STATS 3001 / STATS 4104 / STATS 7054
Statistical Modelling III
Practical 3 - Assumption checking

null

Week 5

GOAL

The purpose of this practical is to explore the application of influence diagnostics in multiple regression. The learning objectives of the practical are

- To demonstrate the practical application of influence diagnostics;
- To demonstrate the correspondence between leverage and the distribution of the x-values;
- To verify the built-in function for calculating leverage;
- To demonstrate the correspondence between Cook's Distance and changes in the parameter estimates.

DATA

The file `hills.csv` contains the record times in 1984 for 35 Scottish hills races.

The dataset contains the following variables:

- `dist`: The total distance in miles
- `climb`: The total climb in feet
- `time`: The record time in minutes

Interest is focused on modelling `time` using `dist` and `climb` as predictors.

STEPS

1. Read in the data
2. Obtain a scatter plot of `dist` vs `climb`. Identify the points that you believe will have high leverage.
3. Calculate the leverage values from the design matrix for the model `{~ climb + dist` using the matrix expression given in lectures.
   Note the command `diag(H)` will extract the diagonal values of a square matrix `H`.
   Identify the points with leverage greater than $2p/n$. Check whether the points you identified on the scatter plot do have high leverage.

4. Calculate the leverage values using the built-in `hatvalues()` function i n R and check that they agree with those calculated from the formula.
   (Note: R also provides functions, `cooks.distance()`, `rstudent()` and `rstandard()` to calculate Cook's distance, the studentized residuals and the standardized residuals, respectively.)
5. Obtain the usual sequence of diagnostic plots from R.
6. Based on the residuals vs leverage plot, identify the most influential point.
7. Identify the point with the largest residual, the point with the highest leverage and the point with the highest Cook's distance, and comment.
8. Fit the same model to the data with the most influential point removed.

9. Calculate Cook's distance for the most influential point according to the formula

$$D_i^2 = \frac{\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(i)}\right)^T \left[\hat{Var}(\hat{\boldsymbol{\beta}})\right]^{-1} \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(i)}\right)}{p}$$

$$= \frac{\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(i)}\right)^T (X^T X) \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(i)}\right)}{p s_e^2}.$$

Check that your value agrees with that produced by the built-in `cooks.distance` function.