Solutions To Homework Assignment 6

(1) Regression Assumption Basics:

(a) There are basically four main assumptions we make when we fit a regression model. There is also a "pre-assumption" which I have listed first below for your reference.

(i) The X values in our data set are fixed and known, or measured without error. I didn't really discuss this in class and you shouldn't worry about it too much but I include it here for completeness. We make this assumption to simplify our calculations. If we have to assume that the X's have errors as well as the Y's it makes things much more complex. Sometimes this assumption is reasonable. For instance in early agricultural experiments, the X variable was often the amount of fertilizer used, and the response, Y, was the yield of the crop. Here the X variable was something the experimenter controlled. However, in many real world situations you do not control X. For instance, suppose I want to study house sales, and my X is the size of the house and my Y is the price of the house. Since I can't control what houses go on the market I can't control their sizes. Similarly, there may be measurement errors–measuring the size of a house exactly is not easy. However, it turns out that violations of this assumption do not usually cause serious problems and we will not worry about it.

Remember that the simpler linear regression model has the form

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where ϵ_i is the error term for the ith subject. You can think of this value as representing the subject's individual variability about the population regression line. The main assumptions we make are about the error terms, ϵ_i, and are very important.

(I) The ϵ_i's are mean 0, $E(\epsilon_i) = 0$. or equivalently a linear model is appropriate for describing the relationsip between Y and X. Obviously if the relationship between Y and X is not linear then there is no point to fitting a linear regression! There are many circumstances in which a linear model is not appropriate. For example, consider a situation where X is the price you charge for a product during a month, and Y is your profits for that month. For a while, increasing the price will increase your profits. (For instance, you will certainly make more money at \$1 per product than \$0 per product!) However, if you start to charge too much then people will stop buying your product and you will eventually lose money. Another example would be when X is the amount you spend on advertising in a month and Y is your total sales. For a while, additional advertising should boost your product, at least if the advertising agency is any good. However, there is a point of diminishing returns. Eventually, no matter how much more advertising you do, no more people will buy your product–the market will be saturated–and so the plot of Y vs X will level off–it will not continue as an increasing straight line forever.

How do we think of this in terms of the errors? If the errors are mean zero this implies that the data points are centered around the population regression line for all X values. This is exactly what it means to say we have the right shape of the relationship which is one reason why it is important. In addition, this assumption helps to ensure that our estimates, b_0 and b_1 for β_0 and β_1 are unbiased–that is that when we fit the estimated regression line we get the right answer on average. There are situations when the ϵ_i's will not be mean 0. For instance, suppose I was fitting a model as above where a straight line was not appropriate. If I fit a straight line to the profit vs price data there will be periods when my errors are all positive (i.e. above the regression line) and periods when my errors will be all negative (i.e. the points are below the regression line.) In both these cases, the average error will not be 0. To see this, draw the profit vs price curve–it looks

like an arch. Now draw in the best possible straight line. In some places the arch will be above the line and in some places below the line.

(II) The ϵ_i's are normally distributed. This is important because we want to be able to use the normal or t distribution to construct confidence intervals or perform hypothesis tests in the regression setting. If we assume the ϵ_i's are normal then it turns out that b_0 and b_1 are also normally distributed and we can use our standard tests and confidence intervals. The error terms, ϵ_i are not always normally distributed. For instance, suppose that our Y variable is really discrete–for instance, maybe it only takes on integer values, such as the number of products sold. Then at any value of X, there are only certain values of Y possible and hence only certain values of ϵ_i possible. But the normal distribution is continuous–it takes on every possible value. Thus if the Y's and the errors are discrete, they can't be normally distributed. You can also come up with situations where the residuals are likely to be skewed. For instance, suppose your Y variable is income and X is age. There is a lower bound on income (0–we hope!) but not an upper bound so we are likely to have more really big values than small ones and correspondingly probably more big positive than negative residuals.

(III) The ϵ_i's are independent. This is just like our assumption earlier in the course that our data come from a random sample. It is important because it allows us to get unbiased estimates of the quantities we are interested in. This assumption can also be violated. For instance, suppose we are studying house prices again, and we take as our sample all the houses in a certain neighborhood that have sold recently. Now houses in a neighborhood tend to be similar. If it is a good neighborhood then all the houses may be above the average price for their size–the prices of the houses are related, not independent. Similarly, consider the example of profit vs price from assumption (i). There will be a clear pattern (lack of independence) to the errors in this model IF WE USE A STRIAGHT LINE because errors in one section of the model will all be positive and errors in another section of the model will all be negative.

(IV) The ϵ_i's have constant variance, $\sigma^2_{Y|X}$. This means that the spread of data points about the population regression line is the same for all values of X. If you look at a plot of the data, the points form a band about the regression line. This assumption says that this band is always the same width. This assumption is important because it makes it much easier for us to come up with formulas for quantities like $s_{Y|X}$ and s_{b_1} which let us compute our confidence intervals and hypothesis tests. However, it can also be violated. For instance, consider the house price example. There tends to be a lot more variability in the price of big houses than of small houses. Just because they are so much more valuable there is a much wider range of possible prices.

(b) The plots are in the accompanying graphics file.

(i) The first figure shows a residual plot, which is essentially a scatterplot turned on it's side so you can see the errors relative to their mean of 0. We can use this to check whether the errors are mean 0, independent, and have constant variance. First, the mean 0 assumption is clearly violated. At the start the errors are all negative. Then for X between about 1.5 to 5 they are all positive. Then they are negative again and then become positive at the end. They are not centered about the 0 line for all X as they are supposed to be. The independence assumption is also violated by the pattern I just described. The errors go down and up and down and up again. It looks as if a curved model (in fact probably a cubic polynomial) would fit this data better. In general patterns in the data that show asymmetry about the 0 line indicate a mis-specification of the model has induced independence violation but one that can usually be fixed by picking a better shape (i.e. by doing a transformation). There are other sorts of independence violations (e.g. caused by outliers) that a transformation may not fix. The constant variance assumption for this model however looks OK. If we draw a band above and below the residuals following the up and down pattern, it stays a fairly similar width for each value of X. Therefore, I think this assumption is OK. Finally, there is one point which is out of line with the others. It is at about X=7.5 and has a positive residual when all the other points have negative

residuals, meaning it was well above the fitted line. This is certainly an outlier. However it probably doesn't have that high leverage since it isn't too far from the center of the data set. Exactly how influential it is is hard to tell without actually seeing the regression line snd getting some statistics like it's Cook's Distance or DFBetas. However given its relatively central X value and the fact that there are so many data points, the effect was probably fairly minimal. It is hard to check normality from this plot–for that we would need a histogram or normal quantile plot of the errors.

(ii) Here we are given a histogram and normal quantile plot of our residuals. These can be used to check normality but not any of the other assumptions. (While it is true that the histogram suggests the overall mean of the errors could be near 0, we need to check mean 0 for ALL X which the histogram can't tell us since it doesn't incorporate the X values!) Here the histogram is very skewed rather than bell-shaped and the points curve way away from a straight line on the normal quantile plot so the normality asusmption is clearly violated.

(iii) On the scatterplot and residual plot shown here, the mean 0 and independence assumptions are fine–the points are centered about the regression line or residual = 0 line for pretty much all the X values. However, the constant variance assumption is clearly violated. The points are widely spread out on the left, and very narrowly spread on the right. As usual, it is hard to check the normality assumption without a histogram or normal quantile plot.

There are several outlier/leverage/influential points here. They are at coordinates (1,-65), (1,43), (1,57) and (15,-90) on the scatterplot. The last point is going to have high leverage because it is so far from the center of the data but it is not actually going to be influential in the sense of changing the coefficients or fitted values of the estimated regression because it lies on the same straight line as the majority of the points. The others outliers are probably pulling the line towards themselves and hence could be considered influential considered singly. If I remove the point at (1, -65) the regression line will become less steep and if I remove one of the other two it will become more steep. Note however, that the effects of these 3 points tend to cancel each other out–if I took them out at the same time it probably wouldn't change things much. If I wanted to formally check this I could calculate statistics such as the standardized residuals, leverage values, Cook's distances, DFBetas, etc. and see which points had abnormally large values. Recall that Cook's distance and DFBetas tell you about the **actual influence of a point on the line**–i.e. how much the predicted values or coefficients would change if you removed the specified point–while the other values tell you more about **potential influence**–i.e. how far away is the point from the main data cloud.

(2) Curvilinear Regression Basics:

(a) The STATA scatterplot is shown in the accompanying graphics file. It looks as if the data may curve somewhat or begin to level off as X gets larger though it is a little hard to tell given the amount of individual variability in the data. Intuitively I would expect a curvilinear model to be better. For a while, giving the plants more fertilizer should be helpful–it gives them added nutrients and so forth. However, a plant can only absorb so many nutrients. Beyond a certain point there should be no gain from adding more fertilizer and we would expect the plot of Y vs X to level off. In fact it is even possible that too much fertilizer could be harmful in which case the plot of Y vs X could turn downward again as X gets very large. Combining the scatterplot with my intuitive reasoning I suspect a curvilinear model will be better for this data.

(b) From the STATA printout below, the estimated curvilinear regression equation is

$$\hat{Y} = 2.91 + 1.91X_1 - .17X_1^2$$

The fitted value at $X_1 = 5$ is

$$\hat{Y} = 2.91 + 1.91(5) - .17(25) = 8.21$$

Note that we had to square the X value before plugging it in for the quadratic term. In other words, when we use 500 pounds of fertilizer per acre (X=5) we expect the alfalfa to grow to an average height of 8.21 inches.

```
. reg growth fertilizer fertilizersq

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  2,    12) =   96.64
       Model |  43.8382878     2  21.9191439           Prob > F      =  0.0000
    Residual |  2.72171251    12  .226809376           R-squared     =  0.9415
-------------+------------------------------           Adj R-squared =  0.9318
       Total |  46.5600004    14  3.32571431           Root MSE      =  .47625

------------------------------------------------------------------------------
      growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  fertilizer |   1.911376   .2507408     7.62   0.000     1.365058    2.457693
fertilizersq |  -.1721461   .0402981    -4.27   0.001    -.2599481    -.084344
       _cons |   2.905534   .3243424     8.96   0.000     2.198853    3.612216
------------------------------------------------------------------------------
```

(c) The estimated simple linear regression is

$$\hat{Y} = 3.86 + .88X_1$$

The fitted value at $X_1 = 5$ is $\hat{Y} = 3.86 + .88(5) = 8.26$ We expect the alfalfa to grow to an average height of 8.26 inches when we use 500 pounds of fertilizer per acre.

```
. reg growth fertilizer

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  1,    13) =   75.23
       Model |  39.6993804     1  39.6993804           Prob > F      =  0.0000
    Residual |  6.86061997    13  .527739998           R-squared     =  0.8526
-------------+------------------------------           Adj R-squared =  0.8413
       Total |  46.5600004    14  3.32571431           Root MSE      =  .72646

------------------------------------------------------------------------------
      growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  fertilizer |   .8784992   .1012884     8.67   0.000      .659679    1.097319
       _cons |   3.864502   .3570947    10.82   0.000     3.093046    4.635959
------------------------------------------------------------------------------
```

(d) Looking at the scatterplot in (a), or for that matter at the original data, we see that when close to X=5 or 500 pounds of fertilizer per acre were used, the alfalfa never grew to a height of more than 8.2 inches.

Therefore the lower prediction obtained from the multiple regression model seems more appropriate. In fact, we have a data point $X_1 = 5.1$ which has Y value 7.7. The amount predicted by the linear regression is further away from this value than that predicted by the curvilinear regression.

