1 3-D representation of Multiple Regression

- Problem 3.2, G&S page 104 *[don't get fancy .. rough sketch is good enough]*

2 More Multiple Regression

- Problem 3.3, G&S page 105

If you didn't keep an electronic version of these data from your earlier homework, and if you are a slow typist, the raw data and the permanent sas dataset can be found on the course website under "datasets". If using sas, download the sas permanent dataset (gst1_1.sd2) file from the website into your "sasuser" directory. You can change data values "live" directly inside INSIGHT -- and, if you wish, save them as to a new sas permanent dataset rather than writing over the original.

3 Extracting estimates of regression parameters (i.e. statistics) from software output

- Problem 3.4, G&S page 105

4 Using Multiple Linear Regression: with more information, can one better predict Old Faithful?

Use the data mentioned in Moore and McCabe's story "Is Old Faithful Faithful?" reached by following the link to "story and newer (1995) data [M&M]" on the class web page.

To save time, I have taken their datafile and created a sas permanent dataset, which is available from the course web page. The sas program that created it is also available; the program has the raw data embedded in it, in case you wish to import or cut and paste the raw data into another statistical or spreadsheet package.

a Using these new data, answer the 4 questions at the end of the M&M story

b In an earlier analysis, you fitted a simple linear model of interval vs. duration to a sample of data collected in 1978 -- before the 1983 earthquake. Test whether the

(i) mean interval has changed since the earthquake [déjà vu, t-test, 607]

(ii) coefficient in the simple linear model has changed since then [déjà vu, pp 27-28]

5 Meaning of coefficients in multiple regression ; using coefficients with "x" differences that are not 1 unit in magnitude

Refer to "electricity bills" data on course web page *[questions are at end of documentation file]*

Do parts 1-9 [i.e. skip the last two parts, 10 and 11]

Remember that CI[$\hat{\beta}_x \times \Delta x$] = CI[$\hat{\beta}_x$] $\times \Delta x$, You might think of the analogy: if you were pretty (95%) sure that a person's height was between 65" and 67", then you should be equally sure that the person was between 65×2.54 cm and 67×2.54 cm.

*The CI calculations requested in parts 8 and 9 are not possible with INSIGHT. So instead, for these last parts use PROC REG from the SAS Program Editor window. Type (or paste) the following program statements into the Program Editor window -- **include the (important!) semicolons**! Then click on the "run" icon (icon is of a person running) to submit the job. The log window will tell you if you have been successful, and if you are the output will be in the output window.*

5 **Meaning of regression coefficients ; using coefficients with "x" differences that are not 1 unit in magnitude** ***continued...***

```
options ls=75 ps=55;
proc reg data=sasuser.electric;
model bill = income persons area;
  output out = stats
  predicted = bill_hat
  l95m      = l95mean
  u95m      = u95mean
  l95       = l95indiv
  u95       = u95indiv;
proc print data=stats;

run;
```

[you can copy the above text from an acrobat pdf file -- use the tool indicated by a T *with a marquee beside it to select the text, then copy it and paste it into the sas editor window.]*

Parts 8 and 9 ask for predictions (and CI's) for a specific covariate pattern than may not be present in the actual dataset. There is a formula for the CI (it is an extension to multiple x's of the formula in the slides for ch 2) -- but it involves matrices. For now, simply use the output for "neighbouring" covariate patterns in the dataset.

By the way, do you get any sense of how the x values affect the CI's?

6 **OPTIONAL** -- prelude to next class

Is there gain without pain? Is more better? *["bodyfat data (Small)" on course web page]*

a Reproduce the calculations in the documentation

b Which single-variable model has the highest r-square? (for now, don't consider powers or other functions of a variable)

c Does the addition of the better of the other 2 remaining variables add to it?

d Examine the larger list of predictor variables available in the other (larger) dataset on bodyfat. Do you think that this larger set of variables will yield more accurate predictions than the 3 available in the dataset above? Why/why not?

No calculations required for part d. And the fact that the numbers of observations also differ too is a separate matter!