Tutorial 7

1. Suppose we use model

$$Y = \begin{cases} a_0 + b_0\mathbf{x}_1 + c_0\mathbf{x}_2 + \varepsilon_0, & \text{if } \mathbf{x}_1 + \mathbf{x}_2 < 0, \\ a_1 + b_1\mathbf{x}_1 + c_1\mathbf{x}_2 + \varepsilon_1, & \text{if } \mathbf{x}_1 + \mathbf{x}_2 \geq 0. \end{cases}$$

to fit data $(\mathbf{x}_{i1}, \mathbf{x}_{i2}, Y_i), i = 1, 2, ..., n$. Write the procedure to calculate the (delete-one-out) CV value.

2. For model

$$Y = 4\mathbf{x}_1\mathbf{x}_2 + \varepsilon,$$

find two PPR models for it. (In other word, the representations of PPR model is not unique.)

3. For **data B** (the first 3 columns are independent variables, the last one response variable). Find the best model among

Linear regression I:	$Y = a + b\mathbf{x}_1 + \varepsilon$
Linear regression I:	$Y = a + b\mathbf{x}_1 + c\mathbf{x}_2 + \varepsilon$
Linear regression I:	$Y = a + b\mathbf{x}_1 + c\mathbf{x}_2 + d\mathbf{x}_3 + \varepsilon$
PPRA:	$Y = \phi_1(\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \alpha_3\mathbf{x}_3) + \varepsilon$
PPRB:	$Y = \phi_1(\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \alpha_3\mathbf{x}_3) + \phi_2(\beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3) + \varepsilon$
PPRC:	$Y = \phi_1(\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \alpha_3\mathbf{x}_3) + \phi_2(\beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3)$ $+\phi_3(\gamma_1\mathbf{x}_1 + \gamma_2\mathbf{x}_2 + \gamma_3\mathbf{x}_3) + \varepsilon$
PPRD:	$Y = \phi_1(\alpha_1\mathbf{x}_1 + \alpha_3\mathbf{x}_3) + \varepsilon$
PPRE:	$Y = \phi_1(\alpha_1\mathbf{x}_1 + \alpha_3\mathbf{x}_3) + \phi_2(\beta_1\mathbf{x}_1 + \beta_3\mathbf{x}_3) + \varepsilon$
PPRF:	$Y = \phi_1(\alpha_1\mathbf{x}_1 + \alpha_3\mathbf{x}_3) + \phi_2(\beta_1\mathbf{x}_1 + \beta_3\mathbf{x}_3)$ $+\phi_3(\gamma_1\mathbf{x}_1 + \gamma_3\mathbf{x}_3) + \varepsilon$

4. Consider the Swiss banknotes again. For **training data** apply CART to built the CART tree. Based on this tree, check the **validation data**. How many banknotes are misclassified? Apply PPR with (component 1) to the same training and validation sets by assigning 1 to genuine and 0 to counterfeit banknotes, if a banknotes' predicted value is greater than 0.5 then it is classified as genuine, otherwise counterfeit. how many banknotes are misclassified?

5. In the Prostate data, predictors are lcavol, lweight, age, lbph, svi, lcp, gleason, pgg45. The response variable is lpsa. For **training data** apply CART to built the classification tree. Based on this tree, predict the lpsa in **validation data**.