Homework Assignment 5

Due Date: Monday, November 14th

Note: There are 7 problems on this assignment. The first 5 provide examples and extra practice. They are relevant for the exams but do not need to be turned in. Solutions for them are available on the class web site. You must turn in Problems 6-7 together with the corresponding STATA/SAS printouts to receive full credit. The assignment is due Monday, November 14th. 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 3:00 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) Still More Height and Weight A doctor is interested in understanding what factors predict the weight of teenagers. As we have seen before, height is a good predictor of weight. The doctor suspects age may also be an important factor. She collects data from the next ten teens who come to her office on weight (in pounds), age (in years) and height (in inches). The data are presented in the accompanying table and in the data files on the class web site.

Weight	Age	Height	Weight	Age	Height
90	16	63	86	13	47.8
72	13	48.6	112	14	55.8
132	16	62.2	122	16	55
100	14	48.6	88	13	50.2
62	10	39.8	98	15	56.6

(a) Fit two separate simple linear regressions, one of weight on height and one of weight on age. Is there a significant linear relationship in each case? Explain briefly.
(b) Which of age and height is the better individual predictor of weight? Justify your answer using three appropriate numbers from the regresison printouts.

(c) Fit the multiple regression of weight on height and age. What is the estimated regression equation? Does the regression overall explain a significant amount of the variability in Weight? Explain using an F test. Make

sure to write out the null and alternative hypotheses, both mathematically and in words, give the test statistic and p-value, say whether or not you reject and why, and explain your real-world conclusions. Use $\alpha = .05$.

(d) In the regression of part (c) are the variables height and age significant? Explain using individual t tests. In each case, write the null and alternative hypotheses, both mathematically and in words, give the test statistic and p-value, say whether or not you reject and why, and explain your conclusions. Use $\alpha = .05$.

(e) Explain what has caused the apparent contradictions in (a)-(d). What other evidence do you see of this in the printout besides the contradictions in the hypothesis tests?

(f) Verify your suspicions from part (e) in three ways: first by calculating an appropriate set of correlations in STATA or SAS, second by calculating the variance inflation factors for the age and height variables in STATA or SAS directly as part of the regression fit, and third by using appropriate simple linear regressions and hand calculations to get the variance inflation factors. Explain what you learn from each of these calculations.

(g) How would you fix the problems described in parts (e)-(f)? Be as specific as you can.

(h) Explain how what we have observed could be thought of as height mediating the effect of age on weight. Does this appear to be a case of full or partial mediation or neither? Can we draw any conclusions about the causal pathway based on this data? Discuss.

(2) Harry Potter and the Sorcerer's Statistic: A longer version of this problem was included in the midterm review set. I focus here on the parts having to do with multicollinearity and indicator variables. Recall that Polygon Pictures, the film-making branch of Mathematical Media Incorporated, is interested in knowing what factors contribute to the profitability of their movies. For their last n=27 films they have recorded Y, the box-office sales (in millions of dollars), X_1, the production costs for the film (in millions of dollars), X_2, the number of theaters in which the film was shown, X_3 the advertising budget for the film (in millions of dollars), and X_4 which is 1 if the movie featured a big name star and 0 if it didn't. They have also classified the films as action/adventure ($X_5 = 1, X_6 = 0$), comedy ($X_5 = 0, X_6 = 1$), or romance ($X_5 = 0, X_6 = 0$). A multiple regression printout for their data is shown below along with some possibly helpful statistics. Use this information to answer the following questions.

```
The regression equation is
Box = - 0.842 + 1.84 Cost + 0.0025 Theaters - 0.628 Ads + 5.47 Star
          + 4.59 Action - 5.14 Comedy

Predictor       Coef       Stdev    t-ratio        p
Constant     -0.8423      0.9715      -0.87    0.396
Cost          1.8365      0.1339      13.71    0.000
Theaters      0.0025    0.001455       1.71    0.102
Ads          -0.6282      0.5574      -1.13    0.273
Star          5.4713      0.7191       7.61    0.000
Action        4.5880      0.6523       7.03    0.000
Comedy       -5.1441      0.6412      -8.02    0.000

RMSE = 0.8970      R-sq = 99.7%     R-sq(adj) = 99.6%

Analysis of Variance
SOURCE       DF          SS          MS         F        p
Regression    6     5665.32      944.22   1173.54    0.000
Error        20       16.09        0.80
Total        26     5681.41
```

```
Correlations:
            Box Theaters      Ads
Theaters  0.900
Ads       0.925    0.889
Cost      0.950    0.912    0.927
**********************************
Summary          N     MEAN   MEDIAN    STDEV
Box             27    35.00    25.00    14.78

               MIN      MAX       Q1       Q3
Box           5.00    70.00    20.00    40.00
**********************************
```

