Exercise on df

Degrees of freedom in regression models

This exercise covers topics that are discussed in Ch. 5 and in Ch. 6.2 in the textbook.

Many of the regression methods that we are discussing are linear in y, that is the fitted value $\hat{f}(x)$ is a linear combination of the observed y_i. Let $\{x_i, y_i, i = 1, \ldots, N\}$ be the training data where $x_i \in R^p$. Let $\hat{f} = (\hat{f}(x_1), \ldots, \hat{f}(x_N))^T$ be the fitted values in the training points. Linearity in y then corresponds to the existence of a matrix S such that $\hat{f} = Sy$.

a) Consider first ordinary linear regression based on the model $y = X\beta + \epsilon$. Find S in this case, and show that $\mathrm{trace}(S) = p$.

In general the matrix S will depend on some complexity parameter λ and we will then write S_λ which we call the *smoother matrix*. We define the *effective degrees of freedom* for a regression model that is linear in y by

$$\mathrm{df}_\lambda = \mathrm{trace}(S_\lambda)$$

b) Argue why this is a reasonable definition for the ordinary linear regression model.

c) Show that also k-nearest neighbor, the Nadaraya-Watson estimate and local linear regression are linear in y.

d) For k-nearest neighbor, show that $\mathrm{df}_\lambda = N/k$. Discuss this result. *Hint:* Identify the diagonal elements of S_λ.