36-401 Homework #3 Due 9/21/00

Download the dataset "televisions.dat" and the descriptor file "televisions.txt" from the class web page. Examine both files before proceeding (e.g. with cat or lpr or xemacs).

The dataset is in a format not directly accessible to Splus because some countries have invalid characters in their names, i.e. spaces, commas and parentheses. (Spaces are especially bad because, by default, spaces indicate the beginning of the next data column.) Edit the file to make all country names one word (embedded periods are OK).

Read the data into a data.frame called "tele". Be sure to use the optional read.table() parameter na.strings="*" because missing data are coded as a asterisks in this file. Note that read.table() drops the "Country" as a data column and uses it as the row names because it is the first non-numeric variable that has unique values. If you need the country names, e.g. as the third argument to identify(), you can use row.names(tele).

1. **10 pts** Divide the plot area into a 2x2 grid with space at the top for an overall title. Plot (Y on X) FLE on PPT, MLE on PPT, FLE on PPP and MLE on PPP using appropriate labels.

   Create two new columns in the data frame using

   ```
   tele$logPPP_log(tele$PPP)
   tele$logPPT_log(tele$PPT)
   ```

   Make a similar set of 4 plots using the log versions of the predictors. Describe what you can learn from the 8 graphs including suitability of linear regression for further analysis of these data.

2. **10 pts** Perform a linear regression of male life expectancy on log of people per physician. Make a scatter/box plot of Y on X (using scatbox), and include the regression line and the identity of the several largest outliers (largest vertical distances from regression line to point). Also include scatter/box plots of residual on X, residual on fitted values, and a quantile normal plot of the residuals. Turn in the plots. In one sentence describe the information in the boxplots of the first plot. Do the same for the second plot, and state which boxplot tells us something about whether the standard regression assumptions are met.

   To my eye, there is only mild evidence in these plots of violation of those assumptions that can be checked using such plots. State exactly what deviations you see from the standard assumptions.

3. **10 pts** Repeat the procedure of the last problem, but using "physicians per person" as the covariate. State exactly what deviations you see from the standard assumptions, and whether you judge each deviation as mild, moderate or severe.

— **Turn Over** —

4. **10 pts** Explain (interpret) all of the numbers in the "Coefficient" block of the summary() of the lm() result for problem 2. Explain the meaning of the residual standard error.

5. **20 pts** Assume these life expectancy data were presented to you, the statistician, by an official in the United Nations. This official wants to present these data to support extra funding to send more physicians to small underdeveloped countries.

   Write a report for this official based on the analyses above. Introduce the dataset and the problem in the first paragraph (5 pts). Describe the key univariate features of the LE and PPP variables in the second paragraph (5 pts). Describe the results of the scatter-plots and bivariate regressions in the third paragraph (5 pts). Summarize and conclude in the last paragraph (5 pts). In your conclusion, include a comment on cause and effect (with reference to the television scatter plot). Also include a comment on extension of these results to e.g. Palau.

   DO use a formal writing style for your report. Do NOT use variable codes, e.g. PPP, in your report. Do NOT use any S-plus terminology such as "data.frame" or "lm". You should include a small number of graphs to which the text must explicitly refer. All graphs must be nicely labeled (e.g. use xlab=, ylab=, mtext(), etc. as needed). Grading is based on clarity of expression, and choice of what is important to include. Inclusion of lots of extraneous material reduces your score. *Do not include any Splus code or direct Splus output.*