(a) In what sense could you think of this problem as an Analysis of Covariance? What variables would represent the base ANOVA and what variables are you adjusting for. Why do you think it would be important to adjust for those factors?

(b) Which type of film is generally most profitable, all other things being equal? An action/adventure film, a comedy, or a romance? Explain briefly.

(c) Do you think there is a significant difference in Box Office take between Action and Comedy Films, all else equal? Explain briefly. If you wanted to test this formally, what procedure would you use in STATA or SAS?

(d) Is there any evidence of multicollinearity in this model? Briefly justify your answer. What variables do you think you would most likely include in a final model for box-office sales and why? And why can you not be sure from the printout that your choice is correct?

(3) Interaction Basics: We will discuss this example in class but I include it here for those who are making an early start on the homework. It is another version of our "exercising heart rate" problem. A researcher is studying the relationship between the length of time a person jogs (in minutes), X_1, and their heart rate at the end of the run, Y (in beats per minute). She suspects that there may be a difference in increase in heart rate between people who run regularly and people who do not so she also records X_2, an indicator that is 1 for people who exercise regularly and 0 for those who do not. She fits a regression model with main effects for length of run and exercise category and an interaction term between the two variables and gets the following estimated regression equation:

$$\hat{Y} = 80 + 1.2X_1 - 15X_2 - .2X_1 * X_2$$

(a) Write down the equation giving the relationship between exercise time and heart rate for (i) a person who does not exercise regularly and (ii) a person who does exercise regularly.

(b) Explain carefully what each of the regression coefficients, $b_1, b_2, b_{1,2}$ means in this model.

(c) Suppose the interaction term was not statistically significant (i.e. you could not reject the hypothesis that $\beta_{1,2} = 0$). What would that tell you about the relationship between exercise time, regularity of exercise and heart rate?

(4) Don't Drink and Derive: The printout below shows a multiple regression in which the outcome of interest, Y, was blood alcohol level or BAC, measured as a percentage. The predictor variables were 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 (more about that on HW6!), 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), the type of alcohol consumed

($X_6 = X_7 = 0$ for wine, $X_6 = 1, X_7 = 0$ for beer and $X_6 = 0, X_7 = 1$ for hard liquor). The printout is included for your reference although you won't need any of the numeric values this time.

```
Multiple Regression Model:
. reg BAC Amount Time TimeSq Weight Meal Beer Liquor

      Source |       SS       df       MS              Number of obs =     122
-------------+------------------------------           F(  7,   114) =  111.80
       Model |   .07850346     7   .01121478           Prob > F      =  0.0000
    Residual |  .011435201   114  .000100309           R-squared     =  0.8729
-------------+------------------------------           Adj R-squared =  0.8650
       Total |  .089938661   121  .000743295           Root MSE      =  .01002

------------------------------------------------------------------------------
         BAC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Amount |    .0045     .0003546    12.62   0.000     .0037716    .0051764
        Time |      .12     .0071479    16.50   0.000      .103769    .1320887
      TimeSq |     -.04     .0020517   -19.38   0.000     -.043817    -.035688
      Weight |  -.00007     .0000255    -2.83   0.006    -.0001225   -.0000216
        Meal |     -.01     .0022286    -4.59   0.000     -.014637   -.0058073
        Beer |   -.0042     .0018844    -2.23   0.027    -.0079439   -.0004778
      Liquor |     .022     .0022718     9.69   0.000     .0175146    .0265153
       _cons |    -.033     .0081909    -4.02   0.000    -.0491158   -.0166636
------------------------------------------------------------------------------
```