(e) The curvilinear regression fits the data better. As we can see from looking at the plot, the data seem to follow a curve. Also, the R^2 and $R^2_{adjusted}$ values are much higher (93% vs 84%) in the curvilinear regression meaning the curvilinear regression is explaining more of the variability in alfalfa growth than the simple linear model. Similarly, RMSE is smaller in the curvilinear regression (.476 inches vs .927 inches) meaning we are making much smaller prediction errors on average when we use the curvilinear model. Finally, we can perform an hypothesis test to determine whether the curvilinear term, X^2 was worth adding to the model. After all, if $\beta_2 = 0$ it means that adding the X^2 term did not explain any additional variability or make our predictions better. Our hypotheses are
$H_0 : \beta_2 = 0$, i.e. The fertilizer squared term does not explain any additional variability in alfalfa growth beyond what is explained by the plain fertilizer term and so we might as well stick with the simple linear model. (Note that it is hard to give a physical interpretation to fertilizer squared!)
$H_A : \beta_2 \neq 0$, i.e. The fertilizer squared term does explain additional variability in alfalfa growth beyond what is explained by just fertilizer. It is therefore worthwhile to use a curvilinear model to predict alfalfa growth instead of a simple linear model.

From the STATA curvilinear regression printout we see that our test statistic is $t_{obs} = -4.27$ and the corresponding p-value is .001. We therefore reject the null hypothesis at $\alpha = .05$, or at almost any α for that matter, and conclude that the curvilinear model does a significantly better job of predicting alfalfa growth.

(f) First I decided to try a transformation of X that would result in the modeling leveling off at a maximum fertilizer growth. A good choice for such a transformation is $X' = 1/X$. If we fit the model $Y = b_0 + b1X' = b_0 + b_1\frac{1}{X}$ then as X gets very large $\frac{1}{X} \to 0$ and the predicted value will approach b_0. In particular I expect $b_1 < 0$ since the predicted values should be getting bigger (I am subtracting off less) as X gets bigger. I created a new variable in STATA using the command

gen invfertilizer = 1/fertilizer

The resulting regression printout is shown below. As expected the coefficient of the inverse fertilizer term is negative. The intercept which is the maximum growth level predicted as the amount of fertilizer gets really large is 7.82. This is a little disturbing since we had some actual values as high as 8.0. Now it is true that this curve tells you about the average Y as X gets very large should be 7.82 so maybe individual observations above 8 aren't too surprising but it is a bit suggestive. More importantly the R^2 and RMSE are actually much worse than even for the simple linear regression. This model is not an improvement! We could also have tried functions like lnX or $\sqrt{(X)}$ that don't ever level off but do not grow as quickly as a linear or quadratic model. However since we can actually see the points starting to turn back down at the highest fertilizer levels this probably won't be an improvement either.

```
. reg growth invfertilizer

      Source |       SS       df       MS              Number of obs =      14
-------------+------------------------------           F(  1,    12) =   18.17
       Model |   17.716222     1   17.716222           Prob > F      =  0.0011
    Residual |  11.7009215    12  .975076795           R-squared     =  0.6022
-------------+------------------------------           Adj R-squared =  0.5691
       Total |  29.4171435    13  2.26285719           Root MSE      =  .98746
```

```
------------------------------------------------------------------------------
      growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
invfertili~r |  -1.878046   .4405956    -4.26   0.001    -2.838022   -.9180711
       _cons |   7.818624   .3582866    21.82   0.000     7.037985    8.599264
------------------------------------------------------------------------------
```

Next I decided to try transforming the Y variable. Given the success of the quadratic model, one reasonable choice is a square root model, $Y' = \sqrt(Y) = b_0 + b_1 X$. This will probably alter the shape of the model in the same way as adding an X^2 term but it will have a rather different effect on the error terms. The command to create the new variable is

gen sqrtgrowth = sqrt(growth)

The resulting printout is shown below. This is much better than the inverse fertilizer model–it has an R^2_{adj} value back up around 80% again compared to the 57% for the inverse fertilizer model. However it still doesn't look as good as the quadratic model from part (b) where the percentage of variability explained was over 90%. One reason for the difference is that sqare rooting Y doesn't allow a completely arbitrary quadratic fit. If $\sqrt(Y) = b_0 + b_1 X$ then $Y = b_0^2 + 2b_0 b_1 X + b_1^2 X^2$. The three coefficients are related to each other in this case, where as in the quadratic model in part (b) they were allowed to vary freely. Also of course, fitting the model on the square root scale has changed the weight put on the different data points which can effect how good the fit looks. While the R^2 values are still in some sense roughly comparable, the RMSE values are certainly not since the units of measurement have changed. Overall the best model for this data appears to be the quadratic fit from part (b) but we could try other models too, perhaps even higher powers of X.

```
. reg sqrtgrowth fertilizer

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  1,    13) =   58.05
       Model |  1.73887906     1  1.73887906           Prob > F      =  0.0000
    Residual |  .389438825    13  .029956833           R-squared     =  0.8170
-------------+------------------------------           Adj R-squared =  0.8029
       Total |  2.12831789    14  .152022706           Root MSE      =  .17308

------------------------------------------------------------------------------
  sqrtgrowth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  fertilizer |   .1838587   .0241322     7.62   0.000     .1317242    .2359932
       _cons |   1.969954   .0850788    23.15   0.000     1.786152    2.153755
------------------------------------------------------------------------------
```

(3) Pregnancy Weight Gain:

(a) The scatterplot is in the accompanying graphics file. It seems that a linear model is NOT appropriate for these data. There is a clear upward curve in the data as X gets larger. There is one point that may be an outlier–a 23 year old woman who gained over 40 pounds while the next highest weight gain was just above 30 pounds. It is possible this point is influential. It's X value is at the high end of the age range and it's Y value is also quite unusual so it will have some leverage to pull the regression line towards itself. However it is not totally outside the main X range or that out of line with the general flow of the points so in fact it

may not be that bad. One ad hoc way to check this we could fit a model with and without the point and see if the answers changed. There are a large number of data points here so I would hope that the effect would not be too big. To check more formally, I have calculated some summary statistics for this model including the raw residuals, studentized residuals, leverage values, DFBeta (for the age variable), Cook's distances and DFFits. A subset of these values are in the graphics file, including the point of interest. This point certainly has the largest residual, studentized residual, DFBeta, Cook's distance and DFfit values. However it does NOT have the highest leverage value which suugests that its position in the overall cloud of points won't allow it to completely dominate the fit. To decide whether these values are in fact large we can refer to our standard rules of thumb. For standardized or studentized residuals which are supposed to have approcimate t-distributions, values above 2 or 3 in absolute value are generally considered large. (What value to use really depends on the size of the data set. With 100 points we would expect about 5% or 5 of our residuals to be above 2 and less than half a percent or probably none of our residuals to be above 3 in absolute value.) The point of interest has a studentized residual of 7.5 which is very large, definitely making it a point of interest. For leverage, h_{ii}, the values lie between 0 and 1 with values closer to 1 indicating higher potential influence. The rough rule of thumb is that points with leverage above $2(p+1)/n$, where p is the number of predictors and n is the sample size, are large. The logic essentially is that you can show that the average leverage should be $(p+1)/n)$ so points with more than twice that leverage are unusual. Here were have p = 1 since we have fit a simple linear regression and n = 100 so using the rough rule of thumb, leverage values above .04 are worthy of consideration. The leverage for our point of interest is right around .025, so it actually doesn't have that high leverage. It is worth noting that there are other points that have greater leverage than this one. There are similar rough rules of thumb for the other influence statistics. DFBetas for a particular variable assess how much the coefficient of that variable changes if a given point is left out. It is generally recommended to use absolute values greater than 1 for small to medium sized data sets and values greater than $2/\sqrt{n}$ for large data sets. Our data set with 100 points is reasonably large. Using the second rule of thumb, values above .2 would be large. Our point of interest has a DFbeta score of .92, by far the largest in the data set. it is having a noticable effect on the slope (coefficient) of the age variable. DFFits measures the impact of removing a point, i, on the fitted or predicted value, for that point, $\hat{Y}_i$. In this case for small to medium data sets the recommendation is to use absolute values above 1 for small to medium data sets and values above $2\sqrt{p/n}$ for large data sets. Since p = 1 this once again gives us a cutoff of .2 for our data set. The DFFits value for the point of interest is 1.19, high by either standard, so whether this point is included or not makes a big difference in our predictors for women aged about 23 and a half. Cook's distance gives a combined measure of the impact of removing point i on all the estimated regression coefficients or predicted values. It can be large because the point has high leverage, a large residual, or a combination thereof. For Cook's distance, there are many rules of thumb. Your books suggests that values greater than 1 are big and values bigger than 4 are really big. However there is some dependence on sample size. One common sample-size dependent approach is to compare the value D_i to an F distribution with p and n-p degrees of freedom and see what percentile value it corresponds to. If the Cook's distance is below the 20th percentile or so the point is not considered influential. If it is above the 50th percentile it is considered quite influential. There is a big grey zone in between. In any case, it is often good to see how large the next biggest Cook's distance is. Here we have $D_i = .457$ for our point of interest which is by far the largest Cook's distance in the data set. For an F distribution with 1 and 98 degrees of freedom this is approximately the 50th percentile (I got this using the Ftail command in STATA), so this point seems to be fairly influential by this measure. All in all it looks like the combination of this point's moderate leverage and large residual will result in it having a fairly large impact on a simple linear regression.

(b) The STATA regression printout is given below. There is clearly a statistically significant relationship between weight gain and mother's age based on this model–the p-value for the t-statistic of the age variable is .000 as is the p-value for the F test which, since this is a simple linear regression, is equivalent. It may seem odd to say that there is a significant linear relationship between the variables when in part (a) we said a linear model was inappropriate. However all that our test is saying is that a linear relationship accounts for SOME of the variability in weight gain. Using age in a linear model is better than not using it. Just

because this model represents an improvement does not mean it is the BEST model. Part (a) suggests that we can get an EVEN BETTER model if we fit a curved shape.

```
. reg weightgain momage

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =  221.36
       Model |  2922.82212     1  2922.82212           Prob > F      =  0.0000
    Residual |  1294.00783    98  13.2041615           R-squared     =  0.6931
-------------+------------------------------           Adj R-squared =  0.6900
       Total |  4216.82995    99  42.5942419           Root MSE      =  3.6338

------------------------------------------------------------------------------
  weightgain |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      momage |   2.316687   .1557118    14.88   0.000     2.007682    2.625692
       _cons |   -32.7315   3.177901   -10.30   0.000    -39.03794   -26.42506
------------------------------------------------------------------------------
```

(c) The plots are shown in the accompanying graphics file. The residual plot can be used to check the mean 0, independence and constant variance assumptions. We see that the mean 0 and independence assumptions are violated because the points on the residual plot are not centered about 0 for each X–first they are positive, then negative, then positive again–and in particular there is a curved shape suggesting that we are not fitting the best possible model. The constant variance assumption also appears to be violated. From the residual plot, if you draw a band above and below the points you see that it gets wider as the mother's age gets larger–the residuals fan out. Finally, we can use the histogram and normal quantile plot to check the normality assumption. The histogram should look bell-shaped and on the quantile plot the points should follow a straight line. (Recall that the quantile plot gives your residuals versus the ones you should get if normality is correct.) The histogram is clearly skewed to the right and on the quantile plot the points do not follow the line–they start above it, then go below it, and curve up above it at the end. Thus the normality assumption is also violated. All four of our assumptions are wrong, further confirming our answer to part (a)!

(d) The multiple regression printout for the quadratic model is given below. We see that it has a smaller RMSE (3.17 pounds versus 3.63 pounds) than the simple linear model and a higher R^2_{adj} (76.38% versus 69.00%) suggesting that this model has substantially improved our fit. We can also check this by performing a hypothesis test to see if the X^2 variable is statistically significant. We have

$H_0 : \beta_2 = 0$, i.e. The age squared term does not explain any additional variability in weight gain beyond what is explained by the plain age term and so we might as well stick with the simple linear model. (Note that it is hard to give a physical interpretation to age squared!)
$H_A : \beta_2 \neq 0$, i.e. The age squared term does explain additional variability in weight gain beyond what is explained by just age. It is therefore worthwhile to use a quadratic model to predict weight gain instead of a simple linear model.

