Homework Assignment 6

Due Date: Wednesday, November 23rd

Note: This is our final homework assignment of the quarter! There are 5 problems. The first 4 provide examples and extra practice. They are relevant for the final but do not need to be turned in. Solutions for them are available on the class web site. You must turn in Problem 5 together with the corresponding STATA/SAS printouts to receive full credit. The assignment is due Wednesday, November 23rd. Officially it is due in class but you may turn it in to the Adam's mail folder (in CHS 51-254) any time before 3:30 with no penalty.

Note: Output from any calculations done in STATA or SAS MUST be included with your assignment for full credit. If I do not specify which way to do a problem you may chose whether to do it by hand or on the computer. All the STATA/SAS commands needed to complete this homework are given at the end of the assignment and will be reviewed in the lab. You do not need to hand in a separate lab report–simply turn in the relevant output as part of your homework.

Note: You are encouraged to work with fellow students on these problems. However, you MUST write up your solution ON YOUR OWN and IN YOUR OWN WORDS. The style of your write-up is as important as getting the correct answer. Your solutions should be easy to follow, and contain English explanations of what you are doing and why. You do not have to write an essay for each problem, but you should give enough comments so that someone who has not seen the problem statement can understand your work. You do not have to type your assignments. However, if they are too sloppy to read, too hard to understand, or give just numbers with no comments, you WILL lose points.

Warm-up Problems

(1) Regression Assumption Basics

(a) List the assumptions that we make when we fit a simple linear regression model. In each case say why we make this assumption, and see if you can think of an example for which that assumption might be invalid.

(b) For each of the plots in the accompanying graphics file, HW6P1 graphs, say which of the four regression assumptions about the errors you can check and explain which ones are violated. Identify any outliers, influential or high leverage points visually, say how you could identify them computationally if you had the data, and what effect if any you would expect to see on the model if the points were removed.

(2) Curvilinear Regression Basics (WBCH 13.56): An agronomist wants to develop a regression model to predict the weekly growth of alfalfa, Y (in inches), from the amount of fertilizer, X (in units of 100 pounds per acre). Data were collected for 15 plots of land where the amount of fertilizer was varied and are presented in the table below as well as in the online data sets.

Growth	Fertilizer	Growth	Fertilizer
2.5	0.0	8.0	3.4
4.2	0.4	8.0	3.9
5.1	0.9	7.7	4.3
4.6	1.3	7.7	4.7
5.0	1.7	7.8	5.1
5.8	2.1	8.2	5.6
7.1	2.6	8.5	6.0
7.3	3.0		

(a) Obtain a scatterplot of the data in STATA or SAS. Based on this plot, do you think a straight line model or a curvilinear model makes more sense? Does this seem reasonable in real-world terms? Explain.

(b) Find the estimated curvilinear regression equation $\hat{Y} = b_0 + b_1 X + b_2 X^2$ in STATA or SAS, and use it to find the predicted height of the alfalfa plants when 500 pounds of fertilizer are used per acre. Do the prediction first by hand to make sure you know how and then verify your answer in STATA or SAS.

(c) Find the estimated simple linear regression equation $\hat{Y} = b_0 + b_1 X$ in STATA or SAS, and use it to find the predicted height of the alfalfa plants when 500 pounds of fertilizer are used per acre.

(d) Based on the scatterplot in (a), which of your predictions seems more sensible? Explain briefly.

(e) Which of the models seems to be fitting the data better? Justify you answer (i) by comparing R^2_{adj} and $RMSE$ for the two models and (ii) by performing an hypothesis test to determine whether or not $\beta_2 = 0$ in the curvilinear regression model. Explain briefly why each of these comparisons tells you which model is better.

(f) Try fitting two more curvilinear models to this data, one transforming X and one transforming Y. In each case explain how you chose the transformation you did and whether you think it is better than the quadratic model from part (b).

(3) Pregnancy Weight Gain: An obstetrical researcher is interested in understanding the relationship between a woman's age and how much weight she gains during pregnancy. For each of n=100 of her patients she records X, their age (in years), and their total weight gain, Y, over the course of the pregnancy (in pounds). The variables can be found in the HW6 data files on the web site. Note that I have labeled the X variable as "momage" and the Y variable as "weightgain".

