Homework Assignment 4

Due Date: Wednesday, October 26th

Note: There are 6 problems on this assignment. The first 4 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 5-6 to receive full credit. The assignment is due Wednesday, October 26th. 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 **noon** with no penalty. Note the difference in the grace period. I want to get the assignment in early so I can post the solutions before everyone leaves for the day to study for the midterm.

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) CIs and PIs:

(a) Explain in words the difference between a **confidence interval** for $\mu_{Y|X}$, the average value of Y for a given X, and a **prediction interval** for a specific value of Y at a given X. Which interval should be wider and why? Give an example of when you might prefer to use each type of interval.

For the remaining parts of the problem use the sedation and cortisol data from warm-up problem 5 of homework 3. The data are repeated in the HW 4 files for your convenience.

(b) Use STATA or SAS to find 95% confidence intervals for b_0 and b_1 and explain carefully what they tell you.

(c) Is there a significant relationship between cortisol level and sedation score? Explain based (i) on your answer to part (b) and (ii) by performing two appropriate hypothesis tests. You should set up the hypotheses and test statistics by hand and base your conclusions on appropriate p-values from the printout.

(d) Now use STATA or SAS to find a CI and PI for the sedation score corresponding to a blood cortisol level of 12 micrograms per deciliter and interpret each of your intervals. Explain how these intervals are different from the ones you computed in part (b).

(2) Interpreting A Multiple Regression Equation: A regression equation was found to be $\hat{Y} = 20 + 14X_1 - 7X_2$. Which of the following statements are correct? You should explain your reasoning in each case. If the statement is not correct, say how you would correct it or how you could check if the statement were true.

(a) A one unit increase in X_1 causes Y to increase by fourteen units.
(b) Variable Y is more highly correlated with X_1 than X_2 since the coefficient on X_1 is positive.
(c) If the value of X_2 is large enough, one can obtain negative predictions of Y.

(3) Condo Prices (WBCH 13.27): A real estate broker with engaged the services of a consultant to develop a multiple regression model to predict the sales price of condominiums, Y, in thousands of dollars from the area of the floor space, X_1, in hundreds of square feet and whether or not the condominium had access to a swimming pool, X_2. ($X_2 = 0$ with no pool access and $X_2 = 1$ with pool access.) Data for n=20 condos are shown below.

Price	Area	Pool	Price	Area	Pool
67.9	12.0	1	59.1	11.1	0
69.5	13.8	1	59.3	10.5	0
67.2	14.6	0	64.3	11.9	1
61.1	12.6	0	56.6	9.7	0
68.3	12.1	1	50.3	9.2	1
43.5	8.0	0	63.7	12.5	0
59.2	11.0	1	65.5	13.4	1
52.6	9.6	0	61.4	12.9	0
56.2	9.9	1	63.2	13.3	0
56.1	10.6	0	64.8	13.5	1

(a) Find the estimated multiple regression equation.

(b) Find the predicted price of a condo that is 1200 square feet and has a pool. (Be careful of your units!)

(c) What is the estimated influence of having a swimming pool on the sales price of a condo? (In other words, what is the interpretation of b_2?)

(d) Find a 95% confidence interval for β_2 and explain its meaning.

(e) Logically, would you expect having a pool to raise or lower the price of a condo? Set up and perform the appropriate hypothesis test to prove your point. Make sure you state your hypotheses, p-value, and conclusions.

(4) Salary Prediction: The manager of Statisticorp, a company based in the city of Los Seraphim, is wondering what factors determine employees' salaries. In the past, number of years spent with the company has been considered the most important predictor. However in recent times gender and age discrimination have become hot-button issues, so she is particularly interested in knowing whether the age and gender of an employee give additional information about salary. She takes a random sample of 20 employees. Let Y be salary (in thousands of dollars), X_1 the number of years with the company, X_2 age, and let X_3 be an indicator which is 1 if the employee is female, and 0 if the employee is male. A data set containing information for the employees is given in the table below. Use it to answer the following questions.

Salary	Years	Age	Gender	Salary	Years	Age	Gender
23	2	25	1	25	3	26	1
27	4	28	1	35	5	29	1
45	6	28	1	47	6	30	1
50	8	40	1	65	10	40	1
70	12	45	1	102	20	45	1
27	2	25	0	30	3	26	0
30	4	28	0	39	5	29	0
47	6	28	0	51	6	30	0
57	8	40	0	70	10	40	0
77	12	45	0	110	20	45	0

(a) Find the correlations among Y, X_1, and X_2. Do you think X_2 (age) is a good predictor of Y (salary)? Why or why not?