From the STATA curvilinear regression printout we see that our test statistic is $t_{obs} = 5.62$ and the corresponding p-value is .000. We therefore reject the null hypothesis at $\alpha = .05$, or at almost any α for that matter, and conclude that the curvilinear model does a significantly better job of explaining weight gain.

```
 reg weightgain momage momagesq
```

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) =  161.07
       Model |  3240.92603     2  1620.46302           Prob > F      =  0.0000
    Residual |  975.903916    97  10.0608651           R-squared     =  0.7686
-------------+------------------------------           Adj R-squared =  0.7638
       Total |  4216.82995    99  42.5942419           Root MSE      =  3.1719

------------------------------------------------------------------------------
  weightgain |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      momage |  -12.32079   2.606697    -4.73   0.000    -17.49436   -7.147216
   momagesq |    .363564   .0646568     5.62   0.000     .2352382    .4918897
       _cons |   112.6111   25.99637     4.33   0.000     61.01549    164.2067
------------------------------------------------------------------------------
```

(e) The error diagnostic plots are shown in the graphics file. From the residual plot it seems as if the mean 0 and independence assumptions are much improved. The points are now mostly centered around the 0 line and there isn't a clear curved pattern. (If you squint it may seem the errors are a little negative for low X's and a little positive for middle X's but this overall is pretty minor–we will see the reason for it later.) Overall I would say these assumptions are OK. Constant variance still seems to be violated–the spread of the residuals is wider at higher ages. For normality, the histogram doesn't look too bad other than the outlier (which you should ignore in assessing the model assumptions) but it is still skewed to the right a bit. This becomes MUCH more obvious looking at the quantile plot where the points curve badly away from the line they are supposed to follow. This is why we bother with quantile plots–they are easier to read! It seems the normality assumption is still an issue. However overall this model is much better than the SLR of part (b).

(f) The regression printout is given below and the error diagnostic plots are in the graphics file. We can see that this version has by far the highest R^2_{adj} of our three models at 89.68%. From the residual plot we can see that the mean 0, independence and constance variance assumptions now all look fine–the points look like an even band of random scatter about the 0 line. Furthermore, the histogram looks much more bell-shaped and the quantile plot follows the straight line very well except for a few points at the edges (which is pretty common). Thus all our error assumptions are now satisfied. The inverse model is clearly the best and indeed is the one I used to create the data!

```
. reg invweightgain momage

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =  861.57
       Model |   .07994652     1   .07994652           Prob > F      =  0.0000
    Residual |  .009093535    98  .000092791           R-squared     =  0.8979
-------------+------------------------------           Adj R-squared =  0.8968
       Total |  .089040055    99  .000899394           Root MSE      =  .00963

------------------------------------------------------------------------------
invweightg~n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      momage |  -.0121162   .0004128   -29.35   0.000    -.0129353    -.011297
       _cons |   .3282028   .0084244    38.96   0.000     .3114849    .3449207
------------------------------------------------------------------------------
```

(g) The RMSE gives the error you make in the units of your Y variable. In the models from parts (b) and (d) we were using the original units of Y–pounds. However in the inverse model from part (f) we have transformed the Y variable–it is no longer on the original scale. Thus the RMSE for this model is not on the same scale as those from parts (b) and (d) and we can not compare them. This is another example of how keeping careful track of your units is important!

(4) Don't Drink and Derive:

(a) The two simple linear regression printouts are shown below. For an SLR we can look at the p-value for either the t or the F test to determine whether there is a significant relationship between X and Y. In this case the p-value for amount is .0012 and the p-value for time is .0001, both much less than $\alpha = .05$, so we conclude that both amount drunk and time since drinking are indivudally significant predictors of blood alcohol level. Not a surprise!

```
. regress bac amount

      Source |       SS       df       MS              Number of obs =     123
-------------+------------------------------           F(  1,   121) =   11.07
       Model |  .007920396     1  .007920396           Prob > F      =  0.0012
    Residual |  .086595255   121  .000715663           R-squared     =  0.0838
-------------+------------------------------           Adj R-squared =  0.0762
       Total |  .094515651   122  .000774718           Root MSE      =  .02675

------------------------------------------------------------------------------
         bac |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      amount |   .0029968   .0009008     3.33   0.001     .0012134    .0047802
       _cons |   .0409389   .0083054     4.93   0.000     .0244962    .0573816
------------------------------------------------------------------------------

. regress bac time

      Source |       SS       df       MS              Number of obs =     123
-------------+------------------------------           F(  1,   121) =   16.60
       Model |  .011401548     1  .011401548           Prob > F      =  0.0001
    Residual |  .083114103   121  .000686893           R-squared     =  0.1206
-------------+------------------------------           Adj R-squared =  0.1134
       Total |  .094515651   122  .000774718           Root MSE      =  .02621

------------------------------------------------------------------------------
         bac |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |  -.0134077   .0032909    -4.07   0.000    -.0199229   -.0068924
       _cons |   .0901186   .0060613    14.87   0.000     .0781186    .1021186
------------------------------------------------------------------------------
```

(b) To determine which variable is the better individual predictor there are many quantities we could look at including their correlations with Y, the R^2_{adj} values, the root mean squared error, F, p-value, etc.. We

see that time is individually the better predictor. It has larger R^2_{adj} and F values and a smaller root mean squared error and smaller p-value for the F/t tests.

(c) The plots are shown in the accompanying graphics file. The assumptions are that the errors are normally distributed, mean 0 (linear model is the right shape), independent and have constant variance. To check normality we look at the histogram and normal quantile plot. The histogram should be symmetric and hump-shaped. The points on the quantile plot should follow a straight line. Here the histogram looks roughly symmetric except for the outlier although it doesn't tail off as much at the edges as one might like. The points do seem to follow the line on the quantile plot fairly well. Thus overall I would say the normality assumption is decent though not fantastic. For the other assumptions we look at the residual and scatter plots. Here all the assumptions are clearly violated. There is a hugely curved pattern to the data indicating that the mean 0 and independences assumption are violated. For low and high X, all the errors are negative while for X values in the middle most of the errors are positive. Note that the errors have to be mean 0 at EACH X value–it is not enough for them to be mean 0 overall. This indicates we should have fit a curved model, probably a parabola. Finally, the band of errors seems to be much narrower for small and large X values than it is for X values in the middle. Thus the constant variance assumption is violated. Our snake has a fat belly and small head and tail!

We were also asked in real world terms why these violations make sense. The mean 0 and independence violations have both resulted from us fitting the wrong shaped model to the data–we should have fit an upside-down parabola shaped model. This makes sense. When little time has passed the alcohol the person has drunk will not have fully gone into the blood stream. At some point all the alcohol will be in the blood and will be starting to be metabolized from which point the BAC will start decreasing again. Eventually after enough time has passed all the alcohol will be gone and the BAC will return to 0. For the constant variance assumption the explanation is similar. For very short and long times there can be very little alcohol in the blood so there just isn't room for much variation. However in the middle once all the alcohol has gotten into the bloodstream the BAC will depend heavily on how much the person has drunk and there will be a lot of variability.

(d) There is one point in the scatterplot that is unusual–a person who has 0 bac and had their drinks 0 minutes ago. If we create the scatterplot for bac versus amount we see that this person has actually drunk a very large amount. From a real-world point of view this makes perfect sense. Although the person has drunk a lot the alcohol has not had time to get into their blood. I calculated the various outlier diagnostics we have learned for this data: studentized residuals, leverage, DFBetas, DFfits, and Cook's distance. The values corresponding to the last few points in the data set including the point of interest are shown below.

leverage	rstudent	dfbeta(time)	dffit	cooks d
.028046	-2.091532	-.2993929	-.355285	.0614013
.0287611	-.9789661	-.1426806	-.1684641	.0141950
.0294888	-1.478777	-.2193766	-.2577693	.0328998
.0302292	-1.436089	-.2167874	-.2535478	.0318635
.0309821	-2.037189	-.3128443	-.3642682	.0646622
.0317477	-.8001387	-.1249652	-.1448863	.0105273
.0534871	-3.716765	.8136246	-.8835403	.3529437

We see that the last point, which is our person with 0 bac 0 minutes after drinking a very large amount, has by far the hiest values of all the outlier diagnostic statistics. We use the same rules of thumb as in problem 3 above to evaluate what has happened. For leverage, h_{ii}, the rough rule of thumb is that points with leverage above $2(p+1)/n$, where p is the number of predictors and n is the sample size, are large. Here were have p = 1 since we have fit a simple linear regression and n = 123 so using the rough rule of thumb, leverage values above .033. are worthy of consideration. The leverage for our point of interest is well above this cutoff, so

we would consider it to have high leverage. None of the other points shown rise to this standard. Our rules of thumb suggest that points with studentized residuals above 3 for a data set this size are unusual. Here our point of interest has a residual of 3.7 in absolute value so it is definitely an outlier. For DFBetas, values greater than $2/\sqrt{n}$ are considered big for large data sets. Here, values roughly above .2 would be large. Our point of interest has a DFbeta score of .81, by far the largest in the data set. It is having a noticable effect on the slope (coefficient) of the time variable. DFFits measures the impact of removing a point, i, on the fitted or predicted value, for that point, $\hat{Y}_i$. Values above $2\sqrt{p/n}$ in absolute value are considered important for large data sets. Since p = 1 this once again gives us a cutoff of .2 for our data set. The DFFits value for the point of interest is -.883, well above the cutoff, so whether this point is included or not makes a big difference in our predictions for people who have just had their drinks (time = 0). Cook's distance gives a combined measure of the impact of removing point i on all the estimated regression coefficients or predicted values. It can be large because the point has high leverage, a large residual, or a combination thereof. One common rule of thumb is to compare the value D_i to an F distribution with p and n-p degrees of freedom and see what percentile value it corresponds to. If the Cook's distance is below the 20th percentile or so the point is not considered influential. If it is above the 50th percentile it is considered quite influential. There is a big grey zone in between. In any case, it is often good to see how large the next biggest Cook's distance is. Here we have $D_i = .353$ for our point of interest which is by far the largest Cook's distance in the data set. For an F distribution with 1 and 121 degrees of freedom this is approximately the 45th percentile, so this point seems to be fairly influential by this measure. All in all it looks like the combination of this point's high leverage and large residual will result in it having a fairly large impact on a simple linear regression. Note however that this does NOT mean the point will have great influence when we fit a curvilinear model. In fact the point seems to be right in line with the basic curved pattern of the data. Whether a point is influential can depend a lot on which model you fit!

(e) The simple linear regression printout is shown below:

```
. regress bac time timesq

      Source |       SS       df       MS              Number of obs =     123
-------------+------------------------------           F(  2,   120) =  104.68
       Model |  .060080355     2  .030040177           Prob > F      =  0.0000
    Residual |  .034435296   120  .000286961           R-squared     =  0.6357
-------------+------------------------------           Adj R-squared =  0.6296
       Total |  .094515651   122  .000774718           Root MSE      =  .01694