(a) Obtain a scatterplot of Y versus X for these data. Does it seem as if the relationship between weight gain and age is linear? Explain briefly. Visually do you see any outliers/influential points? Calculate a set of appropriate statistics (e.g. Cook's distance, DFBetas, etc.) and/or plots to confirm what you see.

(b) Fit a simple linear regression of Y on X. Is there a statistically significant relationship between weight-gain and age? Briefly justify your answer. Your answer may seem to contradict what you said in part (a). Explain how both answers can be right.

(c) For the model from part (b) obtain a residual plot (residuals versus age), a histogram of the residuals, and a normal probability plot of the residuals. Use these plots to explain whether or not each of the four main regression assumptions is violated, briefly justifying your answers. (Note that the scatterplot from part (a) may also provide some guidance.)

(d) Use STATA or SAS to generate a new variable, momage squared (X^2) and fit a polynomial regression of weight gain on momage and momage squared. Is adding the momage squared variable worthwhile? Justify

your answer by (i) comparing the R^2_{adj} and RMSE of your new model to the one from part (b) and (ii) by performing an appropriate hypothesis test. You should carefully write out the hypotheses for your test both mathematically and in words, say whether or not to reject and why, and explain your conclusions.

(e) Obtain the residuals for the model in part (d) and use them to check the four main regression assumptions by generating appropriate plots. In this sense does the model from part (d) represent an improvement over the model in (b)?

(f) It is possible to fit different shapes of models by transforming the response variable, Y, rather than the predictor variable, X. Use STATA to create a the inverse of the weight gain variable, $1/Y$, and fit a simple linear regression of this new response on the age variable. Compare this model to those from parts (b) and (d) (i) by using R^2_{adj} and (ii) by obtaining the corresponding plots for checking the error assumptions and evaluating them. Overall which model (b), (d), or (f) would you chose to use?

(g) Note that in part (f) I did not recommend using RMSE to compare the new model with those from parts (b) and (d). Explain why RMSE would not be a fair choice for comparison in this situation.

(4) Don't Drink and Derive: As we all know, drinking and driving is dangerous. Drinking and doing statistics may be even more so! A public health professor at the University of Calculationally Learned Adults has been asked to provide information about factors that affect peoples' blood-alcohol levels after drinking, in support of her city's drunk-driving standards. During this problem you will help her analyze the data from a study she has conducted. You will find the necessary variables in the accompanying data file.

Study Part I: Intuitively the professor thinks that two of the most important predictors of a person's blood alcohol level should be the amount they have drunk and how long it has been since they had the drinks. She performs a study with n=122 people and records X_1, the number of ounces of alcoholic beverages the person has consumed (which I will call Amount), X_2, the number of hours since the drinks were consumed (which I will call Time) and Y, the person's blood alcohol content (BAC), measured as a percentage. For reference you may find it useful to know that the legal limit for driving in California is BAC = .08%, just under a tenth of a percent. Note also that in STATA the variable names are all lower case.

(a) Fit two simple linear regressions, one of BAC on number of ounces drunk and one of BAC on time since consumption. Is there a significant relationship between BAC and each of the predictors? Briefly justify your answer.

(b) Which variable, Amount or Time seems to be a better predictor of BAC? Briefly justify your answer.

(c) Obtain a scatterplot, histogram, normal quantile plot and residual plots for the regression of BAC on Time. Explain which if any of our main regression assumptions are violated in this model and why. In addition explain why each of the violations you see in the residual plots make real-world sense in the context of the problem.

(d) Do there appear to be any outliers based on the plots in part (c)? Calculate appropriate summary statistics to detect outliers and leverage points and discuss briefly any points identified as unusual and what might have caused them. (Note: The plots in part (c) will show you the BAC and Time values for any unusual points. It may be helpful to also obtain the scatterplot of BAC versus amount drunk and identify the corresponding points on that plot to fiigure out what is going on.)

(e) Now fit a quadratic model of BAC on Time and Time^2. The quadratic variable is included in the data set as timesq but you should know how to generate it. Perform an appropriate hypothesis test to determine whether the curvilinear model is superior to the simple linear model.