(b) Fit the multiple regression of Salary on Years, Gender and Age. Based on the STATA multiple regression printout perform an hypothesis test to determine whether β_2, the coefficient of age, is significantly different from 0. Clearly state the null and alternative hypotheses and the p-value. What does this test tell you about the usefulness of age as a predictor in this model? In light of your answer to (a), how could this have happened? What should you do to resolve the problem? Explain briefly.

(c) Test the null hypothesis $\beta_3 = 0$ versus the alternative $\beta_3 \neq 0$, giving the test statistic and the p-value. What does this say about the respective salaries of men and women?

(d) Suppose you had started with a theory that women make less money than men with the same qualifications and wanted to prove it. How would this have changed your hypotheses in part (c)? What would your new p-value and conclusions have been? Explain briefly.

(e) Perform an overall F test for this regression. Make sure you state 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.

(f) Based on the answers to the previous parts, what final model you would choose to use for this data set? Briefly justify your choice. Note that the model you choose does not need to be one you have actually fit!

Problems To Turn In

(5) Parenteral Nutrition Continued (Based on GS Problems 2.6 and 3.8):

This continues Problem 8 of Homework 3. Recall that 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. 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, in mg per 12 hours) and related it to four factors: dietary calcium (in mg/12 hours, dietary protein (in grams per day), urinary sodium (in meq per 12 hours), and glomerular filtration rate (in mL per minute) which is a measure of kidney function. Their data are given in the accompany file with the names urinarycalcium, dietarycalcium, gfr, urinarysodium and protein. You may use any printouts from homework 3 (available in the online solutions) without rerunning the corresponding analyses. Use $\alpha = .05$ for all parts of this problem.

(a) Give a careful real-world interpretation of the confidence intervals for β_0 and β_1 from the simple linear regression of urinary calcium on dietary calcium. Make sure your explanation includes the units and numerical values from the interval and also explains whether and why the intercept and slope are significantly different from 0.

(b) Note that the coefficient of dietary calcium in the model from part (a) is less than 1. Does this make real-world sense? Explain.

(c) Find a 95% confidence interval for the average urinary calcium level of people who are getting 200mg of dietary calcium per 12 hours using STATA or SAS and carefully interpret it.

(d) Find a 95% prediction interval for the urinary calcium level of Johnny Skims who is getting 200mg of dietary calcium per 12 hours using STATA or SAS and carefully interpret it. Explain in a sentence or two why your answer is different from that in part (c).

(e) Use STATA or SAS to find a 95% prediction interval for Johnny Cream who is getting 1000 mg of dietary calcium per 12 hours and explain in a sentence why the width of the interval is similar to or different from the width of the interval in part (d).

(f) Now fit a multiple linear regression model of urinary calcium on the four predictor variables using STATA or SAS. Is this model overall useful for predicting the variability in urinary calcium? Explain by performing an appropriate test. Write the null and alternative hypotheses mathematically and in words, get the test statistic and p-value from your regresison printout and explain your real world conclusions.

(g) Give a real-world interpretation of the confidence interval for β_1, the coefficient of the dietary calcium variable, in the multiple linear regression model. Is this interval the same as the one in part (a)? Explain carefully what this tells you.

(h) Is dietary protein a significant variable in this model? Explain by performing an appropriate hypothesis test. Carefully state the null and alternative hypotheses mathematically and in words, get the test statistic and p-value from your MLR printout and explain your real-world conclusions. Are the other predictors significant? Explain briefly. You do not need to write out the details of the tests.

(i) On homework 3 we found that all 4 predictors were significantly associated with urinary calcium level in individual simple linear regression models. They are not all significant in the multiple linear regression model. What does this tell you? (Your interpretations from part (g) and (h) may be helpful here.) Verify your conclusion by performing an appropriate set of calculations.

(j) Compare R^2, R^2_{adj} and RMSE for the multiple linear regression to the same values for the simple linear regression of urinary calcium on dietary calcium from homework 3. Do you think that the multiple linear regression is a substantial improvement over the simple linear regression? Discuss. We will learn how to test this formally later on.

(6) College Tuition

A researcher at US Views and World Seaports is conducting a study about tuition at American colleges and universities. So far, he has collected data from 20 schools about their tuition costs, Y (in thousands of dollars), their score on an independent rating scale, X_1 (in points out of 100), their size, X_2 (in thousands of undergraduates), and whether they are a public ($X_3 = 0$) or private ($X_3 = 1$) school. A printout for the multiple regression of Y on the three X variables is shown below. Use it to answer the questions on the following pages.

```
Regression Analysis


The regression equation is
Tuition = - 2.41 + 0.0967 Rating - 0.0192 Size + 16.9 Type

Predictor        Coef       StdErr          t     P>|t|
Constant      -2.4053       0.9257      -2.60     0.019
Rating        0.09671      0.01172       8.25     0.000
Size         -0.01923      0.01606      -1.20     0.249
Type          16.8581       0.3357      50.21     0.001

RMSE = 0.5869      R-Sq = 99.7%     R-Sq(adj) = 99.6%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3     1742.29      580.76   1686.01    0.000
Error             16        5.51        0.34
Total             19     1747.80

Predicted Values

     Fit  StDev Fit         95.0% CI             95.0% PI
  20.236      0.402   (  19.384,  21.088)  (  18.728,  21.744)
```

(a) Interpret b_0, b_1, and b_3 in terms of tuition costs, rating, and type or school.