------------------------------------------------------------------------------
         bac |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |   .1226603   .0106615    11.51   0.000     .1015514    .1437693
      timesq |  -.0405559   .0031138   -13.02   0.000    -.0467211   -.0343908
       _cons |  -.0030843   .0081583    -0.38   0.706    -.0192371    .0130684
------------------------------------------------------------------------------
```

Here we are testing whether the quadratic (parabola) model is better than the straight line so our hypotheses will be about the shape of the relationship between time since drinking and BAC:

$H_0 : \beta_2 = 0$–the quadratic model is not an improvement over the linear model in time. It is not worth adding the Time^2 term.
$H_o : \beta_2 \neq 0$–the curvilinear model does a superior job of explaining the relationship between BAC and time. It is worth adding the Time^2 variable.

Looking at the STATA printout the p-value for the t-test of Time^2 is 0.000 which is certainly less than our $\alpha = .05$ so we reject the null hypothesis and conclude that the curvilinear model does a better job. Note that you can NOT look at the F test in this problem. The F test asks whether the model as a whole is useful–that is whether the combination of a linear and quadratic (time and time^2) variables is better than not using time in the model at all. This does not address the question of whether a curved model is better than the linear model.

(f) This is an example of a situtation where whether or not the intercept is 0 is actually a sensible question. From the printout we see that b_0 is NOT significantly different from 0. The corresponding p-value is .706, much higher than $\alpha = .05$, so we can not reject the null hypothesis ($H_0 : \beta_0 = 0.$) This means that according to this model it is possible that 0 hours after drinking (i.e. right as they drink) a person's BAC will be 0. This makes perfect sense since with no time elapsed the alcohol will not have had time to get into the blood!

(g) For the simple linear regression of BAC on amount, if we remove the outlier the model should fit better. Therefore R^2 and t_{obs} should go up–the percentage of variability explained should go up and our relationship as measured by the t-test should be more significant. In this case the basic relationship is increasing and the outlier is on the right hand edge of the data set and lower than expected. Therefore removing it will make the line slope up more steeply and b_1 should increase.

The situation is a little more complicated for the outlier as part of the curvilinear model of BAC on time and time squared. Note that the point lies essentially EXACTLY on the curve that is fit to the data. Thus the regression coefficients, including b_2, will remain unchanged whether the point is in or out. So will SSE. This is because the errors for all the current points will remain the same and since B lies exactly on the curve, it's error contribution will be 0. SST for this model will actually decrease if we take the point out. This is because its Y value is below the current mean and is at the most extreme distance from the current mean of any point so it can only make the overall variation greater. The RMSE for this model will increase. This is because $RMSE = sqrt(SSE/(n-3))$ in this model. As we have already seen, SSE stays the same, but our sample size has been decreased by 1 so RMSE goes up. The idea is that we have lost an extra point that was confirming that our model is fitting well so we are actually slightly less confident that the predictions are good. Finally, consider F and it's p-value. We see that

$$F = \frac{MSR}{MSE} = \frac{SSR/2}{SSE/(n-3)} = \frac{SSR}{SSE} * \frac{n-3}{2}$$

As we have already seen, SSE stays the same. But SST got smaller and SSR = SST - SSE so this means SSR must also have decreased. Similarly, with the extra point, n-3 has gotten smaller. Thus overall we see that F has gotten smaller. For the p-value, the degrees of freedom for the numerator has stayed the same, but the degrees of freedom for the denominator have decreased. As you decrease the degrees of freedom in the denominator of an F statistic the value of F required to get a particular p-value goes up. This is because with less data we need a biggera difference from the null hypothesis to be convinced. All else equal, the mode data we have, the surer we our of our result. Since our F has gotten smaller, correspondingly our p-value must have gotten bigger.

(h) Multicollinearity means a strong relationship between two or more of our predictor variables. Here our only predictors are time and time squared so we just need to calculate the correlation between them. As we can see from the printout below, the correlation is extremely strong, nearly .98! Thus we definitely have multicollinearity and it may be affecting the stability of our parameter estimates although it hasn't stopped us from correctly concluding the quadratic model is the better fit. This is an example of a structural multicollinearity, induced because of the way we defined our quadratic term. We can create "centered" versions of the time and time squared variable by subtracting off the mean. The corresponding STATA commands are shown below. The correlation between the centered versions of the variables is only -.05, meaning that centering has removed the multicollinearity. The regression printout for the centered variables

is also shown below. The standard error of the quadratic term is about the same at .003 but the standard error of the linear term has decreased dramatically from .01 to .002, suggesting that we have significantly stabilized our fit. The original estimated regression equation was

$$\hat{Y} = -.00308 + .12266time - .040556timesq$$

Using the centered variables we get

$$\begin{aligned}
\hat{Y} &= .08829 - .01491timecent - .04056timecentsq \\
&= .08829 - .01491(time - 1.69610) - .04056(time - 1.69610)^2 \\
&= .08829 - .01491time + .02528 - .04056timesq + .13759time - .11668 \\
&= -.00311 + .12268time - .04056timesq
\end{aligned}$$

We see that the two equations are the same up to rounding, so we get exactly the same predictions. We are just able to make more precise statements about the various parameter estimates in the second version of the model.

```
. cor time timesq
(obs=123)

             |     time   timesq
-------------+------------------
        time |   1.0000
      timesq |   0.9799   1.0000


. summarize time

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        time |       123    1.696098    .7210223          0       2.92

. gen timecent = time - 1.696098

. gen timecentsq = (time - 1.696098)^2

. cor timecent timecentsq
(obs=123)

             | timecent timece~q
-------------+------------------
    timecent |   1.0000
  timecentsq |  -0.0543   1.0000

. regress bac timecent timecentsq

      Source |       SS       df       MS              Number of obs =     123
-------------+------------------------------           F(  2,   120) =  104.68
       Model |  .060080355     2  .030040178           Prob > F      =  0.0000
    Residual |  .034435296   120  .000286961           R-squared     =  0.6357
-------------+------------------------------           Adj R-squared =  0.6296
```

```
       Total |  .094515651   122  .000774718           Root MSE      =  .01694

------------------------------------------------------------------------------
         bac |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    timecent |  -.0149133   .0021302    -7.00   0.000     -.019131   -.0106957
  timecentsq |  -.0405559   .0031138   -13.02   0.000    -.0467211   -.0343908
       _cons |   .0882904   .0022161    39.84   0.000     .0839027    .0926781
------------------------------------------------------------------------------
```

(i) The printout for the multiple regression is shown below. (I included the versions both centering and not centering the time and time squared terms. As you can see only the intercept and the coefficient of time are affected by this and as above if we multiplied out we'd get the same predictive equation.) If wine is the reference category then we must use the indicators for the beer and hard liquor categories and not the wine indicator. (If we include indicators for all three groups we will have perfect multicollinearity.) To decide overall whether this set of variables is useful for explaining blood alcohol levels we look at the F test. The p-value is 0 to as many decimal places as STATA gives it so we conclude that these variables are as a group useful for explaining blood alcohol levels. This is hardly a surprise since many of the variables were individually significant predictors.

```
. regress bac amount time timesq weight meal beer liquor

      Source |       SS       df       MS              Number of obs =     122
-------------+------------------------------           F(  7,   114) =  110.41
       Model |  .078377714     7  .011196816           Prob > F      =  0.0000
    Residual |  .011560947   114  .000101412           R-squared     =  0.8715
-------------+------------------------------           Adj R-squared =  0.8636
       Total |  .089938661   121  .000743295           Root MSE      =  .01007

------------------------------------------------------------------------------
         bac |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      amount |   .0044612   .0003565    12.52   0.000      .003755    .0051673
        time |   .1205641   .0072245    16.69   0.000     .1062525    .1348757
      timesq |  -.0404138   .0020716   -19.51   0.000    -.0445176   -.0363099
      weight |  -.0000715   .0000256    -2.79   0.006    -.0001222   -.0000208
        meal |  -.0042013   .0018949    -2.22   0.029     -.007955   -.0004476
        beer |  -.0095386   .0021562    -4.42   0.000    -.0138101   -.0052671
      liquor |   .0211892   .0022122     9.58   0.000     .0168068    .0255716
       _cons |  -.0252424   .0081502    -3.10   0.002    -.0413879    -.009097
------------------------------------------------------------------------------

. regress bac amount timecent timecentsq weight meal beer liquor

      Source |       SS       df       MS              Number of obs =     122
-------------+------------------------------           F(  7,   114) =  110.41
       Model |  .078377715     7  .011196816           Prob > F      =  0.0000
    Residual |  .011560946   114  .000101412           R-squared     =  0.8715
-------------+------------------------------           Adj R-squared =  0.8636
       Total |  .089938661   121  .000743295           Root MSE      =  .01007
```

```
------------------------------------------------------------------------------
         bac |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      amount |   .0044612   .0003565    12.52   0.000      .003755    .0051673
    timecent |  -.0165274   .0013104   -12.61   0.000    -.0191232   -.0139315
  timecentsq |  -.0404138   .0020716   -19.51   0.000    -.0445176   -.0363099
      weight |  -.0000715   .0000256    -2.79   0.006    -.0001222   -.0000208
        meal |  -.0042013   .0018949    -2.22   0.029     -.007955   -.0004476
        beer |  -.0095386   .0021562    -4.42   0.000    -.0138101   -.0052671
      liquor |   .0211892   .0022122     9.58   0.000     .0168068    .0255716
       _cons |   .0629858   .0059701    10.55   0.000      .051159    .0748126
------------------------------------------------------------------------------
```

(j) The regression without the beer and liquor indicators is shown below. (Note that I have left out the outlier point in the results that follow.) We need to do a partial F test to figure out if the alcohol type variables have made a significant contribution. The hypotheses are

$H_0 : \beta_6 = \beta_7 = 0$–the type of alcohol drunk does not help explain a person's blood alcohol level beyond what is explained by the amount drunk, the time since it was drunk, the person's weight and whether the drinks were consumed with a meal. In other words, it is not worth adding the pair of indicators for beer and liquor to the model.
$H_A : \beta_6 \neq 0$ or $\beta_7 \neq 0$ or both–the type of alcohol consumed does provide additional information about BAC beyond what is explained by the other variables and it is worth adding the indicators for type to the model.

The test statistic for a partial F test is

$$F = \frac{(SSE_{red} - SSE_{full})/(p_{full} - p_{red})}{SSE_{full}/(n - p_{full} - 1)} = \frac{(SSR_{full} - SSR_{red})/(p_{full} - p_{red})}{SSE_{full}/(n - p_{full} - 1)}$$

where the subscripts "red" and "full" refer to whether the quantitities are taken from the full model (the one with the extra variables added) or the reduced model (the one without the extra variables). $p_{full} - p_{red}$ is just the difference in the number of predictors in the model or if you prefer the difference in the degrees of freedom for regression. Notice that we can write the F statistic in terms of either the difference in total error between the models (the SSE values) or the difference in the total variability explained (the SSR values) because SST stays fixed throughout. Here our full model has 7 variables and the reduced model has 5 (taking out the two indicators.) Plugging in the sums of squares for regression (SSR) from the printouts we get

$$F = \frac{(.07838 - .06907)/2}{.01156/114} = 45.91$$

Under the null hypothesis the test statistics should have an F distribution with 2 and 114 degrees of freedom. The corresponding p-value is essentially 0 (checked using the STATA Ftail command) so we reject the null hypothesis and conclude that type of alcohol does help explain additional variability in BAC beyond what is explained by amount, time since consumption, weight and whether the alcohol was drunk with a meal. The corresponding printout of the STATA joint test of the coefficients is also given below and gives the same answer up to rounding. Note that I had to do the test after fitting the FULL model, not after the REDUCED model because the reduced model does not contain the parameters we are trying to test!

```
. regress bac amount timecent timecentsq weight meal

      Source |       SS       df       MS              Number of obs =     122
```

```
-------------+------------------------------           F(  5,   116) =   76.80
       Model |  .069072705     5  .013814541           Prob > F      =  0.0000
    Residual |  .020865956   116  .000179879           R-squared     =  0.7680
-------------+------------------------------           Adj R-squared =  0.7580
       Total |  .089938661   121  .000743295           Root MSE      =  .01341

------------------------------------------------------------------------------
         bac |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      amount |   .0040414   .0004707     8.59   0.000     .0031092    .0049737
    timecent |  -.0165418   .0017398    -9.51   0.000    -.0199876    -.013096
  timecentsq |  -.0396229   .0027469   -14.42   0.000    -.0450635   -.0341823
      weight |  -.0000339   .0000336    -1.01   0.315    -.0001005    .0000327
        meal |  -.0038144   .0025212    -1.51   0.133    -.0088079    .0011791
       _cons |   .0603448   .0077949     7.74   0.000     .0449061    .0757836
------------------------------------------------------------------------------

. test beer = liquor = 0

 ( 1)  beer - liquor = 0
 ( 2)  beer = 0

       F(  2,   114) =   45.88
            Prob > F =    0.0000
```

(k) We did not really need to do the test in part (j) because from the printout for the full model, the individual indicators for both beer and liquor were highly significant, telling us that both β_5 and β_6 were non-zero and that all three types of alcohol were different in terms of their impact on BAC. However it is possible to imagine a situation where it would be useful to do this test as follows. Suppose your reference group is the middle group in terms of average Y value (as indeed it was here with wine–beer was associated with a lower BAC than wine and liquor with a higher BAC). It could be the case that your evidence was insufficient to prove a difference between either of the higher or lower groups and the reference group, but that the extreme groups WERE significantly different from each other. In this case neither indicator would look significant but the F test would tell us that the groups were not all the same.

(l) The STATA commands and output are given below. We need to create an indicator for each of the indicator variables "beer" and "liquor" by multiplying them by the amount variable. To test whether the interaction of type by amount was useful we again need to perform a partial F test, this time for whether the pair of interaction variables was useful. The corresponding STATA test is shown below and is not quite significant with a p-value around .1. This tells us we do not have sufficient evidence to be 95% sure that the relationship between the amount you drink and your blood alcohol concentration depends on what it is you are drinking. This is perhaps a little surprising. Specifically we would expect that your BAC does not go up as fast if you are drinking beer as if you are drinking wine and it goes up faster with hard liquor than with wine, assuming the same amounts of each are being consumed since beer contains less alcohol per unit volume than wine which in turn contains less alcohol per unit volume than most kinds of hard liquor. What actually happened here is that when I constructed the outcome data I didn't include the interaction and so when I added this part of the question after the fact it didn't work. See if you can think up some real-world scenarios that would lead to the same result.....