(f) Is the intercept in this model significantly different from 0? Explain why your model makes real-world sense.

(g) What do you think would happen if you removed any outliers/influential points you found in part (d) from the models you have fit so far? In particular, for the SLR of BAC on Amount say what would happen to R^2, b_1 and t_{obs} and for the quadratic model of BAC on Time and Time^2 say what would happen to SST, SSE, RMSE, b_2, F and its p-value.

(h) Does there seem to be significant multicollinearity between Time and Time-squared in your quadtratic model? Check by performing an appropriate calculation. Then calculate the "centered" versions of the Time and Time-squared variables as we discussed in class and check whether this has reduced the collinearity. Refit the quadratic model with the new centered predictors and expand the equation out to see how it compares to the original curved model in part (e). How do the standard errors of the coefficients in this model compare to those in part (e)?

Study Part II: Of course, there are many factors besides amount drunk and time since alcohol consumption that may affect a person's BAC level. The professor has collected information on several other variables. For the remainder of the problem we will focus on a multiple regression model for which the outcome of interest, is Y, the BAC score, and the set of predictors includes X_1, the amount drunk (in ounces), X_2, the time since the drinks were consumed (in hours), $X_3 = X_2^2$, a quadratic term in time, X_4, the person's weight (in pounds), and some indicators including X_5, whether the alcohol was drunk with a meal (1 = Yes, 0 = No), and the type of alcohol consumed (beer, wine or hard liquor.)

(i) Fit the multiple regression model of BAC on all the predictors using wine as the reference category for type of alcohol. Overall is this model useful for explaining the variability in BAC? Explain briefly.

(j) Fit the regression of BAC on the first 5 predictor variables (that is leaving out the indicators for type of alcohol consumed.) Use the values from this printout and the one in part (i) to determine whether Type of Alcohol overall contributes improtant information to the model. State the null and alternative hypotheses mathematically and in words, compute the appropriate test statistic by hand and then use STATA or SAS to confirm your result and get the final p-value.

(k) Did you really need to perform the test in part (j) to determine whether Type of Alcohol consumed was useful? Describe a situation in which the test might actually have provided helpful information.

(l) Create the variables you would need to test for an interaction between Type of Alcohol consumed and Amount of alcohol consumed. Overall does there appear to be an interaction? You may perform the relevant test in STATA or SAS. Carefully explain your conclusions and what they tell you in the context of the problem.

(m) If you were using backwards stepwise model selection based on the model in part (i) what would happen at the first step?

(m) If you were using forwards stepwise model selection, what would happen at the first step? Explain how you could figure this out without using any regression models and carry out the corresponding calculation.

Problems To Turn In

(5) Drinking the Milk of Paradise: A professor of pediatrics at the University of Calculationally Learned Adults is interested in knowing what factors affect how well an infant sleeps at night.

Study Part I: A common story among parents of young babies is that if you can get them to drink a lot of milk before going to bed they will sleep longer. Our pediatric researcher decides to test this. She has collected information on Y, the number of minutes a baby sleeps (sleeplength) and X, the number of ounces of milk the baby has consumed (milk). The data are given in the accompanying files.

(a) Is there a significant linear relationship between sleeping time and amount of milk drunk? Justify your answer with an appropriate test using $\alpha = .05$. You do not need to write out all the details. Just give your basic reasoning.

(b) Obtain a scatterplot of sleep time versus amount of milk drunk. Obtain the residuals for the simple linear regression from part (a) and use them to create a histogram, qq-plot and residual plot. List the four assumptions we make about the errors in a regression model and use your assorted plots to explain whether each of these assumptions is reasonable and why. (You should indicate which plot(s) you are using to check each assumption.)

(c) Do there appear to be any outliers or leverage points in this data set? Calculate appropriate summary statistics (leverage values, studentized residuals, dfbetas, Cook's distances, dffits) to check and say how large an effect any points you identify are having on the model. Give a possible real-world explanation (other than a data entry error) for what has caused any points you identified to be unusual.