(a) In the model above we included only main effects for Amount and Alcohol Type. Explain why it might make sense to include an interaction of Amount and Type.

(b) Write down carefully the interaction variable(s) you would need to add to the model to do this and what their interpretations would be.

(c) Can you think of a less complicated way of achieving the same goal in this model with a single variable (instead of including amount, type indicators and interactions)?

(5) Short(?) and Sweet: People with Type I diabetes need to carefully monitor their blood-sugar levels and take insulin every day. Professor Sweet, a diabetes researcher, has developed a new device for patients to wear that automatically monitors blood-sugar levels and provides insulin as needed. She wants to see whether people using her device do better at controlling their blood sugar levels (i.e. keeping them down near the normal level of 100) than people using traditional insulin control methods. She is also interested in whether the effectiveness of the device is related to gender. Thus she recruits 32 men and 32 women and randomizes within gender so that 12 men and women track their blood-sugar using the current standard methods (C) while the other 20 of each gender use her new devices (N). After giving the subjects time to adapt to the new devices she records their blood-sugar levels. A STATA printout of the resulting ANOVA for the four groups (MC, FC, MN, FN) is shown below. Use it to answer the following questions which focus on viewing her ANOVA as a regression model.

```
. oneway BloodSugar Group, tabulate
```

Group	Summary of BloodSugar: Mean	Std. Dev.	Freq.
MC	125	20.0	12
FC	120	20.0	12
MN	100	13.9	20
FN	105	14.0	20
Total	110	18.95	64

Analysis of Variance

Source	SS	df	MS	F	Prob > F
Between groups	6400	3	2133	7.90	0.00016
Within groups	16200	60	270		
Total	22600	63			

(a) Suppose Dr. Sweet had fit her data using a regresison model with women using the current insulin control methods as her reference group and a separate indicator for each of the other groups. Define the indicator variables she would have used and write down the corresponding regression equation.

(b) How would your answer to part (a) change if you had men using the new device as your reference group?

(c) Based on the information given in the ANOVA printout, find R^2, R^2_{adj} and the root mean squared error for the regression model, showing your work. Does the model do a good job (i) in terms of the percentage of the variability in blood-sugar levels that is explained by the treatment group and (ii) in terms of making accurate predictions of blood-sugar level? Briefly explain your reasoning in each case.

(d) Overall do you have evidence that either gender or insulin monitoring method are associated with insulin level? Exaplain briefly what test you are performing.

(e) Do we have enough information to fill in the traditional parameter estimates table (coefficients, standard errors, t-tests, confidence intervals) for this model? If not, why not? If yes, explain briefly how you would do it. You do not need to do the calculations though it would be a good idea to be sure you can!

As an alternative to the model in parts (a) or (b), Professor Sweet considers fitting a regression with indicators for gender and treatment, plus an interaction term. Specifically, she lets $X_G = 1$ for males and $X_G = 0$ for females, and lets $X_T = 1$ for people who used her new device and $X_T = 0$ for those who had treatment as usual. The resulting STATA printout is given below. Use it to answer part (f)-(h).

```
. regress Sales Expense
      Source |       SS       df       MS              Number of obs =      64
-------------+------------------------------           F(  3,   60)  =   7.901
       Model |     6400        3     2133              Prob > F      =  0.0002
    Residual |    16200       60      270              R-squared     =
-------------+------------------------------           Adj R-squared =
       Total |    22600       63   358.73              Root MSE      =   16.43

------------------------------------------------------------------------------
       Sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Gender |         5       6.71       0.75  0.459         [ -8.42, 18.42]
   Treatment |       -15       6.00      -2.50  0.015         [-27.00, -3.00]
Gender*Treat |       -10       8.49      -1.18  0.247         [-26.98,  6.98]
       _cons |       120       4.74      25.30  0.000        [110.52, 129.48]
------------------------------------------------------------------------------
```