```
. gen beeramount = beer*amount
```

```
. gen liqamount = liquor*amount

. regress bac amount timecent timecentsq weight meal beer liquor
beeramount liqamount

      Source |       SS       df       MS              Number of obs =     122
-------------+------------------------------           F(  9,   112) =   88.28
       Model |  .078826964     9  .008758552           Prob > F      =  0.0000
    Residual |  .011111697   112  .000099212           R-squared     =  0.8765
-------------+------------------------------           Adj R-squared =  0.8665
       Total |  .089938661   121  .000743295           Root MSE      =  .00996

------------------------------------------------------------------------------
         bac |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      amount |    .005659   .0006932     8.16   0.000     .0042854    .0070325
    timecent |  -.0162346   .0013146   -12.35   0.000    -.0188393     -.01363
  timecentsq |  -.0402505   .0020712   -19.43   0.000    -.0443544   -.0361466
      weight |  -.0000649   .0000258    -2.51   0.013     -.000116   -.0000138
        meal |  -.0035674   .0019014    -1.88   0.063    -.0073349    .0002001
        beer |   .0024254    .008053     0.30   0.764    -.0135306    .0183813
      liquor |   .0263959   .0075409     3.50   0.001     .0114546    .0413372
  beeramount |   -.001351   .0008741    -1.55   0.125    -.0030829     .000381
   liqamount |  -.0006403    .000849    -0.75   0.452    -.0023226    .0010419
       _cons |   .0508148   .0082663     6.15   0.000     .0344362    .0671935
------------------------------------------------------------------------------

. test beeramount = liqmount = 0

 ( 1)  beeramount - liqamount = 0
 ( 2)  beeramount = 0

       F(  2,   112) =    2.26
            Prob > F =    0.1087
```

(m) In backwards stepwise regression you start with all the possible predictors in your regression model and check to see if any of them are not significant. If there are some you remove first the one with the highest p-value. In the model in part (i) all the predictors are significant so our model is fine as is and the stepwise procedure wouldn't remove anything. If we instead used the model above with the interaction terms the stepwise procedure would first want to remoce the liquor by amount interaction but we need to be careful about that because it is really part of a "collective" variable, the interaction between type and amount, and the hierarchical principal tells us that we want to treat that as a single unit rather than as two separate variables. Viewed that way the F test in part (m) would tell us to remove the pair of interaction variables as its p-value is not significant.

(n) In forward stepwise selection you start with no predictors in the model and you add variables one at a time, using at each step the one that most improves the current model. At the first step this is just the variable that is the best individual predictor of Y which we can determine using a set of simple linear regressions or correlations. The STATA correlation printout is shown below. The time variables have the strongest correlations with bac so we would add them first. However again we need to be a little careful as

they are really a unit and it doesn't make sense to put in time squared without time. Probably we would put them in as a group and continue the procedure from there or else start qith the linear term and seeing whether time-squared or something else was the next best predictor.

```
. cor bac amount timecent timecentsq weight meal beer liquor
(obs=122)

             |      bac   amount timecent timece~q   weight     meal     beer
-------------+---------------------------------------------------------------
         bac |   1.0000
      amount |   0.3623   1.0000
    timecent |  -0.4139   0.1046   1.0000
  timecentsq |  -0.6854  -0.0355   0.0441   1.0000
      weight |  -0.0590  -0.0524  -0.0656   0.0580   1.0000
        meal |  -0.0391   0.0924  -0.0021   0.0268  -0.2511   1.0000
        beer |   0.0011  -0.0309   0.0665  -0.0754   0.0461   0.0172   1.0000
      liquor |   0.2088  -0.1304   0.0101   0.0133   0.1647  -0.0174   0.4522
```

Turn-In Problems

(5) Drinking the Milk of Paradise:

(a) The printout for the simple linear regression of sleep length on amount of milk drunk is shown below. The relationship IS significant since the p-value for the F test (or the t-test for the Milk variable) is 0.0000 which is certainly much less than $\alpha = .05$. Thus we reject the null hypothesis that $\beta_1 = 0$ and conclude that there is a significant relationship between amount of milk drunk and length of time slept.

```
. regress sleeplength milk

      Source |       SS       df       MS              Number of obs =     101
-------------+------------------------------           F(  1,    99) =   26.31
       Model |  1006037.51     1  1006037.51           Prob > F      =  0.0000
    Residual |  3785774.02    99  38240.1416           R-squared     =  0.2099
-------------+------------------------------           Adj R-squared =  0.2020
       Total |  4791811.53   100  47918.1153           Root MSE      =  195.55

------------------------------------------------------------------------------
 sleeplength |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        milk |    36.4985   7.115864     5.13   0.000     22.37908    50.61791
       _cons |   382.9712   39.08246     9.80   0.000     305.4231    460.5193
------------------------------------------------------------------------------
predict milkres, residuals
(22 missing values generated)

. scatter sleeplength milk

. scatter milkres milk
```

```
. histogram milkres
(bin=10, start=-533.5899, width=95.924539)

. qnorm milkres
```

(b) The various plots are shown in the accompanying graphics file. The commands used to create the residuals and obtain the plots are shown after the SLR fit above. We assume that the errors are mean 0, independent, constant variance, and normally distributed. The first three are checked with the residual plot (or with a scatterplot though it's easier to see on the residual plot because the tilted line has been removed) while normality is checked with a histogram and/or normal quantile plot. Here the histogram is fairly symmetric and hump-shaped and the points on the normal quantile plot follow a straight line pretty well so the normality assumption seems reasonable. However on the residual plot, the points are not centered around 0 for all X–for low values of mulk consumption more of them are negative, then for middle values more of them are positive and there are higher values near 400, and then for high values they are more negative again. Thus the mean 0 assumption is violated. The independence assumption is also violated by this pattern–we see a bit of a curved shape in the residual plot suggesting we have fit the wrong model though this is a little obscured by the outlier. Finally to check the constant variance assumption we draw a band above and below the points and see if it is of even width for all X. It looks to me as if the band is a bit narrower for low values of milk consumption, in which case the assumption is violated, but this is not as severe a problem as the mean 0 or independence assumptions. We gave credit on this as long as you explained your reasoning correctly.

(c) There does appear to be at least one severe outlier, point 101, a baby who has drunk about 5.5 ounces of milk but who has slept less than an hour. Because we have fit the wrong shaped model there may be points at the edges that are highly influential in determining the ultimate slope of the line but other than that one point none of them look particularly "out of line." This baby apparently had something bothering it besides being hungry. Maybe it was awakened by a loud noise or a bad dream. I even had someone suggest once that the baby spat up all it's milk after drinking so that although it had officially drunk a lot in practice it didn't get the benefit of this. We accepted any plausible explanation. To check my conjectures about what points were unusual more systematically, I obtained various regression diagnostics including the studentized residuals, leverage values, dffits, dfbetas and Cook's distances. I actually dumped these into Excel and sorted by the various measures to identify the worst points in each category, identified by their number in the original data set. You can get STATA to list points with diagnostic values above a certain value by using the list statement with the if option as well. The worst values of each type are shown below. Note that just because a point is flagged by one of these diagnostics as unusual doesn't mean it IS a PROBLEM or necessarily has to be removed. It just means it is worth further investigation. Here with the exception of the baby who barely slept at all, none of these points look extraordinary or as if they are out of line with the final model I want to fit. In fact if you refit the diagnostics after obtaining the more appropriate curvilinear models later on they will not looks as bad. I therefore recommended that you eliminate only point 101. Below I provide a more detailed analysis.

```
Highest Studentized residuals:
89      2.234925
79      2.099927
96      2.054799
55     -2.445838
101    -2.839412

Highest leverage:
98      0.0451884
```

```
55      0.0433032
64      0.0422855

Highest DFBetas (milk):
96      0.2810325
90      0.240408
43      0.2006251

Highest DFFITS:
96      0.3492722
90      0.2811900
89      0.2579806
84      0.2222547
93      0.2153620
79      0.2139226
95      0.2030985

Highest Cook's Distances:
55      0.1288984
96      0.0590729
17      0.057432
101     0.0403884
```

According to our various rules of thumb, points with studentized residuals above 2-3 are severe outliers. How big a value you should look for depends on the size of the data set. Since the studentized residuals have roughly a t-distribution, with a large data set you expect about 5% of points to have a value above 2 in absolute value and less than 1% to have a value above 2.5. Thus with 100 points we would expect 5 points above 2 by chance but wouldn't necessarily expect points above 2.5 in absolute value by chance. By this standard the only unusually high outlier is point 101, the point that we flagged visually. I listed the other points with studentized residuals above 2 in absolute value for reference. Our rules for leverage say that points with leverage above $2(p+1)/n = 2(2)/101 = .04$ are high. There are only 3 points that meet this standard (though there a number of others that are close). These babies are all ones that drunk a very large amount of milk, putting them right at the edge of our plot which is why they have high leverage. However none of them seem hugely out of line in terms of their sleep times. For DFBetas points above $2/\sqrt{n}$ are considered large according to the ruls of thumb we learder for large data sets. Here we have n = 101 so this gives a cutoff of roughly .2. There are three points that meet this standard, one of them just barely. These babies drank quite large but not the largest amounts of milk and had the highest sleep times. Since our simple linear regression line sloped up these points had the largest effect on pulling the line up. However none of them are really separated from the rest of the cloud of points. For DFFITS, values above $2\sqrt{p/n}$ in absolute value are considered important for large data sets. Since p = 1 this once again gives us a cutoff of .2 for our data set. There are a larger number of such points but the top 2-3 are much more severe than the others and are the same points that gave us the high DFBetas. Finally we look at the Cook's distances. For this there are many rules of thumb. I usually look to see if there are points with much bigger Cook's distances than the others. By this standard, point 55 is far and away the most severe. It is a baby who drank a lot (high leverage) and had quite low sleep time which doesn't fit with the upward trend of the simple linear regression. This combination makes it look extreme. Howevere when we fit the correct quadratic model to the data this point will be right in line with the model and will no longer look like a problem. Alternatively one can compare the Cook's distance to the percentiles of an F distribution withe degrees of freedom corresponding to the overall F test for the model being fit. Values below the 20th percentile are considered negligible and values above the 50th percentile are considered problematic with a big gray zone in the middle. Here we have $F_{1,99,.2} = .064$ and $F_{1,99,.5} = .458$ which I got using the inverse Ftail command

in STATA. Only point 55 is even in the gray zone–it falls at the 28th percentile, so in fact none of our points are that worrisome by the Cook's distance standards. The points with the next highest Cook's distances are the high leverage/DFBETAs points noted earlier plus point 101, the one with the huge residual. There are only really three points that stand out in this discussion: point 101 which is the one that is visually inconsistent with the others and has the highest studentized residual and a high Cook's distance, point 55 which has a relatively speaking unusual Cook's distance, the second highest leverage and is the only other point with a unusually high studentized residual, and point 96 which seems to have the highest leverage. As discussed above point 101 is the only one that seems likely to be truly problematic but we might want to keep an eye on the others.

(d) Taking out point 101 will make our model fit **better**. This means our errors will be smaller, leading to decreased RMSE and SSE, we will do a better job of explaining the variability in sleep times, so R-squared and F will go up, and correpondingly the p-value for the F test will go down (though it is already pretty low. If it is exactly 0 now we would say it would stay the same, but in fact if you go out to a large enough number of decimal places it would ne non-zero and would decrease.) The point has a low value of Y so removing it will make $\bar{Y}$ increase. Finally, the point is low and will pull the line towards itself, making the slope less steep than it otherwise would be. Removing the point will make b_1 increase. These outcomes are listed below along with the STATA printout that verifies the results.