(b) The author of the study wants to know whether the combination of rating, size, and type of school is OVERALL useful for predicting tuition costs. Explain your reasoning briefly. You do not need to write out the details of the test.

(c) There are two numbers that give, approximately, percentage of the variability in college tuition that is explained by the regression on ratings, size, and type. What are these two numbers? Which is more appropriate, in the multiple regression setting, for measuring how good the model is? Explain your choice.

(d) Does this model do a good job of **predicting** college tuition? Briefly explain your reasoning.

(e) Use the printout to compute a 95% confidence interval for β_2. Explain what the resulting confidence interval tells you about the usefulness of size as a predictor of college tuition.

(f) Does it appear from this model that private schools cost significantly more than public schools? Perform an appropriate hypothesis test, making sure to state the null and alternative hypotheses mathematically and in words with a justification of your choice, give the p-value and your real world conclusions. Use $\alpha = .05$

(g) A student is interested in attending the University of Southern North Dakota at Hoople, a private school which supposedly has 1000 students and a rating of 60 on the scoring system used in this study. Unfortunately, the student's parents are having a hard time finding out what the tuition at Southern North Dakota is. (This may have something to do with the fact that the school is completely fictitious! I will be impressed if you know where it comes from...) First verify the estimate of the tuition at Southern North Dakota given on the printout by plugging the appropriate values into the regression equation and then give a range of values in which you can be 95% certain that the true tuition lies. Explain why you chose the interval that you did.

(h) Optional Bonus: The value labeled StDevFit at the bottom of the printout is $s_{\hat{Y}_0}$ the standard deviation associated with the average Y at the specified X values. Explain how you could get from this and other values on the printout the standard deviation associated with an individual value of Y at that set of X values.

(i) Suppose you have a private school and a public school of the same size. According to your best estimate from this model how much more highly does the public school have to be rated than the private school to have the same tuition?

(j) Donald Dimwit, the President of Southern North Dakota, notices that the coefficient of the Size variable is negative, implying that smaller schools have higher tuition (maybe because of their small class sizes). He decides to cut the enrollment at SND to 100 students. Explain (at least) 2 things that are wrong with his reasoning.

STATA and SAS Commands

To do this assignment you need to be able to fit simple and multiple linear regression models in STATA and SAS and get them to give you confidence and prediction intervals for a given set of X values. Adam has created several lab files to teach you these and other regresison commands. I repeat the basics here for completeness.

(1) STATA COMMANDS:

The **regress** command: This command is used to fit not only a simple linear regression model but also a multiple regression model. Type **reg** (or regress) followed by the name of your response variable (Y) followed by the names of all your predictor variables. For example in warm-up problem 3 we would type
reg price area pool.

After using the regress command you can use the **lincom** or **adjust** commands to get confidence intervals for β_0, β_1 and any linear combination thereof, including CIs and PIs for predicted values. Specifically:

lincom _cons gives the estimated value and CI for β_0.
lincom cortisol gives the estimated value and CI for β_1.
lincom _cons + 10.5*cortisol gives the predicted value and confidence interval for the average value of sedation when the cortisol level is 10.5
adjust cortisol = 10.5, se ci gives the predicted value, standard error and CI for the average value of sedation when the cortisol level is 10.5.

adjust cortisol = 10.5, stdf ci is the same except it gives the standard error and range of the corresponding prediction interval for the sedation of an individual with cortisol level 10.5. (stdf stands for standard error of the forecast meaning a new prediction at that value.)

(2) SAS COMMANDS:

For fitting multiple regression models **proc glm** is the best choice in SAS as it handles all sorts of data types easily. For the pool example, telling SAS that pool is a categorical variable (which isn't strictly necessary) we would type

```
proc glm data = work.hw4;
class pool;
model price = area pool;
run;
```

There are several options you can add to the basic **proc glm** command to get confidence intervals for the β's and CIs and PI's for Y. Specifically

model price = area pool/clb clm cli;

will include confidence intervals for the $\beta's$ (clb), confidence intervals for the average Y associated with each set of X's in the data set (clm) and prediction intervals for individual Y's for each set of X's in the data set (cli). To get predictions for new values of the X's you just include rows in your data set for the combinations of X's you want with the Y value missing. I have included an extra row with a 1200 foot condo with a pool but no price in the data set so you can try this out.