  illnessdur |   .8771872   .0329639    26.61   0.000     .8096637    .9447108
       _cons |  -1.399109   .9896937    -1.41   0.168    -3.426405    .6281864
------------------------------------------------------------------------------

. regress memorynormed activation

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  1,    28) = 2605.78
       Model |  1717.98585     1  1717.98585           Prob > F      =  0.0000
    Residual |  18.4603616    28  .659298629           R-squared     =  0.9894
-------------+------------------------------           Adj R-squared =  0.9890
       Total |  1736.44621    29  59.8774556           Root MSE      =  .81197

------------------------------------------------------------------------------
memorynormed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  activation |   .9927719   .0194483    51.05   0.000     .9529339     1.03261
       _cons |  -4.757516   .5798436    -8.20   0.000    -5.945272    -3.56976
------------------------------------------------------------------------------

. regress activation illnessdur

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  1,    28) =  649.28
       Model |   1671.0305     1   1671.0305           Prob > F      =  0.0000
```

```
    Residual |   72.062905    28  2.57367518           R-squared     =  0.9587
-------------+------------------------------           Adj R-squared =  0.9572
       Total |  1743.09341    29  60.1066691           Root MSE      =  1.6043

------------------------------------------------------------------------------
  activation |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  illnessdur |   .8773537   .0344318    25.48   0.000     .8068234     .947884
       _cons |   3.561956   1.033764     3.45   0.002     1.444387    5.679525
------------------------------------------------------------------------------

. regress memorynormed illnessdur activation

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  2,    27) = 1410.85
       Model |  1719.98822     2  859.994109           Prob > F      =  0.0000
    Residual |  16.4579944    27  .609555348           R-squared     =  0.9905
-------------+------------------------------           Adj R-squared =  0.9898
       Total |  1736.44621    29  59.8774556           Root MSE      =  .78074

------------------------------------------------------------------------------
memorynormed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  illnessdur |   .1493684   .0824126     1.81   0.081    -.0197282    .3184651
  activation |   .8295615    .091971     9.02   0.000     .6408527     1.01827
       _cons |  -4.353971   .6003543    -7.25   0.000    -5.585796   -3.122146
------------------------------------------------------------------------------
```

(k) Optional Bonus: The following site has a calculator for the Sobel test for the strength of the idirect path: http://www.people.ku.edu/ preacher/sobel/sobel.htm

If you input the requested values from the regressions of activation on illness duration and memorynormed on activation you get a test statistic of 22.82 for the Sobel test and a p-value of 0 meaning that there is significant evidence of a significant indirect effect–i.e. mediation. This is not surprising given what we have seen from the printouts.