STAT 400 Data Analysis October 20, 2003

Today: Test feedback
Intro to Ch 5: Hypervariate Data
Reminder of Assignment Due

Test Answers:

1. In the "Ubi" exercise, what assumption needs to be satisfied in order to undo the confounding of species and species plot location?

The adjustment for a location effect cannot be done on each plot in isolation since the 5 measurements are not enough to estimate this reliably. However, if all the plots can be assumed to be affected by the location effect by the same linear effect, then we can pool the information from all ten plots to make the correction. This amounts to modeling the location effect using a single plane over the area of the ten plots.

2. In the "Mercedes" exercise, how would you choose "alpha" in the loess fit, for detecting the seasonal trend?

The data was collected approximately weekly for four years – about 200 data points. To estimate a seasonal effect, we need to include as much data as possible relating to the season, but not lumping seasons together. Perhaps a window of about three months is appropriate. This would correspond to about 13 weeks, and 13/200 = 0.065. This would be a good starting value for alpha.

3. The "histo" exercise in density estimation and the fitting of the surface to the "galaxy" data both used kernel estimation. What is this common technique called "kernel estimation"?

Kernel estimation uses a weighting function to accumulate contributions from data according to the distance from a grid point. The weights decrease as the distance increases. The contributions can be pseudofrequency or values of a dependent variable.

4. In the analysis of the "Bimbo Bakery" data, the actual and simulated sales distributions are compared using overlaid ECDF plots. What advantage does this approach have (in the context of the exercise assigned) over examination of a Q-Q plot?

The ECDF overlays made it easy to adjust the guess to move the simulated sales ECDF closer to the actual sales ECDF. The shape of the Q-Q plot also provides this information but the actual values of mean and SD are harder to see.

5. What characteristic of a sequence of coplots for trivariate data indicates the presence of interaction? Explain the general case but with reference to the coplot on page 189.

If the slope of the response of Z to changes in X is different for different values of Y,

then X and Y are said to interact in the effect on Z. The NOX on p 189 depends on CR with a slope near zero for high values of ER, but is strongly positive for low values of ER. So CR and ER interact in determining NOX.

6. How does "brushing" improve the utility of matrix plots?

It is hard to see interactions from matrix plots – but it it can be done with brushing. If one chooses a brush that includes a limited range of X values, in the plot of Z against X, then the linked points in the plot of Z against Y show how Z depends on Y. Changing the limited range of X then would show if the plot of Z against Y changes slope. One gets information similar to a coplot (but without loess usually).

7. What does one look for in a spread-location (s-l) plot?

Monotone Spread. We want to know if we can treat the variability in Y|X as constant over the range of X.

8. Compare and contrast wireframe plots and contour plots for the depiction of trivariate data, when there is one dependent variable of interest and two independent variables.

Wireframe: easier to explain, shows general nature of curve, and is amenable to various views. Difficult to relate to axis scales. Hard to see whole curve at once.
Contour: a bit more difficult to explain. Ambiguity of pits and peaks. Scale can be read in fair detail. Not good a depicting the nature of a surface where it is fairly flat. Can see whole surface at once.

9. Explain how the cycle plot on page 164 is constructed, and what it shows about the CO_2 data?

The 32 years of data is shown by month – each 32 Jan values is one series, etc. These 32 values appear to be shown by a loess curve (otherwise one would expect more variability) for each month. What the cycle plot shows, apart from the obvious seasonal effect, is that the amplitude of the cycle is increasing over time (since spring values are getting higher and fall values lower).

10. Why is computation of residuals to a loess fit a non-trivial procedure?

The complication is that residuals are computed at the data values, while the loess is computed on a grid. So interpolation of the loess to the data positions is needed in determining the $\hat{Y}$ that goes with the observed Y in computing the residual Y-$\hat{Y}$.

Intro to Hypervariate Data – Ch 5

Air Pollution Data: O3, R, T, W

How many Coplots?
What is lost in adding fourth variable (p 279)?
Fifth variable possible?

Reminder:
Assignment 5: Analyze galaxy data using coplots. (code, graphs, and words necessary in report.) How does this compare with the level plot analysis for revealing important information about the data? This will not be due until Wednesday, October 22, 2003. (5 days after midterm). It would be a good idea to learn how to do this in R or Splus - it is possible in MINITAB but more difficult.