STAT 400 Data Analysis November 5, 2003

Today: Assignment 6 feedback
1-1 Plots
Bootstrap

Recall Assignment 6: Analyze the Ozone data using regression methods, with an aim of determining a good model for predicting Ozone Concentration from Solar Radiation, Temperature and Wind Speed. Use coplots to analyze the residuals. Provide a verbal summary of your results. (Note that the Cleveland text gives some information about this data – but your report is on your own analysis of the data.) Your report should include your final regression output and the coplots.

Some comments arising from this:

0. Objective – Remember to state the objective of your analysis – it guides what you do.

1. Verbal report – "There exists and interaction …." is not a verbal summary of results. Jargon is for teaching and learning, not for summarizing results. Your analysis may not show exactly what Cleveland's does, although it should be close if you have a good model, but in this case a verbal summary might be:

"Ozone increases with increasing radiation but with a decreasing marginal effect. This increase (of ozone with increasing radiation) is accentuated with increasing temperature. Ozone decreases with increasing wind speed but with a decreasing marginal effect. This relationship appears almost unrelated to the temperature. "

Cleveland's pp 286-289 gives the words and pictures on which this is based.

2. Selecting a good subset of variables in a regression model. It is not a sin to include independent variables in a model that are not statistically significant! If the sample size is much larger than the number of variables considered, then the beta estimates should be small for a variable that is not doing any valid predicting, and so its inclusion/exclusion will have little effect. If the variable is a plausible predictor, and if prediction is the aim, inclusion of nonsignificant variables is a good idea, since a variable can have an effect without it being significant.

Stepwise Procedures are very misleading – correlated variables can mask each others' effect – there is a preempting that happens. Stepwise procedures are better at picking variables that should be included than they are at picking variables that should be excluded. In the exercise, the R-T interaction was eliminated if the W-T interaction was entered in, but this did not lead to the best model or interpretation of results. There is no automatic way to pick the best subset in a regression – you need to use the context to guide the analysis.

3. The p-value is not a measure of importance of the alternative hypothesis. In

regression, small value of p does not say predictor is important - The p-value measures the strength of evidence against the coefficient being 0. Any p-value can occur with any variable no matter how important, since this p-value is also a function of the sample size.

4. Transformations of dependent variables are usually done to symmetrize the dependent variable so that least squares is a reasonable criterion (i.e. minimizing the sum of squared residuals in units of the dependent variable). But transformations of independent variables are also useful – it can often simplify the predictive relationship, by linearizing it, or removing an interaction.

5. Reports of analyses need to be selective – words to say in general terms what you did, a graph to illustrate a key finding that you came across in your exploratory work, an equation or graph to show the ultimate model you derived, and some jargon-free words to describe the outcome of your analysis. Some things to avoid in a report: anova tables, a series of regression outputs, dotplots of all the variables, 4 significant digits in p-values or coefficient estimates!

6. Why use coplots for the residual analysis? Because if there is interaction revealed by these coplots, the model is not adequate. The coplot analysis of residuals is much better than the plot of Y-Yhat against Yhat.

1-1 Plots: Methods for graphical representation of the data without preliminary "signal" extraction or "noise" reduction. The advantage is that eyeball signal identification is more flexible than parametric signal extraction. The disadvantage is that sometimes even the eyeball cannot see the signal for the noise (woods for the trees).

Star plots
Chernoff Faces
Profile Plots
Augmented Scatter Plot (and GIS)

Bootstrap