```
(i)   R-squared      Increase
(ii)  RMSE           Decrease
(iii) b1             Increase
(iv)  Y-bar          Increase
(v)   F              Increase
(vi)  p-value for F  Decrease
(vii) SSE            Decrease

. regress sleeplength milk

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =   28.98
       Model |  1034501.63     1  1034501.63           Prob > F      =  0.0000
    Residual |  3497999.72    98  35693.8747           R-squared     =  0.2282
-------------+------------------------------           Adj R-squared =  0.2204
       Total |  4532501.35    99  45782.8419           Root MSE      =  188.93

------------------------------------------------------------------------------
 sleeplength |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        milk |   37.02467   6.877371     5.38   0.000     23.37675    50.67259
       _cons |   385.8047   37.77206    10.21   0.000     310.8473    460.7621
------------------------------------------------------------------------------
```

(e) We want a **prediction interval (PI)** rather than a confidence interval because we are talking about what my particular baby will do on this particular night, not about the average sleep time of babies who do not drink their milk. We could get this by hand if we wanted to as follows. The formula for a PI is

$$\hat{Y}_0 \pm t_{\alpha/2,n-2} * RMSE * \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SSX}}$$

We are given $X_0 = 0$ since the baby drinks no milk. Thus our predicted value is just $\hat{Y}_0 = b_0 = 385.8$.

From the printout, RMSE = 188.9. We can also get summary statistics for X which tell us that $\bar{X} = 4.75, n = 100, SSX = (n-1)s^2 = 99(2.761)^2 = 755$. Finally, we need $t_{\alpha/2,n-2} = t_{.975,98}$. Unfortunately our t-table has no row for 98 degrees of freedom. The value we want must be somewhere between $t_{.975,60} = 2$ and $t_{.975,120} = 1.98$. If we use the inverse ttail command in STATA we get that the exact value is 1.9845. Plugging in all these numbers gives a standard error of 192.7 and an interval of [3.5, 768.1]. Thus my baby is going to sleep somewhere between 3.5 and 768.1 minutes tonight. This is a huge range–it seems the sleeping habits of young babies are very variable!

If we want to take the shortcut and let STATA do all the work we just use the adjust command as follows and get the same answer up to rounding:

```
. adjust milk = 0, stdf ci

------------------------------------------------------------------------------
     Dependent variable: sleeplength     Command: regress
 Covariate set to value: milk = 0
------------------------------------------------------------------------------

------------------------------------------------------------
      All |         xb        stdf          lb          ub
----------+-------------------------------------------------
          |    385.805   (192.667)    [3.46314    768.146]
------------------------------------------------------------
     Key:  xb         =  Linear Prediction
           stdf       =  Standard Error (forecast)
           [lb , ub]  =  [95% Prediction Interval]
```

(f) 10 hours corresponds to 600 minutes of sleep. This value is in my PI so it IS possible I will get this much sleep. Remember the interval gives the range of possible values–since 600 is in the interval it is a possible value. In fact I might get as much as 769 minutes which is over 12 hours. However, I can not be 95% sure of getting 10 hours of sleep. My interval contains many values that are less than this–in fact I might get as little as 2 minutes of sleep–ouch! Note that my best estimate is 385 minutes or just over 6 hours, not even close to the amount that I want :(

(g) The STATA commands for creating the new variables X^2, $1/X$ and the shifted inverse, $1/(X + .5)$ that we suggested later, are shown below along with the corresponding regression models.

```
. gen milksq = milk*milk

. gen invmilk = 1/milk

.gen invmilkshift = 1/(milk+.5)

. regress sleeplength milk milksq

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) =   16.98
       Model |  1175171.96     2  587585.981           Prob > F      =  0.0000
    Residual |  3357329.39    97  34611.6432           R-squared     =  0.2593
```

```
-------------+------------------------------           Adj R-squared =  0.2440
       Total |  4532501.35    99  45782.8419           Root MSE      =  186.04

------------------------------------------------------------------------------
 sleeplength |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       milk |   85.42296   24.94405     3.42   0.001     35.91593      134.93
     milksq |  -5.174514   2.566726    -2.02   0.047    -10.26875   -.0802745
      _cons |   311.7171   52.28784     5.96   0.000     207.9403     415.494
------------------------------------------------------------------------------

. regress sleeplength invmilk

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =    4.17
       Model |  184847.018     1  184847.018           Prob > F      =  0.0439
    Residual |  4347654.33    98  44363.8197           R-squared     =  0.0408
-------------+------------------------------           Adj R-squared =  0.0310
       Total |  4532501.35    99  45782.8419           Root MSE      =  210.63

------------------------------------------------------------------------------
 sleeplength |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     invmilk |  -2.138985   1.047891    -2.04   0.044     -4.21849     -.05948
      _cons |   569.2835   21.37207    26.64   0.000     526.8713    611.6957
------------------------------------------------------------------------------

 regress sleeplength invmilkshift

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =   18.55
       Model |  721376.988     1  721376.988           Prob > F      =  0.0000
    Residual |  3811124.36    98  38889.0241           R-squared     =  0.1592
-------------+------------------------------           Adj R-squared =  0.1506
       Total |  4532501.35    99  45782.8419           Root MSE      =   197.2

------------------------------------------------------------------------------
 sleeplength |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
invmilkshift |  -201.1454   46.70275    -4.31   0.000    -293.8255   -108.4653
       _cons |   632.0177   25.57396    24.71   0.000      581.267    682.7684
------------------------------------------------------------------------------
```

(h) In a curvilinear model with powers of X you can't talk about the terms independently. It is not possible to hold X fixed and change X^2 and vice versa. Rather you have to interpret the coefficients as a group in terms of what they say about the shape of the relationship between Y and X. A quadratic model (X, X^2) gives the shape of a parabola. The fact that the coefficient of X^2 is negative means that the parabola opens down. Thus for a while as the amount of milk (X) the baby drinks increases the length of time they sleep increases. However eventually if the baby drinks too much the length of time they sleep will start to go down again. This makes real-world sense. A hungry baby (X too small) won't sleep, but a baby that has drunk too much will get a tummy ache or wet its diaper and also won't sleep! You can in fact use the numeric

values of the coefficients to figure out the optimal amount of milk to give the baby to maximize sleep and you can also look at the roots of the quadratic equation to see where it crosses Y=0 (i.e. the points at which the baby stops sleeping). Obviously a quadratic model can't be right indefinitely as it will become negative and a baby can't sleep a negative amount of time. Specifically to find teh peak of the parabola you differentiate $Y = b_0 + b_1X + b_2X^2$ with regard to X and set the result equal to 0. This gives $b_1 + 2b_2X = 0$ or $X = -b_1/2b_2$. From the printout we have $b_1 = 85.42$ and $b_2 = -5.17$ which gives an estimated optimal value of $X = -85.42/2(-5.17) = 8.26$ ounces of milk which is quite a plausible value.

For the inverse model we can use more or less our normal interpretations except that as the amount of milk drunk gets very large, $1/X$ or $1/(X + .5)$ approaches 0 and as the amount of milk drunk gets very small $1/X$ gets very large. Since the coefficient of the inverse term is negative in either case we see that as the amount of milk drunk gets large, $b_0 + b_1(1/X)$ or $b_0 + b_1(1/(X + .5))$ approaches b_0 from below. That is, the amount of time the baby sleeps increases as the amount of milk drunk increases but it levels off with a maximum value of $b_0 = 569$ minutes, a little under 10 hours with the simple inverse, or $b_0 = 632$, around 10 and a half hours based on the shifted inverse model. The value of b_1 tells us how quickly the function levels off. Thus function makes sense if we think there is an upper limit to how long a baby would sleep and that drinking milk would only help up to a certain point. This does not, however, allow for the possibility that too much milk is a problem and also as the amount of milk drunk gets very small this function starts to predict negative sleep times which is not reasonable.

Note that as suggested in my e-mail to the class, the simple inverse model does not fit very well. In fact the R^2 value is only around 3% which is horrible compared to the roughly 25% we get from the quadratic model or even the 22% from the simple linear model. The problem is that for values of milk drunk near 0, the function $1/X$ "blows up" or gets very large. The result is that the model has to curve down drastically as the amount of milk shrinks and will even become very negative which is not realistic. Adding the shift prevents this "blow up" since the smallest the inverse variable can get is then $1/(0 + .5) = 2$. This leads to a more stable and realistic model and we see that the R^2 value is up to about 15%–still not great but not as hideously bad. If you fool around with the exact shift you can make the fit even better although it will never be as good as the quadratic model which was what I actually used when generating the data.

(i) The diagnostic plots (residual plot, histogram, qq-plot) for the two models are shown in the accompanying graphics file. For the quadratic model the residual plot looks perfectly like random scatter centered about the 0 line. The histogram looks roughly bell-shaped though looking at the qq-plot we see that it is maybe a little heavy in the tails. Thus all the assumptions except possibly normality look fine and even normality isn't bad. For the inverse model the residual plot does not looks mean 0. Rather there is an overall upward slant with a bit of a curve at the end suggesting we have gotten the wrong shaped model and the constant variance assumption also looks a little off. The histogram and the qq plot don't look too bad. inf act if anything they may be very slightly better than for the quadratic model, but this is definitely outweighted by the problems with the residual plot.

(j) Overall I prefer the quadratic model. It has the much higher R^2_{adj}, lower RMSE and according to the residual plots it provides the right shape (no mean 0/independence violations) while the inverse model doesn't seem to fit quite right. Note that although I didn't ask you to do so we could formally test whetehr the quadratic model is superior to the simple linear model we fit originally. Our hypotheses would be:

$H_0 : \beta_2 = 0$–It is not worth adding the Milk squared term to the model when we already are using the Milk variable. The curvilinear model is not an improvement over a simple linear model.
$H_A : \beta_2 \neq 0$–It is worth adding the Milk squared term to a model that already has the Milk variable. A curvilinear model is superior to a simple linear model for explaining sleep time.

The p-value for the t-test of the Milk squared term is .047, less than our significance level of $\alpha = .05$ so we

reject the null hypothesis and conclude that the curvilinear model fits the data significantly better than a straight line. This is not a surprise given the results of our residual plots!

(k) You could check for multicollinearity here either by calculating the correlation between milk and milk squared or by looking at the variance inflation factors. Since there are only two predictors in the quadratic model these two things are essentially equivalent. Both calculations are shown below. We see that there is a huge correlation of r = .9624 and teh variance inflation factors are 13.57, way over our book's rule of thumb of 10 for an extremely bad variance inflation. This we definitely have severe multicollinearity. The milk squared term in our model seems relatively stable. However the linear term is a huge coefficient and standard error, possibly representing instability caused by the multicollinearity.

```
. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
        milk |     13.57    0.073712
      milksq |     13.57    0.073712
-------------+----------------------
    Mean VIF |     13.57

. corr milk milksq
(obs=100)

             |     milk   milksq
-------------+------------------
        milk |   1.0000
      milksq |   0.9624   1.0000
```

To fix the multicollinearity one option is to center the milk variable. That is we subtract off the mean of the milk variable to create a new predictor. Then we square that predictor to obtain a new quadratic term. The commands for creating these new variables (summarize to get the mean value of milk in the data set, gen to create the new variables, correlation or estat to check multicollinearity of the centered variables) are given below. We see that the centered variables have a correlation very close to 0 and that after fitting the regression with the centered variables the variance inflation factors are 1 (the minimum possible–no inflation) out to two decimal places. Thus centering has definitely removed the multicollinearity. As a result we see that the coefficient of the centered milk variable also has a much lower standard error (about 6.8) than it did in the uncentered model (24.9).

