Leverage and Mahalanobis Distance

Contents

1 Relationship between Leverage and Mahalanobis Distance in Predictor Space

Given a data matrix X, the 'full' $n \times p$ data matrix including an intercept has the form $[1X]$ where 1 is a $n \times 1$ column of 1s.

The vector of leverages in a regression in which the predictor variables are given by X plus an intercept term is the diagonal of the projection matrix (often known as the 'hat matrix') onto the linear space spanned by the unit vector and the columns of X, $\mathcal{L}(1, X)$.

Under the assumption that $[1X]$ is of full column rank:

$$\begin{aligned} H &= [1X] \left([1X]'[1X]\right)^{-1} [1X]' \\ &= [1X] \begin{pmatrix} n & 1'X \\ X'1 & X'X \end{pmatrix}^{-1} [1X]' \end{aligned}$$

As a projection matrix, H is invariant under location scale transformations of the data

matrix X. Replacing X with X_c in which each column is centered so that $X_c'1 = 0$

$$\begin{aligned} H &= [1X_c]\left([1X_c]'[1X_c]\right)^{-1}[1X_c]' \\ &= [1X_c]\begin{pmatrix} n & 0 \\ 0 & X_c'X_c \end{pmatrix}^{-1}[1X_c]' \\ &= \frac{1}{n}11' + X_c(X_c'X_c)^{-1}X_c' \\ &= \frac{1}{n}11' + \frac{1}{n}X_c\Sigma_X^{-1}X_c' \end{aligned}$$

The diagonal elements of $X_c\Sigma_X^{-1}X_c'$ are the squares of Mahalanobis distances of individual observations standardized by the maximum likelihood estimate of variance using division by n.

Thus, using the diagonal elements:

$$h_{ii} = \frac{1}{n}(1 + Z_i^2)$$

This is a result that is valid for linear multiple regression for any number of predictor variables with an intercept term. In simple regression, $p = 1$ and Z_i, with an appropriate

sign, is simply the the 'Z-score' for ith predictor observation.

2 Some consequences

1. It is easily shown that $\sum_i h_{ii} = 1 + p$:

$$\begin{aligned}
\sum_i h_{ii} &= \text{trace}(H) \\
&= \text{trace}\left([1X]\left([1X]'[1X]\right)^{-1}[1X]'\right) \\
&= \text{trace}\left(\left([1X]'[1X]\right)^{-1}[1X]'[1X]\right) \quad \text{since trace}(AB) = \text{trace}(BA) \\
&= \text{trace}\, I_{(p+1)\times(p+1)} \\
&= p + 1
\end{aligned}$$

so that:

$$\sum_{i=1}^{n} Z_i^2 = np$$

and

$$\overline{Z_i^2} = p$$

which concords with expectation since Z_i^2 has a distribution that is approximately χ_p^2.

2. We also know that $\frac{1}{n} \leq h_{ii} \leq 1$ so that

$$0 \leq Z_i^2 \leq n-1$$

3. If X is a $n \times p$ matrix of predictor variables in a least-squares regression, then Z_i is the Mahalanobis distance of the ith case in **predictor** space.
4. The 'residual-leverage' plot, which is the fourth plot produced by the 'plot' command in R applied to 'lm' objects, plots $h_{ii} = \frac{1+Z_i^2}{n}$ on the horizontal axis.
5. Mahalanobis distance is the generalization of the univariate Z-score to all dimensions. Strictly speaking, in one dimension, Mahalanobis distance is the *absolute value* of the Z-score.
6. The ellipse in predictor space:

