Multivariate Statistics: Exercise 5

November 7, 2018

Robust covariance estimation and regression diagnostics:

Take the data `fat` from the `library(UsingR)`. For detailed information see `help(fat)`. In the following, we will use the variable `body.fat` as response variable for linear regression, and all remaining variables except `case`, `body.fat.siri`, and `density` as explanatory variables.

1. Investigate if there are leverage points by using

   (a) classical diagnostics based on the diagonal elements of the hat matrix,

   (b) robust diagnostics based on robust Mahalanobis distances. Use the MCD estimator (`covMcd()` from the package `robustbase`) for this purpose.

   What do you conclude?

2. Now apply linear regression by not using the first 100 observations, with the

   (a) the least-squares estimator (`lm()`) and the robust MM-estimator (`lmrob()` from `library(robustbase)`). Interpret the results of `summary()` and `plot()`. Is robustness recommendable?

   (b) Compute the Cook distances from the least-squares solution as

   $$D_i = \frac{r_i^2}{s^2} \frac{h_{ii}}{p \cdot (1 - h_{ii})^2},$$

   for each used observation with index $i = 1, \ldots, n$. Here, r_i is the i-th residual, s^2 the residual variance, p the number of estimated parameters, and h_{ii} the i-th diagonal element of the hat matrix – all referring to the classical linear model. A potential influential point could refer to a Cook distance larger than $4/n$. Compare graphically the Cook distances with the weights resulting from the robust regression approach. Do both types of diagnostics lead to the same conclusions?

   (c) Use the models to predict the responses of the remaining first 100 observations. Compare the classical and robust predictions graphically and numerically using an appropriate measure of prediction accuracy.

Principal component analysis:

1. Load the data `Auto` from the package `ISLR`, and use only the variables `mpg, displacement, horsepower, weight, acceleration`.

(a) Compute the principal components from the standardized data (`princomp()` using the option `cor=TRUE`). Apply `plot` and `summary` on the result object and interpret the results.

Interpret the directions of the principal components (stored as `$loadings` in the result object).

The `$scores` in the result object are the data values projected on the principal components. Show pairwise plots of these scores of the most important principal components, and visualize the variable `origin` by different choices of symbols or colors in the plots. Which multivariate relations are visible?

(b) Perform the above analysis with a robustly estimated covariance matrix using `covMcd()` from the `library(robustbase)`. The MCD result object can be used in `princomp()` via the argument `covmat`. How do the results change compared to (a)?

Save your (successful) R code together with short documentations and interpretations of results in a text file (= R script file), named as *Matrikelnummer_5.R* (no word document, no plots). Submit this file to Exercise 5 of our tuwel course (deadline November 6).