Homework Assignment 3

Due Date: Monday, October 17th

Note: There are 8 problems on this assignment. The first 5 problems should provide you with basic practice on the material. They are relevant for the exams, but do NOT need to be turned in. You must turn in problems 6-8 to receive full credit. The assignment is due Monday, October 17th. 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. Problems labeled (GS) are adapted from our optional text, *Primer of Applied Regression & Analysis of Variance* by Glantz and Slinker.

Warm-up Problems

(1) Birthweight and Smoking: This continues Problem 3 of homework 2 where a researcher was looking at the birthweights of mothers who had never smoked, were former smokers, were current light smokers or were current heavy smokers. For your reference the data are reproduced in the accompanying Excel, STATA and SAS files. The variables are called "birthweight" and "smokingstatus".

(a) Suppose we wanted to use the Bonferroni procedure to correct for multiple testing when performing all pairwise comparisons of the groups at an overall significance level of $\alpha = .05$. How many comparisons are there and what significance level should we use for the individual tests?

(b) Suppose we calculated confidence intervals for each of the pairwise comparisons first unadjusted and then using the Bonferroni procedure. Which set of intervals would be wider? Explain.

(c) On homework 2 you obtained the unadjusted p-values for the hypothesis tests corresponding to the pairwise comparisons. Which groups were significantly different from each other? Do those results change if you use the Bonferroni procedure? Explain. (Note: This should not require any additional calculations.)

(2) Relationship Basics:

(a) Explain the difference between the simple linear regression model, the simple linear regression equation, and the estimated simple linear regression equation.

(b) Explain why the method of least squares gives you a "good" estimate of the simple linear regression equation.

(c) List three variables that you think might be useful in predicting a person's cholesterol level. State how you think each of these variables would be correlated with cholesterol level individually (positive or negative, linear or curved). Write down one possible model for the relationship between cholesterol and the three variables as a group. (You do not need to include numbers–in fact you really can't since you have no data and in any case we haven't done multiple regression yet!)

(3) Midterms Make You Sick: Note that this was the regression problem on an old midterm I gave and is fabulous practice for our exam.

A statistics professor at the University of Calculationally Learned Adults has noticed that the number of cases of statisticitis (a special form of math anxiety!) in her large lecture class seems to go up the closer it gets to exam time. She has surveyed her students on n=12 different days and recorded Y, the number of cases of statisticitis on a given day (per 100 students) and X, how many days it is until an exam. For instance, X=2 would mean that there is an exam in two days time, and Y=50 would mean that 50 out of every 100 students, or 50%, are sick with statisticitis. Some useful numbers are given below.

$\bar{X} = 5$
$\bar{Y} = 25$
SCP = -200
SSX = 40
SST = 1250
n=12

(a) Find the estimated regression equation for predicting Y, the number of statisticitis cases per hundred students, based on X, how many days it is until an exam. (In other words, find b_0 and b_1.)

(b) Give the units and real-world interpretations of b_0 and b_1.

(c) Find SSR, SSE, R^2, and Root MSE from the given information. (Note: You may find it helpful to know that $SSR = b_1 * SCP$.)

(d) Find the correlation between X and Y. Is there a strong relationship between time until an exam and the number of statisticitis cases? Explain briefly.

(e) Does time until the exam, X, explain a large **percentage** of the variability in the number of statisticitis cases, Y? Explain briefly.

(f) Does time to the exam give good **predictions** of the number of cases of statisticitis that will occur? Explain your reasoning briefly.

(g) Predict the number of statisticitis cases that would occur (i) a week before the exam and (ii) 20 days before the exam. Do these answers make real-world sense? Explain what has happened.

(4) Height and Weight:

(a) A UCLA biostatistics professor wants to know what factors are predictive of a person's weight. She surveys her class and obtains data about weight, height, age, and number of hours spent exercising per week. Below are the correlations and covariances she finds for each of the three predictors with weight. Explain

which pair you think goes with each of the predictors, say which variable has the strongest relationship with weight, and which variable has the weakest relationship. In each case briefly justify your answer.

	Predictor A	Predictor B	Predictor C
Covariance	70.000	-30.000	35.000
Correlation	0.700	-0.750	0.100

Now suppose that the professor obtains a regression equation of $\hat{Y} = 40 + 2X$, where Y is a person's weight in pounds and X is the person's height in inches. Use this information to answer parts (b)-(e).