(d) For each of the quantities below indicate whether and why you think they would increase, decrease, stay the same, or it cannot be determined what would happen without actually refitting the model if one of the most major outerliers from part (c), point 101 (the baby who drunk 5.5 ounces of milk but slept less than an hour) was removed from the data set. Then exclude the point and verify your answers.

(i)	R-squared	Increase	Decrease	Same	Can't Tell
(ii)	RMSE	Increase	Decrease	Same	Can't Tell
(iii)	b1	Increase	Decrease	Same	Can't Tell
(iv)	Y-bar	Increase	Decrease	Same	Can't Tell
(v)	F	Increase	Decrease	Same	Can't Tell
(vi)	p-value for F	Increase	Decrease	Same	Can't Tell
(vii)	SSE	Increase	Decrease	Same	Can't Tell

Note: For all subsequent parts of the problem use the data set from part (d) with the major outlier, point 101, excluded.

(e) I note with horror that my baby has not drunk any of his milk tonight before going to sleep! Based on the simple linear regression model, find an appropriate 95% interval for how long he will sleep tonight. You should explain briefly why you chose the type of interval you did.

(f) I have a big presentation tomorrow and feel I need at least 10 hours sleep. Based on your answer to part (e) is it possible I can get this much sleep before my baby wakes up? Can I be 95% sure I will get this much sleep? Explain briefly in each case.

Study Part II: The pediatrics processor has decided to see if a curvilinear model would fit this data better. Specifically she decides to fit (i) a quadratic model with predictors X (milk) and X^2 (milk-squared) and (ii)

an inverse model of the form $Y = \beta_0 + \beta_1 \frac{1}{X}$ where the predictor is 1/milk.

(g) Create the transformed variable the professor needs and fit the two curvilinear models in either STATA or SAS.

(h) Provide the best interpretations you can for the various regression coefficients for the models in part (g). What does the negative sign on the milk-squared term in quadratic model tell you about the nature of the relationship between sleep time and milk consumption and how might this make real-world sense? Why in real-world terms, might the professor's inverse model make sense?

(i) Obtain the diagnostic plots for the two models from part (g) and explain what they tell you about whether the regression assumptions are valid in each case.

(j) Overall, which model do you prefer overall and why?

(k) Does there seem to be significant multicollinearity between Milk and Milk-squared in your quadtratic model? Check by performing an appropriate calculation. Then calculate the "centered" versions of the Milk and Milk-squared variables as we discussed in class and check whether this has reduced the collinearity. Refit the quadratic model with the new centered predictors and expand the equation out to see how it compares to the original model in part (g). How do the standard errors of the coefficients in this model compare to those in part (g)? Do you prefer the centered or uncentered model? Discuss briefly.

Study Part III: The pediatrics professor realizes that there are many factors other than milk drinking that affect how a baby sleeps. She has therefore collected information on several other variables and plans to fit a multiple regression model. Her variables are Y, the length of time slept in minutes, X_1, the baby's age (in days), X_2, the amount of milk drunk at bedtime in ounces, $X_3 = X_2^2$, a Milk Squared term, X_4, the time of night the baby goes to bed (8:00 pm would be recorded as 8, 8:30 as 8.5, and so on) and some indicators for whether or not the baby is sick. There are three categories: not sick, mildly sick, and very sick. I have included indicators for all three sickness categories as well as a single character variable for illness that can be used in SAS. Use this data set to answer the remaining questions.

(l) Fit the multiple regression of sleep time on the other variables, using "mildly sick" as the reference category. Overall is this model useful for explaining variation in how long an infant sleeps? Briefly justify your answer using $\alpha = .05$.

(m) On average, all other things equal, who would **wake up earlier in the morning**, a baby who went to sleep at 8:00 or a baby who went to sleep at 9:00? Can you be 95% (really 97,5%) sure of your answer? Explain.

(n) Is there any evidence of multicollinearity in this data? If so, does it appear to have caused any problems in the model? Is there evidence of overfitting? Explain briefly in each case, showing any relevant calculations needed to verify your answers.

(o) Overall does the health of the infant seem to affect how long it sleeps? Determine this by fitting the MLR without the health variables and using the results plus those from part (l) to perform an appropriate test. State the null and alternative hypotheses mathematically and in words, compute the appropriate test statistic by hand and then use STATA or SAS to confirm your result and get the final p-value.

