Statistik 2 2008, 8.4.08

Övning 2a

Hämta följande datafil: http://www.abo.fi/fak/mnf/mate/jc/statistik2/bostonh.dat
Filen innehåller sammanlagt 506 observationer på 14 variabler som karaktäriserar olika distrikt i Boston-området i USA. Variablerna i datamaterialet betraktas i detalj på s. 44-52 i Härdle & Simar (2003).
Importera datat till R commander och rita låddiagram (boxplot), samt spridningsdiagram (scatterplot matrix) för samtliga variabler. Studera hur de observerade värdena fördelar sig. Anpassa en regressionsmodell, där variabel 14 (median av husprisen i ett distrikt) är den beroende variabeln och de övriga variablerna regressorer (Statistics-Fit models-Linear regression). Vilka variabler verkar koppling till huspriset enligt p-värden? Jämför dessa resultat med spridningsdiagrammen. Undersöka modellanpassningen med hjälp av ett outlier-test (Models-Numerical diagnostics-Bonferroni...) och grafisk diagnostik (Models-Graphs, sedan Basic...,Residual...,Effect...). Verkar modellen lämplig?
Anpassa en ny modell, där endast de variabler som hade signifikanta regressionskoefficienter tas med som regressorer. Notera att varje modell anges ett namn och att man kan betrakta resultat för en valfri modell genom att välja Models-Select active model. Använd igen diagnostiken för att avgöra modellens lämplighet. Har situation ändrats från den tidigare?

Övning 2b

Fortsätt med samma datamaterial som ovan och gör de variabeltransformationer som föreslås i Härdle & Simar (2003) s. 51 (använd Data-Manage...-Compute new variable, och ge lämpliga namn åt variablerna). Notera att en log-transformation kan behändigt utföras för flera variabler på en gång, genom att skriva följande kommando i Script-fönstret (här antas det att datasetet namngetts 'husen' vid importering):

```
husen[,c(1,3,5,6,8,9,10,14)]<-log(husen[,c(1,3,5,6,8,9,10,14)])
```

och sedan måla över kommandot med musen och klicka på Submit-knappen. R commander transformerar alltså nu alla indikerade värden (alla raderna i kolumnerna 1,3,5,6,8,9,10,14) i matrisen 'husen' genom att ersätta dem med logaritmen av det ursprungliga värdet.
Gör sedan samma analyser som ovan för de nya variablerna. Kan den enklare modellen (endast de ursprungligen signifikanta regressorerna är inkluderade) nu anses vara lämplig för datat?