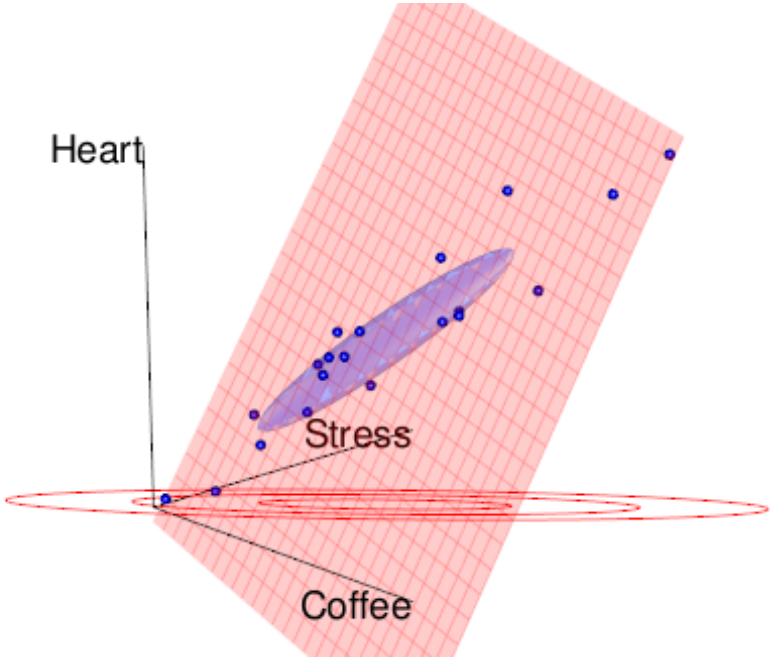
$$\mathcal{E}_r = \left\{ x \in \mathbb{R}^p : (x-\bar{x})' \hat{\Sigma}_X^{-1} (x-\bar{x}) = r^2 \right\}$$

contains the points (if any) with leverage equal to $\frac{1}{n}(1+r^2)$

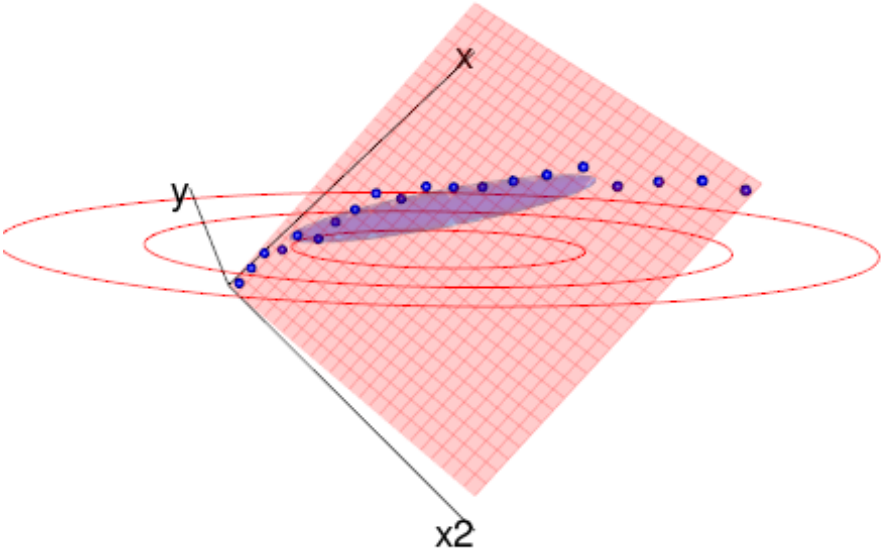
7. The concept works for *linear regression in βs* even if not linear in the predictors.

To compute Mahalanobis distance, you can write your own function, or you can consider the 'mahalanobis' function in base R.

For illustration, the following plot shows the predictor data ellipses of radii 1, 2 and 3, corresponding to Mahalanobis distances of 1, 2, and 3, and leverages of 0.1, 0.25, and 0.5 since $n = 20$.


Heart
Stress
Coffee

The next plot also shows the predictor data ellipses of radii 1, 2 and 3, corresponding to Mahalanobis distances of 1, 2, and 3, and leverages of 0.1, 0.25, and 0.5 since $n = 20$ but in a quadratic regression. Note that although the **predictors** are not normally distributed, the ellipse captures the first and second moments of the predictors, and the least-squares linear regression only depends on the first and second moments.


x
y
x2