(f) How did Dr. Sweet define the interaction variable and how would she have created it in STATA or SAS from the indicators for gender and treatment? Explain briefly.

(g) Explain carefully what the interaction term between gender and treatment means.

(h) Is there evidence of a significant interaction in this model? What does that tell you?

(i) The coefficient of the interaction term is equivalent to a particular contrast. Explain what this contrast is and therefore show how I was able to calculate the standard deviation, s_{b_3} and the p-value for the interaction term using the information from the ANOVA table.

Problems To Turn In

(6) Relationships of the HAART: Professor U. R. Helpful and his AIDS data from Homework 2 are back. Previously we analyzed them using ANOVA methods. Now we are going to attack the problem using two different regression approaches. Recall that Dr. Helpful is interested in how different forms of "highly active antiretroviral therapy" or HAART affect patients' viral loads. He has selected n=124 HIV positive individuals and randomly assigned them to four treatment groups: nonHAART, HAART A, HAART B, and a combination of HAART A and B). The accompanying data file contains his outcome variable, "vload", which is the $\log_{10}$ viral load and a set of indicator variables, N, Aonly, Bonly and Combo which are 1 if the subject is getting the treatment indicated by the variable name and 0 otherwise. Use the data to answer the following questions.

(a) Fit a multiple regression of $\log_{10}$ viral load on treatment group, using the nonHAART subjects as the reference group.

(b) Suppose you had instead wanted to use the HAART A only subjects as your reference group. Show how you could find the estimated regression equation **by hand** from the printout in part (a) and verify your result using STATA or SAS.

(c) Is there evidence that mean $\log_{10}$ viral load differs across the treatment regimens? Justify your answer by performing an appropriate **overall** hypothesis test. State the mathematical hypotheses in the regression framework, give the p-value, and your real-world conclusions.

(d) Does the model do a good job (i) in terms of the percentage of the variability in viral load that is explained by the treatment group and (ii) in terms of making accurate predictions of viral load level? Briefly explain your reasoning in each case. Are your answers surprising in real-world terms? Explain briefly.

(e) Test for differences among each pair of groups based on the model you fit in part (a). Note that you can get answers for three of the tests directly from the printout (explain why) but for the other three you will need to do follow up "tests" or "contrasts" in STATA or SAS. Make sure you indicate in these cases what contrasts you are using. Write out all the details of the test comparing the combination group to HAART B only (math and word hypotheses, test statistic, p-value) but for the others you may just give briefly justified conclusions. Explain as carefully as you can what this suggests about the relative merits of the treatment regimens.

(f) Suppose Professor Helpful wants a range of viral loads which will include 95% of patients taking the combination therapy. What sort of interval should he compute and why? (Note: You don't have to actually get the interval.)

(g) Suppose Professor Helpful wants to test whether patients taking HAART A have lower viral loads than patients not taking HAART A (regardless of whether or not they are taking HAART B). Write down the appropriate contrast in terms of the regression coefficients based on the model from part (a). You should give the null and alternative hypotheses mathematically and in words with a justification of your choice, and then use STATA or SAS to find the p-value and explain your real world conclusions using $\alpha = .05$. Perform a similar calculation for HAART B. You do not need to write out all the details. Just give your best estimate for the effect of HAART B and explain whether it has a significant effect.

In parts (h)-(j) Professor Helpful is interested in testing whether there is an interaction between HAART A and HAART B:

(h) Carefully define indicator variables corresponding to "takes HAART A or not" and "takes HAART B or not" and the interaction between them and create these variables in STATA or SAS.