(b) What are the units of b_0 and b_1 the slope and intercept of the estimated regression equation? What are the real world interpretations of b_0 and b_1?

(c) If a person is 5 feet tall what is their predicted weight according to this model?

(d) According to this model, how much will a person who is 0 inches tall weigh? How much would a newborn infant who is 18 inches long weigh? Do these predictions make real-world sense? Explain.

(e) Why did the equation give such bad predictions in part (d)?

(5) The Tranquilizing Effect of Homework? (Based on GS Problem 2.2): Tranquilizers such as Valium work by binding to specific receptors in the brain which in turn causes changes in nerve activity. However it is hard to measure receptor binding and nerve activity directly in humans. Hommer et al (*Archives of General Psychiatry* vol 43: 542-551, 1986) instead looked at the effects of different doses of Valium on easily measured physiological variables and then looked at the correlations among these variables to attempt to identify which were most strongly linked to the effect of the drug. Two variables they measured were a subject's sedation and their blood level of cortisol, a hormone. The data are given in the accompanying file. Use it to answer the following questions, performing all calculations in STATA or SAS.

(a) Obtain a scatterplot of sedation level, Y, versus cortisol level, X. Is the relationship positive or negative? Explain in real-world terms what this means. Based on the plot, do you believe this is a strong relationship? Explain.

(b) Find the correlation and covariance between sedation and cortisol level. Is your impression about the strength of the relationship verified? Explain which number you are using to answer this question.

(c) Give the units and real-world interpretations of b_0 and b_1 for this model.

(d) Is there a significant relationship between cortisol level and sedation score? You should set up the hypotheses and test statistics for the appropriate test by hand and then base your conclusions on appropriate p-values from the printout.

(e) Suppose a physician wants to be able to predict what a subject's sedation level will be based on their blood-cortisol level. Will they be able to do so accurately? Explain what the typical error would be in percentage terms.

Problems To Turn In

(6) Fun with Multiple Comparisons:

(a) As a supplement to our discussion in class, read the handout in the course information section of the class web site on multiple comparisons (courtesy of my colleague Professor Belin with some additions from me).

(b) Explain briefly the difference between the Tukey and Scheffe procedures and when they can be applied.

(c) Explain briefly the idea behind the omnibus approach to multiple comparisons and the false discovery rate approach. How do these methods differ philosophically from the other methods in the handout?

For the remainder of the question we will use the pollution data from Problem 6 of homework 2. Recall that Professor Urtha Green was comparing particle levels at 5 different locations at local elementary schools. The data are included for your convenience in the accompanying Excel, STATA and SAS files. The variables are named "particle" and "location". (I have updated the location variable so it is a factor rather than a string.)

(d) Suppose you were using the Bonferroni procedure to create 95% confidence intervals for the pairwise differences in means. (i) What value of t would you use and (ii) how much wider would your intervals be (on a percentage basis) than if you used uncorrected intervals? Briefly explain your reasoning.

(e) Use STATA to get the "Bonferroni adjusted" p-values for this data (see instructions below–this was originally part (b) of Problem 5 on HW2) and use them to answer the following questions:

(i) Based on the printout, which pairs of locations do **not** have significantly different particle concentrations? Use this information plus your results from HW2 to describe as precisely as you can the ordering of the locations in terms of air quality for the students.

(ii) Are your results any different from what you obtained on homework 2, part (c) using the uncorrected p-values? (Note: You don't have to redo the analysis from HW2–simply use the p-values I provided in the solutions.)

(iii) Explain briefly why STATA provides adjusted p-values instead of doing the Bonferroni procedure as we learned it in class.

(iv) **Optional Bonus:** Suppose you wanted to calculate the unadjusted p-values from the adjusted ones. Can you always do so exactly? Does it matter if you can't? Discuss briefly.

(f) The Holm procedure is a modification of the Bonferroni procedure for multiple testing. Explain why the Holm procedure may be preferable to the Bonferroni procedure and then use the original (unadjusted) p-values for the pairwise comparisons from HW2, part (c) to obtain the Holm results for the particle data. Discuss briefly how your results compare to those from HW2 and from the Bonferroni procedure you used in part (e).