(p) Repeat the test in part (o) by fitting a glm in SAS with the character variable for illness in place of the dummy variables.

(q) The researcher believes there may be an interaction between the age and sickness variables. (i) Create the variables she would need to check this using the dummy variables and rerun the regression using all of the original predictor terms plus the interaction variables. (ii) Refit the interaction model in SAS using the character variable for illness level plus SAS's interaction notation in the model statement. (See below for instructions.)

(r) Is the model in part (q) better than the one in part (l)? Explain in terms of (i) the percentage of variability explained, (ii) the accuracy of the predictions and (iii) an appropriate test.

(s) What do the results of parts (q) and (r) tell you about whether there is a significant interaction between health and age and if there is an interaction, what levels of health seem to be driving it?

(t) Suppose a baby drinks 5 ounces of milk and is goes to bed at 8pm. Write down equations for the length of time the baby will sleep as a function of age for each of the three illness categories (not sick, mildly sick, very sick) and use them to sketch a plot of sleeping time (Y axis) versus age (X axis). Be careful to label the intercept and slope for each line with the actual numbers. Don't forget to indicate which line corresponds to which state of health. Explain briefly what this plot tells you about the interpretation of the interaction variables.

(u) Nora Numbskull looks at the printout from the interaction model with dummy variables in part (q) and says that using backwards stepwise regression you should remove the interaction between age and the indicator for being severely sick. Explain (i) why Nora thinks what she does and (ii) why she is wrong, befitting her name.

(v) If you were using forwards stepwise model selection, what would happen at the first step? Explain how you could figure this out without using any regression models and carry out the corresponding calculation.

STATA and SAS Commands

For this assignment you need to be able to create new variables (in order to do transformations), perform partial F tests tests and obtain various diagnostic plots and statistics related to regression assumptions and outliers. The corresponding commands are given below.

IN STATA:

Creating New Variables

This is a recap from last time. New variables are created using the **gen** command. You type "gen" followed by the name of the new variable followed by the mathematical expression for the new variable. For instance, in warm-up problem 3 if you wanted to create the age squared term so you could fit a quadratic model to the weightgain data you would type

gen momagesq = momage*momage

Similarly to create the inverse variable for the later part of the problem you would type

gen invmomage = 1/momage

Partial F Tests

Partial F tests are just simultaneous tests that several of your regression coefficients are 0. They are obtained using the **test** command directly after fitting the model of interest. For example, in warm-up problem 4 you are asked to do a test of whether it is worthile to add to indicator variables for type of alcohol consumed to the model. These indicators are named "beer" and "liquor." To do the partial F test you would simply type

regress bac amount time timesq weight meal beer liquor
test beer = liquor = 0

Diagnostics

In order to assess the regression assumptions and check for outliers or influential points you need to be able to store certain outputs from the regression and create plots. First, to calculate and save the residuals and diagnostic values you use the **predict** command. The predict command is followed by the name of the variable where you wish to store the output, a comma, and then the option specifying which sort of thing you want to store. Below I show the commands (in order) for storing the fitted values ($\hat{Y}$), the residuals, the studentized residuals, the leverage values (h_{ii}) the dfbetas for a particular predictor, the dfits and the Cook's distances (D_i) based on a simple linear regression of pregnancy weightgain on age using the data from warm-up problem 3. Note that these commands must be executed following the fit of the model for which you wish to obtain the various diagnostics! I have tagged wt on the names of all the outputs to emphasize that this is for the weight gain data set but you can use whatever names you like. Just make sure you can tell one set of residuals apart from that for another model!

predict wtyhat **Note–this is the default prediction, no option needed!**
predict wtresids, residuals
predict wtstresids, rstudent
predict wtleverage, leverage
predict wtdfbmomage, dfbeta(momage)

predict wtdfit, dfit
predict wtcooksd, cooksd

To look at the various values simply go into the data editor and you will see the columns corresponding to the diagnostics you have created. Or of course you can get STATA to print them to its analysis window.

