TMA4267 - Linear Statistical Models 2009
Week 11: Model Validation and Diagnostics

Theoretical exercises

Rencher & Schaalje:

- 9.1 a)-d)

Computer exercises

Outliers and influential observations

Consider a model on the form

$$y_i = \boldsymbol{\beta}^T \mathbf{x}_i + e_i, \quad i = 1, \ldots, n, \tag{1}$$

where $e_1, \ldots, e_n$ iid and $e_i \sim N(0, \sigma^2), \forall i$, y_i is the response and $\mathbf{x}_i$ covariates. For the data on 35 Scottish hill races in the data frame `hills` in `library(MASS)`, let `time` be the predictor.

a) Compute $\hat{y}_{,i}$, $\hat{\epsilon}_i$, h_{ii}, r_i, t_i and D_i.

   Are there any outliers or potentially influential observations? Compute PRESS and compare to SSE.

   If you find outliers or very influential points in the data set, try without these observations and check the model for outliers or potentially influential observations again. Compare the results.

R-hints:

```
model1 = lm(time~.,data=hills)

par(mfrow=c(2,2))
plot(model1)
```

```
par(mfrow=c(2,3))
plot(fitted(model1))
plot(residuals(model1))
plot(lm.influence(model1)$hat)
plot(stdres(model1))
plot(studres(model1))
plot(cooks.distance(model1))

model2 = update(model1, subset = -c(1,14))
```

Use `help(command)` in R to find out what the commands do.

Checking model assumptions

Some of the model assumptions can be evaluated by calculating the residuals and plotting or otherwise analyzing them. The following plots can be constructed to test the validity of the assumptions:

- Residuals against the explanatory variables in the model. The residuals should have no relation to these variables (look for possible non-linear relations) and the spread of the residuals should be the same over the whole range.
- Residuals against the fitted values.
- A time series plot of the residuals, that is, plotting the residuals as a function of time.
- Residuals against the preceding residual.
- A normal probability plot of the residuals to test normality. The points should lie along a straight line.

There should not be any noticeable pattern to the data in all but the last plot.

b) Check the model assumptions.