Stat 482 Homework Set #12

A hospital surgical unit is interested in predicting survival in patients undergoing a particular type of liver operation. A sample of $n = 108$ patients is available. The data is available from Table 9.1 on the course website. From each patient record, the following information is gathered, listed in order of column:

x1 = blood clotting score,
x2 = prognostic index,
x3 = enzyme test,
x4 = liver test,
x5 = age,
x6 = sex (0=male,1=female),
x7 = alcohol use moderate (0=no,1=yes),
x8 = alcohol use heavy (0=no,1=yes),
y = survival time (in days),
log.y = ln(y) (transformation to achieve normality)

1. Describe the goal of the discrepancy function approach to model selection.
2. How is the best model defined?
3. What are the two sources of model error?

4. Plot the Cp statistic against the model dimension.
5. Plot the Cp statistic against the candidate models.
6. Which variables are included in the selected model?

7. Test the selected model against its best competitor having fewer parameters. (Compute the test statistic and the p-value.)
8. How does discrepancy based model selection compare to p-value based selection?