```
. summarize milk

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        milk |       100    4.755834    2.760939   .0052087   9.925487

. gen milkcen = milk - 4.755834
(23 missing values generated)

. gen milkcensq = milkcen*milkcen
(23 missing values generated)
```

```
. corr milkcen milkcensq
(obs=100)

             |  milkcen milkce~q
-------------+------------------
     milkcen |   1.0000
   milkcensq |  -0.0600   1.0000


. regress sleeplength milkcen milkcensq

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) =   16.98
       Model |  1175171.97     2  587585.986           Prob > F      =  0.0000
    Residual |  3357329.38    97  34611.6431           R-squared     =  0.2593
-------------+------------------------------           Adj R-squared =  0.2440
       Total |  4532501.35    99  45782.8419           Root MSE      =  186.04

------------------------------------------------------------------------------
 sleeplength |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     milkcen |    36.2047   6.784511     5.34   0.000     22.73932    49.67007
   milkcensq |  -5.174514   2.566726    -2.02   0.047    -10.26875   -.0802744
       _cons |   600.9376   26.85723    22.38   0.000     547.6335    654.2418
------------------------------------------------------------------------------

. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
     milkcen |      1.00    0.996406
   milkcensq |      1.00    0.996406
-------------+----------------------
    Mean VIF |      1.00
```

We can expand the quadratic equation for the centered variables out to see that both models will give us the same predictions. If we let $M = X - 4.756$ where M is the centered milk variable and X the original milk variable we see that the second regression equation can be written as

$$\begin{aligned}
\hat{Y} &= 600.94 + 36.20M - 5.17M^2 \\
&= 600.94 + 36.20(X - 4.756) - 5.17(X^2 - 2(4.756)X + 4.756^2) \\
&= 600.94 + 36.20X - 172.17 - 5.17X^2 + 49.12X - 22.62 \\
&= 406.15 + 85.32 - 5.17X^2
\end{aligned}$$

Up to rounding this is the same as our original equation so it will give the same predictions and of course all the values like R^2, RMSE, F, etc. are the same. What are different are the intercept and the coefficient of the milk term which are more stable in the centered model. note that the quadratic term has not changed–the highest power of X will always have the same coefficient in the centered and uncentered model because it only appears once when you expand the equation The drawback of the centered model is that it is a little

harder to interpret. The milk variable is expressed in terms of how much more or less the baby drank than 4.76 ounces which is not a natural reference point. This often happens when you center. What is more, when the data set changes the optimal centering point changes which is also a disadvantage. In this particular situation I don't think it makes a big difference whether we cventer or not as we are not very focused on trying to precisely estimate β_1.

(l) We are now asked to fit a multiple regression with all the predictor variables using mildly sick as the reference category for the health variables. This means we will use the indicators for healthy and severely sick and leave out the mildly sick indicator. I will use the uncentered milk variables for ease of interpretability but it won't really change anything in the rest of the problem. To tell whether the model is useful overall we need to perform an F test. Our hypotheses are

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$–None of the prdictors (age, milk, bedtime, level of illness) helps explain the variability in sleeptime. Overall our model is not useful.
H_A : At least one $\beta_i \neq 0$–i.e. at least one of the predictors does help explain how long a baby sleeps and so the model as a whole is useful.

From the printout the F statistic is a whopping 121.4 and the p-value is essentially 0 so we conclude the model is (very) useful for explaining variability in how long infants sleep. At least one of age, milk consumption, bedtime and health status is related to how long the babies sleep.

```
. regress sleeplength milk milksq age bedtime healthy sicksevere

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  6,    93) =  121.40
       Model |  4019317.99     6  669886.332           Prob > F      =  0.0000
    Residual |  513183.359    93  5518.10063           R-squared     =  0.8868
-------------+------------------------------           Adj R-squared =  0.8795
       Total |  4532501.35    99  45782.8419           Root MSE      =  74.284

------------------------------------------------------------------------------
 sleeplength |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        milk |   91.61008   9.973989     9.18   0.000     71.80372    111.4164
      milksq |  -5.435133   1.026692    -5.29   0.000     -7.47394   -3.396326
         age |   1.921524   .2590049     7.42   0.000     1.407192    2.435857
     bedtime |  -26.26565   12.68286    -2.07   0.041     -51.4513   -1.079999
     healthy |   264.5049   16.06949    16.46   0.000      232.594    296.4157
  sicksevere |  -154.4349   23.07161    -6.69   0.000    -200.2505   -108.6192
       _cons |    279.736   121.5916     2.30   0.024     38.27924    521.1928
------------------------------------------------------------------------------
```

(m) This is one of those nasty interpretive problems based on a CI for a regression slope and caused a lot of confusion! From the printout, the confidence interval for β_4, the ebdtime variable, is [-51.45, -1.08]. This means that on average, putting the baby to bed an hour later decreased sleep time by between 1 and 51 minutes. This decrease is less than the hour later the baby was put to bed so we can be 95% sure that the baby that goes to bed later will wake up later too–by between 9 and 59 minutes. If you find that confusing try it with some specific numbers. Suppose the baby who goes to bed at 8:00 is predicted to sleep for 12 hours so they wake up at 8am. The equivalent baby who went to sleep at 9:00 will according to our CI sleep 1-51 minutes less or for between 11:09-11:59 minutes. This yields a wake-up time of 8:09-8:59. You can not say that the 9:00 baby will wake earlier because it sleeps less–after all it stayed up later too!

(n) Multicollinearity refers to the presence of linear relationships among two or more of the predictor variables. We can check for pairwise relationships using correlations and we can check for higher order relationships using variance inflation factors. Below I show the correlations and VIFs for the model with all 6 predictors (milk, milksquared, age, bedtime and the indicators for current health.) We know from part k that there is a huge correlation between milk and milk-squared. The real purpose here was to determine whether there were any other correlations that were worrisome since we know the problem with milk and milk-squared can be fixed by centering. Looking at the correlation table the only high correlations we see are between some of the predictors and sleeplength (these are good since we want the X's to be correlated with Y or they won't be good predictors!), between milk and milk-squared (which we already knew about) and between some of the indicator variables for health. These last are actually not that worrisome because they don't really represent separate variables. The baby has to have exactly one of the three health status levels–healthy, mildly sick, or severely sick–this indeed is why we don't include all three indicators–it would give us perfect multicollinearity! However even if we only include 2 we will get some collinearity. This is rarely enough to cause a problem but if it does we can center the indicators as well. Other than that there are no bad looking correlations and the VIFs for everything except milk and milk squared are very close to 1 so it doesn't seem as if adding the new variables has created any additional multicollinearity problems. Moreover all the new variables are significant and have the signs and magnitudes we would expect.

```
. corr sleeplength milk milksq age bedtime healthy sickmild sicksevere
(obs=100)

             | sleepl~h     milk   milksq      age  bedtime  healthy sickmild
-------------+---------------------------------------------------------------
 sleeplength |   1.0000
        milk |   0.4777   1.0000
      milksq |   0.4120   0.9624   1.0000
         age |   0.3017   0.0350   0.0326   1.0000
     bedtime |  -0.0924  -0.1403  -0.1486  -0.0736   1.0000
     healthy |   0.6610  -0.0752  -0.0726  -0.0021   0.0774   1.0000
    sickmild |  -0.3581   0.0159   0.0204   0.0519  -0.0302  -0.7543   1.0000
  sicksevere |  -0.4362   0.0851   0.0748  -0.0709  -0.0678  -0.3576  -0.3433

             | sickse~e
-------------+---------
  sicksevere |   1.0000

. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
      milksq |     13.61    0.073449
        milk |     13.60    0.073503
  sicksevere |      1.16    0.861008
     healthy |      1.15    0.867250
     bedtime |      1.03    0.966816
         age |      1.01    0.987184
-------------+----------------------
    Mean VIF |      5.26
```

Overfitting means including variables in your model that are not useful. There is no evidence of overfitting in this model since all our predictors have p-values less than $\alpha = .05$–they are all worth keeping in the

model. People often try to answer this part by comparing R^2 and R^2_{adj}. While it is true that the closer they are the less you suspect overfitting, this isn't a perfect answer since it is hard to tell how big a difference is significant. The p-values are a more reliable guide. Additionally, some people tried to tell me there was a problem because R^2_{adj} was lower than R^2. This is not correct. In EVERY model R^2_{adj} is lower than R^2–just look at the respective formulas. The important question in terms of overfitting is whether when you ADD a variable to a model the R^2_{adj} for the NEW model is LOWER than the R^2_{adj} for the OLD model. To check this way you would have to fit multiple models adding variables one at a time to see if R^2_{adj} went up or down. You can do this but it's a lot of extra work when you can tell that the existing variables are all significant.

(o) We suspect we already know the answer to this question since the indicator variables for healthy and severely sick both had p-values $< .05$, implying that infants in these conditions had significantly different sleep times from mildy sick infants of the same age, milk consumption and bedtime. However, to check this formally we need to do a partial F test to see whether adding the two indicators **as a group** improved our model. To make the comparison we need to fit the model without these indicators and compare it to the model from part (l). The "reduced model" without the health indicators is shown below:

```
. regress sleeplength milk milksq age bedtime

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  4,    95) =   12.26
       Model |  1542768.96     4   385692.24           Prob > F      =  0.0000
    Residual |  2989732.39    95  31470.8672           R-squared     =  0.3404
-------------+------------------------------           Adj R-squared =  0.3126
       Total |  4532501.35    99  45782.8419           Root MSE      =   177.4

------------------------------------------------------------------------------
 sleeplength |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        milk |   84.37943   23.78903     3.55   0.001     37.15223    131.6066
      milksq |  -5.162641   2.450714    -2.11   0.038    -10.02792   -.2973577
         age |   2.091232   .6164454     3.39   0.001     .8674333    3.315031
     bedtime |  -5.355956   30.19127    -0.18   0.860     -65.2932    54.58128
       _cons |   196.7187   289.1654     0.68   0.498    -377.3472    770.7846
------------------------------------------------------------------------------
```

We want to test whether the full model from part (l) is superior to the reduced model fit above which is equivalent to testing whether the coefficients of the extra variables are all 0 or not. Our hypotheses are

$H_0 : \beta_5 = \beta_6 = 0$–the health indicators do not explain any additional variability in sleep time beyond what is explained by milk consumption, age and bedtime. This group of variables is not worth adding to the model.
$H_A : \beta_5 \neq 0, \beta_6 \neq 0$ or both–the health indicators as a group do explain some of the variability in sleep time. The full model using all 6 variables is superior to the reduced model without the health indicators.

Since we are adding two extra variables to the model, the test statistic that goes with these hypotheses is

$$F = \frac{(SSE_{red} - SSE_{full})/2}{SSE_{full}/(n-6-1)} = \frac{(SSR_{full} - SSR_{red})/2}{SSE_{full}/(n-6-1)} = \frac{(R^2_{full} - R^2_{red})/2}{(1-R^2_{full})/(n-6-1)} = \frac{(.8868 - .3404)}{(1-.8868)/93} = 224.4$$

You can think of this statistic as representing a standardized reduction in error or increase in explanatory power from adding the extra variables. It is sometimes called the R^2 difference test because of the third

way it can be written. Under the null hypothesis it has an F distribution whose degrees of freedom for the numerator are the difference in the number of predictors between the two models and whose degrees of freedom in the denominator are the error degrees of freedom for the full model. Here we have 2 and 93 degrees of freedom. The associated p-value, obtained using the Ftail command in STATA, is $P(F_{2,93} \geq 224.4) = 0$ to 35 decimal places! We therefore reject the null hypothesis and conclude that health status does explain a significant amount of the variability in sleep times. We can replicate this test directly in STATA as follows:

```
. test healthy = sicksevere = 0

 ( 1)  healthy - sicksevere = 0
 ( 2)  healthy = 0

       F(  2,    93) =  224.40
            Prob > F =    0.0000
```

(p) The purpose of this part was to demonstrate how some programs directly accomodate a multilevel categorical variable, treating it as a single X rather than several separate variables. Of course what SAS actually does is to create its own dummy variables which you can tell by noting how many degrees of freedom it assigns to the categorical variable. The printout is shown below. You should get exactly the same answer as from the partial F test shown above and you do up to several decimal places. If you look hard to enough decimal places you will see a slight difference. There was apparently a typo for one of the subjects in the categorical health status variable but it hasn't changed the basic result.