(i) Fit the multiple regression with "main effects" for HAART A and HAART B and their interaction using the variables you defined in part (h).

(j) Is there evidence of a significant interaction between HAART A and HAART B? What does this tell you about the effects of these medications? Suppose your test had come out the other way. Explain carefully what your interpretation would have been in that case.

(k) (Optional Bonus) We never learned how to calculate CIs and PIs for Y in a multiple regression because the formulas for the standard deviations $s_{\hat{Y}_0}$ and s_{Y_0} are messy. However in an ANOVA context where there are only indicator variables it is possible to write down simple formulas. (i) Explain what the CI and PI formulas ought to be using ANOVA notation, (ii) get the interval Professor Helpful wants in part (f) and (iii) for the very brave, verify that if your ANOVA has two groups then these formulas reduce to the formulas we learned for simple regression involving $\bar{X}$ and SSX. (Note: To do the latter part you may find it useful to remember that your X values are all 0's and 1's denoting group membership, and to let n_1 be the number of 1's and n_2 the number of 0's, with $n_1 + n_2 = n$, the total sample size.)

(7) If Memory Serves.... Dr. Brain, a neuropsychologist, is interested in memory and aging, particularly as it relates to patients with a rare mental disorder. He has conducted a study with 30 patients with the disorder and 30 healthy controls in which he has recorded age (in years) and performance on a standard memory test. (Possible scores on the test range from 0 to 120 points with higher scores being better.) For the patients with the disease he has also recorded illness duration (the number of years they have had the disease) and a brain activation score (higher activation values in this particular brain region under these test conditions are an indication of neural degeneration.) In the accompanying data file the variables are labeled memory, age, group (1 = patient and 0 = healthy control), illnessdur and activation. Use $\alpha = .05$ for all hypothesis tests.

(a) It is well known that memory deteriorates with age, even in healthy subjects. Dr. Brain wants to know if there is a difference in the rate of deterioration with age between healthy subjects and people with the disorder he is studying. Explain why what Dr. Brain is interested in corresponds to an interaction between age and group, create the desired interaction variable, and fit a multiple regression of memory on age, group, and your interaction term.

(b) Is there evidence of a significant interaction between age and group in terms of their effect on memory? Justify your answer by performing an appropriate hypothesis test. Give the null and alternative mathematically and in words, the test statistic and your real-world conclusions using $\alpha = .05$.

(c) Write down the estimated relationship between memory test score and age for (i) a healthy control and (ii) a patient with the mental disorder Dr. Brain is studying based on your model from part (a), and provide a rough sketch of these relationships over a reasonable age range.

(d) Interpret as carefully as possible all the four regression coefficients from your model in part (a), making sure to include the units, numerical values and signs in your descriptions.

(e) Dr. Brain is interested in studying how various aspects of illness severity and duration are related to memory loss in the patients in his sample. His colleague suggests that he should adjust for age as a covariate in his patient only models. Explain why, based on your answer to part (b), this may not be a good idea.

(f) After thinking carefully about part (e), Sandra Superbrainy, Dr. Brain's doctoral student (who is good at statistics), suggests that he "regress out" the effects of normal aging to create a new memory score as follows. First, fit a simple linear regression of memory on age **using only the healthy controls.** Then use the coefficients from this model to obtain the predicted memory scores the **patients** would have based on their age if they were healthy. Finally, take the difference between the predicted scores and the actual scores of the patients to get new "age adjusted" memory scores. Follow Sandra's procedure to obtain a new variable and verify that you get the same values that are in my column labeled "memorynormed." Explain briefly whether you think Sandra's idea is a good one or not and why.

Note: For the remainder of the problem we will focus only on the 30 patients, treating the "age adjusted" memory score, memorynormed, as the outcome of interest.