(g) Optional Bonus: Use SAS to get the Tukey multiple comparison confidence intervals for these data (see instructions below). How do your results compare to those from the Bonferroni and Holm procedures? In general how would you expect these methods to compare in terms of how conservative they are?

(7) A Healthy Homework Exercise:

Two of the variables in my class databases for the last few years have been weight (in pounds) and time spent exercising (in hours per week). In this problem we will build a model to predict weight, Y, based on hours exercised, X using a sample of size n=101 students. Some summary information is given below. (I have somewhat modified the observations from my databases to simplify the question.)

$\bar{X} = 4$
$\bar{Y} = 150$
$SCP = -900$
$SSX = 600$
$Var(Y) = 900$

(a) Find b_0 and b_1, the estimates we would get for the slope and intercept if we fit a simple linear regression to this data and write down the corresponding estimated regression equation.

(b) Give the units and real-world interpretations of b_0 and b_1 and say briefly whether they make practical sense. Your answer should incorporate the numerical values from part (a).

(c) According to this model, what is the average weight of a person who exercises 10 hours per day? Does your answer make real-world sense? If so why? If not, what has gone wrong?

(d) According to this model, if you wanted to lose 6 pounds (and keep it off) how much additional exercise would you need to do each week beyond what you currently do? What assumption are you making to do this calculation and why might it not be correct?

(e) Find the covariance and correlation between X and Y. Show your work. Is there a strong relationship between X and Y? Explain what number(s) you are using to answer this question and whether your answer is surprising in real-world terms. (Note: You may find it helpful to remember that $Var(X) = SSX/(n-1)$ and that $SD(X) = \sqrt{Var(X)}$.)

The ANOVA table for these data is given below. Use it to answer the rest of the problem. Note: You may need to perform some simple calculations from the numbers given in this table to get your final answers!

Source	SS	df	MS	F	Prob > F
Regression	1350	1	1350	1.51	.2221
Error	88650	99	895.45		
Total	90000	100			

(f) What proportion of the variability in weight is explained by how much the students exercise? Show your work and explain whether you think the model is doing a good job in this regard.

(g) Does amount of time exercised do a good job of predicting weight? Carefully explain your reasoning.

(h) Is there a significant relationship between weight and amount exercised? Justify your answer by performing an appropriate hypothesis test. Give the null and alternative hypotheses mathematically and in words and state your real-world conclusions.

(i) Optional Bonus for the algebraically inclined: Explain how you could have calculated the ANOVA table by hand from the information given at the beginning of the problem.

(8) Parenteral Nutrition (Based on GS Problem 2.6):

When patients are unable to eat for long periods, they must be given intravenous nutrition, a process called parenteral nutrition. Unfortunately, patients on parenteral nutrition show increased calcium loss via their urine, sometimes losing more calcium than they are given in their IV fluids. Such calcium loss might contribute to bone loss as the body pulls calcium out of bones to try to keep the calcium level in the blood within normal range. In order to better understand the mechanisms of urinary calcium loss, Likin et al (*American Journal of Clinical Nutrition*, vol. 47:515-523, 1988) measured urinary calcium (Y) and related it to four factors: dietary calcium, dietary protein level, urinary sodium, and glomerular filtration rate (gfr) which is a measure of kidney function. Their data are given in the accompany file with the names urinarycalcium, dietarycalcium, gfr, urinarysodium and protein. Use these data to answer the following questions. All caculations should be done in STATA or SAS and you should mark on the printouts the pieces used to answer each part of the problem.

(a) Obtain scatterplots of urinary calcium versus each of the other four measurements. Based on these plots which relationship appears to be the strongest? Explain.

(b) Find the correlations between urinary calcium and each of the other four measurements. Which variable has the strongest relationship and which has the weakest and why?

(c) Fit a simple linear regression of urinary calcium on each of the other four measurements.

(d) Note that SST and the degrees of freedom in the ANOVA tables from your four regressions are all the same. Explain briefly why this is the case.

(e) Which of the four predictor variables have significant linear relationships with urinary calcium level? Explain in terms of the p-values of an appropriate set of tests. You do not need to write out all the details of the tests.

