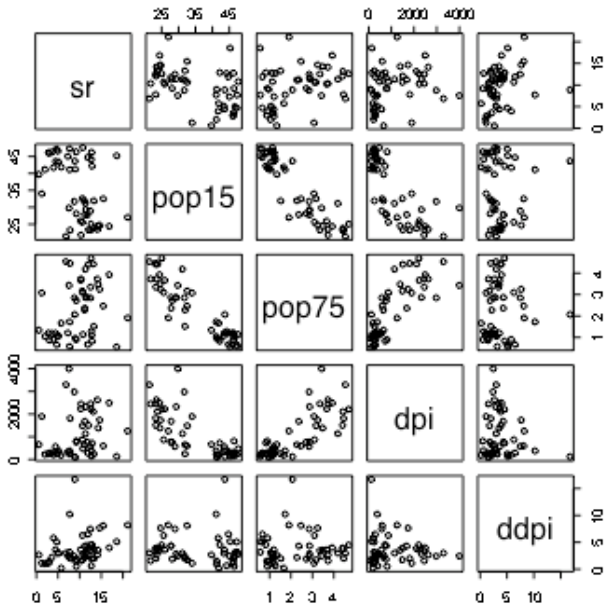
STAT 400 Data Analysis Oct. 28, 2005

Feedback on the test.

1. Based on Fig 4.1 (p 180) alone, what levels of Tensile Strength and Hardness would tend to provide a relatively small amount of abrasion loss? Provide numerical guidelines in your answer.

Hardness above 70 Shore and Tensile Strength above 180 kg/cm^2. (In the case of Tensile Strength, one needed to look at the means in vertical strips, or think of where the smooth curve would pass).

2. Here is a matrix plot of the LifeCycleSavings data described in the notes for Oct 24. What features of the data set observable in this matrix plot would influence your subsequent analysis strategies? (Just a couple of obvious things!)



There are separated clusters of pop15 values. And ddpi has a clear outlier. (One would certainly look at the labels of these countries when seeking an explanation for the variability observed.)

3. In ch 4, some coplots show data points in them and some do not. Why?

The coplots without data points are coplots of fitted surfaces. These fitted surfaces are estimated at grid points, not at the observed data values. Since all the data is used to estimate the fitted surface, one would have to include all the data in each panel, which would just be confusing. If one tried to use the local data only, it is not clear

what data to use, since there is no interval of the conditioned values to select the data subset to plot in a coplot panel.

4. The parametric analysis of the ethanol data (the one I posted and discussed in class) results in a statistically significant interaction term with a negative coefficient. Explain why this would be expected from an examination of Fig 4.6.

The (NOX vs CR given ER) slope decreases as ER increases, so the response due to the main effect of CR and ER needs to be reduced, and a negative interaction does this.

5. Figure 4.13 shows the relationship of NOX to ER for various CR values. One of the CR values used is 10.2, and yet there are no engines with this value of CR. Should we "crop" Fig 4.13 to only include the panels for CR values that are present in the data set?

The surface fit has a smooth response over the 5 CR values. The intermediate values in the grid are not represented in the data but are well estimated by the surrounding data. So there is no concern that the fit at intermediate values of CR are unsupported (as there would be if these were extrapolated values). Do not crop.

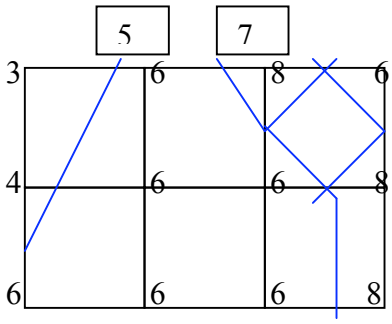
6. Why does Cleveland recommend cropping of the Abrasion Loss surface for the coplots on p 203?

There are portions of the rectangular grid that contain no data, and these portions are in areas where the surface would have to be extrapolated if used. So the portion of the curve that is to be believed is the portion that is well-occupied by data. The surface is estimated using all the data but the coplot that is shown is only for portions of the surface where data is present.

7. How does Figure 4.50 (p 239) relate to Figure 4.55 (p 247)

If one shaded in the regions between the contour lines of Fig 4.50, one would get shapes like the ones shown in separate panels in Fig 4.55. If you collapse the plots in Fig. 4.55 onto one plot, you get a contour plot (as long as you can still see the boundaries of the regions from 4.55). (Note that the regions in each panel of Fig 4.55 are not curves, they are areas – the "curves" have varying width).

8. For the grid values shown here, draw the estimated contour lines at levels 5 and 7. (Copy the grid to your answer sheet – freehand is OK).



Lines are approx due to graphics limitations.

9. The galaxy data is portrayed with a level plot, while the ethanol and rubber data sets use the coplot. Why were these methods chosen for these particular applications? (i.e. what is different about the galaxy data set compared to the other two. Note that I am **not** asking for the difference between a coplot and a level plot.)

The galaxy data analysis had no interest in the differential effect of the two directions, since the aim was to describe the spatial (jointly described by the two coordinates) velocity pattern. So no conditioning was called for in this application. In ethanol and rubber applications, the conditioning was of great interest since the separate effect of each predictor was sought, as well as the interaction if present.

10. Cleveland talks about Direct Manipulation in almost every chapter. What does he mean by this term? Which of Fig 4.50 (p 239) and Fig 4.58 (p 251) would be enhanced by Direct Manipulation?

Direct Manipulation is examination of data displays while varying parameters in real time. It would have been useful with the wireframe plots (varying the direction of observation) but not as much for the contour plots (since the only thing that might be varied would be the smoothing constant and that can be assessed easily with a small number of ordinary plots.