(g) Fit a simple linear regression of memorynormed on age. Is the relationship significant? Briefly justify your answer and say why it makes sense given your findings in the earlier parts of the problem.

(h) Fit a multiple linear regression of memorynormed on age and illness duration. Is age a significant predictor of memory performance in this model? State the null and alternative hypotheses for the relevant test methmatically and in words and give your real-world conclusions.

(i) What do you think has caused the difference between your answers in (g) and (h)? Verify your suspicions

in the following four ways: (i) note any unusual features about the age predictor from the printout (ii) calculate an appropriate set of correlations in STATA or SAS (iii) calculate the variance inflation factors for the age and illness duration variables in STATA or SAS directly as part of the regression fit and (iv) use appropriate simple linear regressions and hand calculations to get the variance inflation factors. What should you do to resolve the apparent problem.

(j) Last but not least....Dr. Brain believes that the relationship between illness duration and age-adjusted memory score is mediated by the brain activation score he recorded for the patients. Draw the mediation diagram that corresponds to this scenario and fit the sequence of regression models he would use test it. Is the data consistent with the mediation hypothesis, either fully or in part? Explain. Can Dr. Brain conclude based on these tests that his mediation hypothesis is correct, or is there an alternative theory that might fit the data? Explain.

(k) Optional Bonus: For those who want to try a formal mediation test, go to the web site of David Kenny, http://davidakenny.net/cm/mediate.htm, and read the main page. Then follow the link to the Sobel test and perform the test for the strength of the indirect path corresponding to part (j).

STATA and SAS Commands

For this assignment you need to be able to create new variables, perform tests about the $\beta's$ in a multiple regression model and obtain variance inflation factors as part of a regression output. You may also find it useful to recall how to obtain correlations. The corresponding commands are given below.
IN STATA:

(i) Creating a new variable from existing variables: The easiest way to do this is with the **gen** or **generate** command. For instance, if you have two predictor variables, X and W, and you want to create the interaction variable Z = X*W you simply type

gen Z = X*W

(ii) Tests about regression coefficients: Just as we could perform tests about linear combinations of means by using the **test** command following the **anova** command, we can perform tests about combinations of coefficients by using **test** after **regress**. As before, make sure you use the test statement immediately after fitting the relevant model as STATA always refers back to the most recent model. You simply type the expression in terms of the relevant variable names. So for example, to compare the coefficients of the age and height variables in warm-up problem 1 (though I can't see why you would want to!!) you would type the following sequence of commands:

regress weight age height
test age = height

You can make the expressions as complicated as you like, just as after the anova command.

(iii) Obtaining variance inflation factors: After the regress command has been executed simply type

estat vif

and STATA will give you the variance inflation factors for all the variables in the model.

(iv) Obtaining correlations: To get all pairwise correlations among a set of variables, say var1, var2 and var3 simply type

cor var1 var2 var3

IN SAS:
(i) 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 product of height and age (i.e. an interaction variable) in warm-up problem 1 and store it as part of the same data set we would type

```
data work.hw5; set work.hw5;
heightage = heaight*age;
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.

(ii) Tests about regression coefficients: 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, do duplicate the test comparing the coefficients of age and height in the warm-up problem 1 model two different ways with the names "eqcoeff" and "eqcoeffv2" we would type:

```
proc reg data = work.hw5;
model weight = age height;
eqcoeff: test age = height;
eqcoeffv2: test age - height = 0;
run;
```

(iii) Obtaining variance inflation factors: In SAS, variance inflation factors are obtained as part of the **model** statement in **proc reg** by adding a slash and the option **vif**. For example, if you are doing a regression of Y on X1, X2, and X3 you would simply type

```
proc reg data = work.hw5;
model Y = X1 X2 X3/vif;
run;
```

(iv) Obtaining correlations: As we have seen before, in SAS correlations are obtained using **proc corr.** For example, to get the correlations amoung var1, var2 and var3 simply type

```
proc corr data = work.hw5;
var var1 var2 var3;
run;
```