Next you will need to create various plots. The commands are as follows:

Scatter Plot: To create a scatter plot you just type **scatter** followed by the Y variable followed by the X variable. For instance for warm-up problem 3 you would type

scatter weightgain momage

Histograms: You can either go through the graphics dropdown menu and select histogram and enter the variable name in the resulting menu or you can type the syntax directly at the command line. The basic command is **histogram** followed by the name of the variable for which you wish to get a histogram. For instance to get a histogram of the residuals for the pregnancy weightgain SLR createed above you would type

histogram wtresids

Normal Probability or Normal Quantile Plot: To check the normality of any variable you can use the command **pnorm** or **qnorm** followed by the name of the variable whose normality you wish to check. The first command produces a standardized plot of the percentiles of a normal distribution represented by your data compared to the (uniform) percentiles you would expect for a sample of size n. The second version plots the values you observed versus the ones you would have expected to see if your data had really been normal. In both cases you would expect the plot to be a straight line if your data were really normal. For example, for the residuals of the pregnancy weightgain data we would type

qnorm wtresids

There are several versions of the basic residual plot that you might wish to get. You can plot residuals versus the X variables, residuals versus the fitted values, residuals versus the observed Y's, and so on. All of these plots can be obtained using the scatter command:

scatter wtresids momage
scatter wtresids wtyhat
scatter wtresids weightgain

You can also use the graphics menu to get many of the regression diagnostic plots automatically and there is a shortcut command to get the plot of the residuals versus the fitted values. After fitting a regression model just type **rvfplot**.

Excluding values

Sometimes you want to be able to exclude a value (say an outlier) from an analysis. You can type most of the major STATA commands followed by **if** followed by an expression which says which points are to be included in the analysis. For instance, in warm-up problem 4 the outlier is a point with both a time and a bac value of 0. We could type

regress bac time if time > 0
regress bac time if time $!= 0$

Both of these would fit the regression without the point in question. (! = means not equal, == means exactly equal). You can also of course just go into the data editor and delete the point.

IN SAS:

Creating a new variable from existing variables:

In SAS you can use a **data** statement to create new variables from variables in an existing data set. For instance to get the age squared variable for warm-up problem 3 to fit a curvlinear model we would type

```
data work.hw6; set work.6;
momagesq = momage*momage;
run;
```

If we wanted to create a new file we would give its name immediately after the word "data". The "set" command tells SAS what data set is being drawn from to create the new variables.

Partial F Tests and Interactions:

In SAS as part of **proc reg** you can include a **test** statement very similar to the one on STATA. You can include as many test statements as you want as part of a model and name them to avoid confusion in the printout. For instance to do the test for the alcohol type indicators in warm-up problem 4 as part of our multiple regression command we would type:

```
proc reg data = work.hw6;
model bac = mount time timesq weight meal beer liquor;
eqcoeff: test beer=liquor=0;
run;
```

The test would appear under the name "eqcoeff" but including a name label is optional.

A second approach which gets you partial F tests for multicategory qualitative variables is to use **proc glm**. You can create a single column which has category names in it for the variable in question instead of a whole set of dummy variables. You then use the class command to warn SAS this is a categorical variable. SAS will automatically generate the dummy variables (though you won't see them) and produce the partial F test corresponding to including them all in the model. Note that SAS can also build product variables and interactions for you automatically without you having to create the product variables. You just use the * for product) or — (for interaction) symbol in your model statement. The code below is for the turn-in problem, and gives both the categorical variable for the babies' health and its interaction with age:

```
proc glm data = work1.hw6;
class healthstatus;
model sleeplength = milk milk*milk bedtime healthstatus|age;
run;
```

Diagnostics

In SAS you can easily generate residual and influence diagnostics by adding options to the model subcommand in proc reg. The option **r** produces an analysis of residuals and the option **influence** produces all our favorite measures for getting outliers. FOr instance to get both in warm-up problem 4 we would type

```
proc reg data = tmp1.hw6;
model bac=amount time timesq weight meal beer liquor/r influence;
run;
```

You can also get SAS to save these values in a data set using the **output** subcommand (see SAS help under proc reg) and create diagnostic plots using the **plot** subcommand. Plotting in SAS is a bit of a pain though so I recommend STATA for this part :)