1 **Fit a regression using an existing SAS permanent data set in SAS INSIGHT**

- Do problem 2.3, G&S page 46-7 [p50 in 2nd ed] (download .sd2 file from website into your "sasuser" directory). Test both the slope and the correlation against zero. Comment.

2 In each of the following, which is more appropriate: **confidence interval for mean response OR prediction interval for a new observation?** (Q's i-iii are from Neter et al, page 88, rest are "homegrown")

i What will be the humidity level in this greenhouse tomorrow when we set the temperature level at 31°C?

ii How much do families whose disposable income is $23,500 spend, on the average, for meals away from home?

iii How many kilowatt-hours of electricity will be consumed next month by commercial and industrial users in Montreal, given that the index of business activity for the area remains at its present level?

iv question "d" (how you would answer a judge) at end of Blood Alcohol and Eye Movements *[see documentation under datasets on web page]*

v Questions c and d in "Predicting when babies first sleep through the night article and q's" *[under Resources on web page]* don't do any calculations, just identify the "target" of each question]

vi Analysis of Rates of Fatal Crashes on rural interstate highways in New Mexico in the 5 years 1982-1986 (55 mph limit) and in 1987 (65 mph limit). The authors fitted a regression line to the 5 "55-mph" years, **calculated the "projected" value for 1987** and the expected range of variation around this fitted value, and determined where, relative to this range of possible values, the observed value in 1987 lies.

Full story and analysis in "Intro to multi-variable analysis (incl. SHARPER and FAIRER e.g.'s) from 1995" under Resources.

3 **Using Simple Linear Regression for Prediction: How Faithful was Old Faithful?**

(Which Interval ??? CI for mean response ??? prediction interval for a new observation???)

Use the (now outdated) data collected in1978, to be found under "Datasets" on the web page. . Read the "documentation" that describes these older data. (for a more colourful description, follow the link "newer data on Old Faithful [M&M]" on the class web page to obtain the story "Is Old Faithful faithful?" We will use these newer, and more multivariate, data in the homework to go with the next class.

Exclude the observations taken in 1979. In INSIGHT you can do this via the menus

Edit -> **Observations** -> **Find** YEAR>1 then **Edit** -> **Observations** -> **Exclude in Calculations** and **Hide in Graphs**

Alternatively, find the observations you want (YEAR=0), then **Extract** them (via menu found by clicking on black triangle sign at top left corner of spreadsheet)

How Faithful was Old Faithful? ... *Q's adapted from M&M*

Explore the relationship between duration of the current eruption and the length of time between the current eruption and the next eruption.

a Create a scatterplot of the duration of the current eruption vs. the length of time between the current eruption and the next eruption. Do the data exhibit a linear relationship? Explain.

b Use simple linear regression to quantify the linear relationship observed in Part a.

i State the regression equation and the likely size of prediction error associated with that equation.

ii How much of the variability in the time between eruptions can be explained by the duration of the previous eruption?

c Based on the regression equation found in Part b, predict the amount of time between the current eruption and the next eruption, given that the duration of the current eruption is 4 minutes.

*[To save manual claculations, in the window showing the results of the fit, use the **Curves->Confidence Curves** menu to obtain the appropriate 95% band (? mean ?prediction)-- and then (in the column that gets added to the dataset) find an observation that has a duration of 4 minutes]*

Why are the band limits not very "bow-shaped" in this example [in the examples below, they will be!] ?

d Find the standard deviation of the time between eruptions in the data set and the standard deviation of the residulas from the regression in part b).

[if you wish, you can get the overall (unconditional) standard deviation by fitting the "null" model i.e., don't click a variable into the "X" panel in the FIT dialog box].

Explain why one of these standard deviations is smaller than the other.

e Does the scatterplot of the residuals versus predicted values for the regression in Part b suggest any problems with the regression?

*Note that in INSIGHT, the predicted values and the residuals from the Fit are added to the dataset, and the residuals are plotted vs the predicted values by default. You could also use the **Scatter Plot (Y X)** to plot the residuals against the X values.*

Comment on the pattern of the Residual Normal QQ Plot available under the **Graphs** menu on the window with the results of the FIT *(look ahead to pp 128-129, 127-133 in 2nd Ed.)*

4 **Using a software package to obtain a confidence interval for mean response OR a prediction interval for a new observation?** ***Two smaller datasets.***

Use INSIGHT or equivalent (and a bit of manual interpolation or extrapolation) to obtain appropriate intervals for parts iv (*alcohol*) and vi (*speed limit*) examples in Q2.