(f) Which of the four variables do you think is most important for explaining urinary calcium? Explain in terms of (i) the percentage of variability explained, (ii) the quality of the predictions and (iii) your answer to part (e).

STATA and SAS Commands

To do this assignment you will need several new commands. The first relate to multiple comparisons. Both SAS and STATA have add-ons to their ANOVA commands that make doing multiple comparisons easy. Second, you will need to be able to obtain correlations, covariances and simple linear regressions and produce scatterplots. The commands needed to do this are given below.

(1) STATA COMMANDS:

Recall that in STATA you can perform an ANOVA using either the **anova** or **oneway** commands. The command **oneway** has several options for multiple comparisons including **bonferroni** which can be shortened to just **b** and **scheffe** which can be shortened to **sc**. To get p-values for all the pairwise comparisons on the birthweight and smoking example (warm-up problem 1) using these methods one would type one of the following:

oneway birthweight group, bonferroni
oneway birthweight group, b

oneway birthweight group, scheffe
oneway birthweight group, sc

Note that what STATA provides are ADJUSTED p-values that get compared directly to the overall significance level you want! For instance, for the Bonferroni procedure it multiplies the unadjusted p-values by the number of pairwise comparisons. Think about why it might do this....

Next we consider the commands for relationships among variables:

The **correlate** command: This command is used to calculate correlations or covariances. You simply type **cor** followed by the list of variables and STATA will give you all the pairwise calculations. If you use the optional command **c** or **cov** at the end, STATA will give you covariances instead of correlations. For example for warm-up problem 5 I have named the variables "sedation" and "cortisol." The corresponding commands are

cor sedation cortisol
cor sedation cortisol, cov

The latter gives the covariance. Don't forget the comma after the last variable name!

The **regress** command: This command is used to fit a linear regression model. For now we will just use it to obtain a simple linear regression fit. Type **reg** (or regress) followed by the name of your response variable (Y) followed by the name of your predictor variable (X). For example in the sedation-cortisol example of warm-up problem 5 we would type

reg sedation cortisol

The **scatter** command: To get a scatter plot of X versus Y STATA has a shorthand. Just type **scatter** followed by the Y variable and then the X variable. For the sedation-cortisol example you would type

scatter sedation cortisol

In STATA you can also use the menus to create plots. In version 9 you select Graphics/Easy Graphs/Scatter plot and enter your Y and X variables. In version 11 you select Graphics/Two-way graph and then click "Create" in the resulting pop-up box and select from the various options.

(2) SAS COMMANDS:

In SAS we used either **proc anova** or **proc glm** to do analysis of variance. The anova procedure has options for doing multiple comparisons of MANY types. The basic set up is

```
proc anova data = work.hw3;
class group;
model birthweight = group;
means group/bon;
run;
```

The "/bon" after the means command tells SAS you want Bonferroni multiple comparisons of all the pairwise means. SAS will produce all the adjusted pairwise confidence intervals and tell you which tests are jointly significant but it does not give adjusted p-values like STATA. The default overall significance level is .05. Other options that can be used after the slash include duncan, dunnett, scheffe, lsd, snk, and tukey. The

same add-ons can be used in **proc glm**.

Next we consider the commands for relationships among variables.

Correlations and covariances with **proc corr:** To get correlations and covariances the basic set up is

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

The option "cov" after "proc corr" indicates you want covariances as well as correlations and can be omitted if you just want correlations. The "var" and "with" subcommands can each have multiple variable names on them. You are given the correlations of each entry on the "var" list with each entry on the "with" list. If you want correlations among all possible pairs of variables you can simply list them after "var" and omit the "with" line altogether. However if you have many variables it can sometimes be useful to restrict the set of comparisons.

Fitting a simple linear regression with **proc reg** or **proc glm**: SAS has several procedures for fitting statistical models, some of which are special cases of others. The **proc reg** procedure is the most similar to the **reg** command in STATA but is less flexible than **proc glm**, especially in terms of handling categorical variables. That isn't important for this assignment but will become so later. The commands are very similar:

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

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

Getting graphics in SAS: In general SAS is not the world's best graphics program (or at least I am not good at getting nice plots out of it!) For this assignment you can add the line

plot sedation*cortisol;

to **proc reg** between the model and run commands and it will include a scatterplot with the estimated regression line drawn in and some basic summary statistics.