```
proc glm data = tmp1.hw6;
class healthstatus;
model sleeplength = milk milk*milk age bedtime healthstatus;
run;
```

```
                                        The SAS System    12:47 Saturday, November 27, 2010   2

                                     The GLM Procedure

Dependent Variable: SleepLength   SleepLength

                                           Sum of
      Source                       DF         Squares     Mean Square    F Value    Pr > F

      Model                         6     4019317.938      669886.323     121.40    <.0001

      Error                        93      513183.329        5518.100

      Corrected Total              99     4532501.267


                  R-Square     Coeff Var      Root MSE    SleepLength Mean

                  0.886777      13.22042      74.28392           561.8879


      Source                       DF       Type I SS     Mean Square    F Value    Pr > F

      Milk                          1     1034501.589     1034501.589     187.47    <.0001
```

Milk*Milk	1	140670.372	140670.372	25.49	<.0001
Age	1	366606.589	366606.589	66.44	<.0001
Bedtime	1	990.420	990.420	0.18	0.6728
HealthStatus	2	2476548.967	1238274.483	224.40	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Milk	1	465520.103	465520.103	84.36	<.0001
Milk*Milk	1	154642.716	154642.716	28.02	<.0001
Age	1	303714.355	303714.355	55.04	<.0001
Bedtime	1	23666.364	23666.364	4.29	0.0411
HealthStatus	2	2476548.967	1238274.483	224.40	<.0001

(q) The STATA and SAS printouts for the interaction model are shown below. Note that just as we need a reference category for health status we also need a reference category for the healthstatus by age interaction. It makes most sense to create interaction terms for healthy and severely sick with age since those are the categories being used for health status as a main effect.

```
. gen healthyage = healthy*age
(23 missing values generated)

. gen sickseverage = sicksevere*age
(23 missing values generated)

. regress sleeplength milk milksq age bedtime healthy sicksevere healthyage sick
> severage

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  8,    91) =  113.98
       Model |  4121223.73     8  515152.967           Prob > F      =  0.0000
    Residual |  411277.615    91  4519.53423           R-squared     =  0.9093
-------------+------------------------------           Adj R-squared =  0.9013
       Total |  4532501.35    99  45782.8419           Root MSE      =  67.227

------------------------------------------------------------------------------
 sleeplength |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        milk |   93.93993   9.068173    10.36   0.000     75.92712    111.9527
      milksq |  -5.715107   .9329975    -6.13   0.000    -7.568392   -3.861822
         age |   1.076291   .3357409     3.21   0.002      .409383    1.743199
     bedtime |  -34.12419   11.86222    -2.88   0.005    -57.68703   -10.56136
     healthy |   94.86435   42.41308     2.24   0.028     10.61597    179.1127
  sicksevere |  -95.64924   72.65317    -1.32   0.191    -239.9658    48.66735
  healthyage |   2.095442   .4903281     4.27   0.000     1.121466    3.069419
sickseverage |  -.8644753   .9126223    -0.95   0.346    -2.677287    .9483369
       _cons |   418.0405   114.3557     3.66   0.000      190.887    645.1941
------------------------------------------------------------------------------

Note: The | indicates to SAS that you want an interaction.  I also created
the milk squared term as part of the model statement just by typing
```

```
milk*milk.

proc glm data = tmp1.hw6;
class healthstatus;
model sleeplength = milk milk*milk age bedtime healthstatus|age;
run;
```

The SAS System 12:56 Saturday, November 27, 2010 2

The GLM Procedure

Dependent Variable: SleepLength SleepLength

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	4121223.680	515152.960	113.98	<.0001
Error	91	411277.586	4519.534		
Corrected Total	99	4532501.267			

R-Square	Coeff Var	Root MSE	SleepLength Mean
0.909260	11.96457	67.22748	561.8879

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Milk	1	1034501.589	1034501.589	228.90	<.0001
Milk*Milk	1	140670.372	140670.372	31.12	<.0001
Age	1	366606.589	366606.589	81.12	<.0001
Bedtime	1	990.420	990.420	0.22	0.6408
HealthStatus	2	2476548.967	1238274.483	273.98	<.0001
Age*HealthStatus	2	101905.743	50952.871	11.27	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Milk	1	485014.4117	485014.4117	107.32	<.0001
Milk*Milk	1	169582.6801	169582.6801	37.52	<.0001
Age	1	93772.3850	93772.3850	20.75	<.0001
Bedtime	1	37401.3138	37401.3138	8.28	0.0050
HealthStatus	2	41286.5064	20643.2532	4.57	0.0129
Age*HealthStatus	2	101905.7427	50952.8714	11.27	<.0001

(r) The model in part (q) is definitely better than the model in part (l). We can see this in several different ways. First it has a larger R^2_{adj}. For the model without the itneraction we have $R^2_{adj} = .8795 = 87.95\%$ while for the model with the interaction we have $R^2_{adj} = .9013 = 90.13\%$. Similarly we RMSE = 74.284 for the model in part (l) and RMSE = 67.227 for the model in part (q) indicating that our prediction errors

have improved substantially. However it is hard to tell whether either of these differences is "significant"– i.e. couldn't just be due to chance. To formally check this we use another partial F test to see if the two interaction variables have added something to the model or equivalently whether the age*healthstatus interaction term in the SAS version of the model is significant. Looking at the SAS printout we see that the F test for the interaction term has F = 11.27 with 2 and 91 degrees of freedom and a p-value of ¡ .0001. Therefore the interaction term was definitely worth adding to the model. We can see the same thing from the STATA partial F test shown below.

```
. test healthyage = sickseverage = 0

 ( 1)  healthyage - sickseverage = 0
 ( 2)  healthyage = 0

       F(  2,    91) =   11.27
            Prob > F =    0.0000
```

(s) The results of parts (q) and (r) tell us there is a significant interaction between healthstatus and age. We know that the model has significantly improved by adding the interaction terms. Thus means that the relationship between age and sleeping time depends on the baby's health or equivalently that the differences in sleep time between babies in different health conditions depends on how old the babies are. To tell what this effect looks like it is actually easiest to look at the interaction terms we created in STATA. The variable healthyage has a positive coefficient meaning that for healthy babies average sleep time goes up faster with age than for mildly sick babies while for sicksevereage the coefficient is negative meaning that average sleep time goes up less quickly with age than for mildly sick babies. The coefficient for healthy age is much bigger in absolute value than that for sicksevereage and is statistically significant while the coefficient of the sicksevereage variable is NOT statistically different from 0. This suggests that there is a difference in the sleep-age relationship between healthy and sick babies but that the level of severity of the sickness may not have a significant effect. In other words, what is driving the interaction is whether or not the baby is sick, not how sick they are. (Note also that my description is not meant to imply that I have longitudinal data where I follow a healthy or sick baby and see how its sleep time changes. Rather I have cross sectional data and am saying that the average difference between, say, size month old and seven month old babies depends on how sick they were when the measurement was taken.)
(t) Part (t) is an extension of part (s), looking more closely at exactly what the differenecs in the sleeptime-age relationship are as a function of health. Basically to get the equations we have to plug in the given values for milk and bedtime and the different possible health combinations to get a relationship between sleep time and age. We are given Bedtime = 8, Milk = 5 and therefore MilkSquared = 25. The equations are as follows. For a healthy baby, healthy = 1 and sicksevere = 0 so we get

$$\begin{aligned}\hat{Y} &= 418.04 + 1.08Age + 93.93(5) - 5.72(25) - 34.12(8) + 94.86(1) - 95.65(0) + 2.10(1)Age - .864(0)Age \\ &= 566.59 + 3.18Age\end{aligned}$$

For a mildly sick baby we have healthy = sicksevere = 0:

$$\begin{aligned}\hat{Y} &= 418.04 + 1.08Age + 93.93(5) - 5.72(25) - 34.12(8) + 94.86(0) - 95.65(0) + 2.10(0)(Age) - .864(0)Age \\ &= 471.73 + 1.08Age\end{aligned}$$

For a very sick baby we have healthy = 0 and sicksevere = 1:

$$\begin{aligned}\hat{Y} &= 418.04 + 1.08Age + 93.93(5) - 5.72(25) - 34.12(8) + 94.86(0) - 95.65(1) + 2.10(0)Age - .864(1)(Age) \\ &= 376.08 + .21Age\end{aligned}$$

The plot is shown in the accompanying graphics file. We see that the lines all appear to have different slopes and intercepts. This tells us that the effect of illness on sleep time depends on how old the baby is and that the rate of change in sleep time with age depends on how sick the baby is. Babies who are severely ill sleep nearly the same average amount no matter how sick they are (the slope is almost flat) and babies who are mildly sick sleep roughly an extra minute per extra day of age. From our hypothesis tests above we know that this difference may not be significant. Babies who are healthy sleep almost 3 minutes extra per day of age on average which is a significant effect. The base sleep times (which corresponds to newborns) are significantly different for all three health categories (per part q) with hjealthy newborns slwwping the longest followed by mildly sick babies and then severely sick babies. This all makes perfect sense. Note also that the interaction does not just say there is a difference in sleep times between sick and not sick babies. The key is that the combination of age and sickness is important.

Note: You did not have to turn in parts (u) and (v) because we had not completely covered them in class by the time the homework was due. However I include the solutions below as they are relevant for the final exam.

(u) In backwards stepwise regression you fit a model that includes all the possible predictor variables and if any of them are not significant you remove them one at a time, taking out first the one that contributes least to the model, and then refitting at each step to see how the significance of the other variables change. The reason Nora thinks we wold remove the sicksevere-age interaction term first in a backwards stepwise procedure is that it is the only variable in the model that has a p-value greater than $\alpha = .05$. However we have to remember the hierarchical principle. The sicksevere-age variable is NOT really an independent variable. It is part of the healthstatus-age interaction that has to be represented by multiple model terms but is really one concept. As we saw earlier this interaction is significant and we shouldn't remove just "part" of it unless we want to redfine our health status variable as just healthy versus not healthy. If we did that then the sicksevereage term would not appear in the model.

(v) In forward stepwise selection at each stage you add the variable that will most improve the current model. You start with no variables entered into the model so at the first step you add the best individual predictor. You can figure out the best individual predictor by getting correlations and seeing which variable has the strongest correlation with the outcome. The correlation table for our variables is shown below. We see that in terms of the relationship with sleeplength the best variable is actually the indicator for being healthy, followed by the linear term in milk. The only tricky part of this is how you handle the hierarchical principle. In fact "healthy" should not be added on its own. It is part of the multicategory variable, healthstatus. If you wanted to you could fit a model a la our SAS procedure above with healthstatus as the only predictor and look at its R^2 value and compare that to the other individual variables. Here we know healthstatus would win since even just using the single indicator it came out the best. However in general stepwise procedures deal badly with variables that are represented using multiple predictors. They also deal rather badly with multicollinearity. In backwards stepwise you get unstable estimates for the coefficients of the correlated variables which may lead to removal of the wrong variables and in forward stepwise which of a multicollinear set gets added first is very dependent on the sample, again leading to unstable selection procedures.

```
. corr sleeplength milk milksq age bedtime healthy sickmild sicksevere
(obs=100)

             | sleepl~h     milk   milksq      age  bedtime  healthy sickmild
-------------+---------------------------------------------------------------
 sleeplength |   1.0000
       milk |   0.4777   1.0000
     milksq |   0.4120   0.9624   1.0000
        age |   0.3017   0.0350   0.0326   1.0000
    bedtime |  -0.0924  -0.1403  -0.1486  -0.0736   1.0000
```

```
    healthy |   0.6610  -0.0752  -0.0726  -0.0021   0.0774   1.0000
  sickmild |  -0.3581   0.0159   0.0204   0.0519  -0.0302  -0.7543   1.0000
sicksevere |  -0.4362   0.0851   0.0748  -0.0709  -0.0678  -0.3576  -0.3433
```