Speed limit data : Once you have brought the speed limit data into SAS INSIGHT, click on the square symbol at the beginning of the row corresponding to 1987 and turn off the "Include in Calculations" flag. Fit a line to the 5 remaining datapoints; note (in the column that gets added to the dataset) the predicted value for 1987. In the window showing the results of the fit, use **Curves=>Confidence Curve** to obtain the appropriate 95% band (? mean ?prediction) for this estimate. Extend the curve by hand to see where the actual 1987 datapoint falls relative to the expected band.

[try as I might, i can't get INSIGHT to fit the line using the 5 points but have it extend the bands to all 6.. it is possible if we use PROC REG from the PROGRAM EDITOR]

We could also have tested this new data point via a multiple regression, as is described in "Intro to multi-variable analysis (incl. SHARPER and FAIRER e.g.'s) from 1995" under Resources. but that is for a subsequent class.

Alcohol data : Of interest is the band at an alcohol level of 80. Once you have brought the alcohol data into SAS INSIGHT, fit a line to the 12 datapoints; note (in the column that gets added to the dataset) the predicted values for two points that "straddle" 80. In the window showing the results of the fit, use **Curves=>Confidence Curve** to obtain the appropriate 95% band (? mean ?prediction) for this estimate. Interpolate by hand/eye to obtain the appropriate band for an alcohol level of 80.

Here is the SAS code to produce bands using PROC REG from the PROGRAM EDITOR.

```
proc reg data = sasuser.alcohol;
  model decrease = alcohol;
  output out = bands
                  predicted = fitted
                  l95       = l95indiv
                  u95       = u95indiv
                  l95M      = l95mean
                  u95m      = u95mean;
proc print round data=bands;
run;
```

5 **Non-linear relationships**

(after reading "heat exchange in gray seals", G&S pp 42-45, pp 45-48 in 2nd Ed.)

Download the raw data or sas permanent file for the ***caries/fluoride*** dataset in course 697. (same username and password as for course 678). These data are from before the "added fluoride" era i.e. these are levels of naturally occurring fluoride.

Graphically examine the relationship. What message do the data have for public health officials who wish to select the amount of fluoride to add to the water supply?

Suggest a few non-linear forms for the relationship, and fit them [in INSIGHT you can use **Edit=>Variables** to add transformations of the "Y" (or the "X") variable to the dataset]. Don't expect to get a perfect fit -- there are a number of unmeasured influential variables. One is temperature (e.g., children in warmer communities drink more water -- a factor which municipalities now consider when setting fluoride levels).

6 **Non-linear relationships** - vocabulary data for 1 child

(from KKMN problem 5_12 under datasets on main web page ; KKMN say that the data are from M. E. Smith, "An Investigation of the Development of the Sentence and the Extent of Vocabulary in Young Children," Studies in Child Welfare (University of Iowa) 3(5) (1926))

a Does a linear growth model fit the data? [In their q's, KKMN ask if the point "0.0 years, 0 words" is on the line of vocabulary growth; they also ask that one add this point and re-fit!]

b Suggest and fit a better model.

c Then, from the FIT window in INSIGHT, fit the highest possible degree polynomial to the data -- and explain why the vocabulary "dips" between ages 5 and 6!! [KKMN ask: if the data included observations through age 30 years, what would the extrapolated curve look like?]

Remember to ask me about the investigator who fitted the highest possible degree polynomial to daily WBC's (White Blood Counts)! And ask yourself HOW the investigator determined that at age 5 the child has a vocabulary of EXACTLY 2072 words.

d KKMN ask : "The data appear to be from one child. If this is true, what assumption of the least-squares approach is most likely violated, and why?

They are probably getting at the INDEPENDENCE of the "error" components in the model. If these were 15 observations from 15 DIFFERENT children, and one had a good model for the expected (MEAN) values, this clearly would not be a problem.

But the only purpose here is to describe the progression of THIS ONE child, and to use the model for interpolation FOR THIS ONE CHILD.

My Question: Do you think that the serial "errors" or residuals would be correlated? i.e., is it likely that observations on one side of the curve would tend to be immediately followed by observations on the same side of the curve (positive serial correlation) or by ones on the the opposite side (negative serial correlation)? Why?