Project V

Parametric/Nonparametric Nonlinear Regression

(Due on 11/08/2022 Tuesday by 11:59pm)

Instructions: hile discussion with classmates are permitted and encouraged, please try to work on the project independently and direct your questions to me. Please present computer output in the main report only when necessary. Always remember to interpret your analysis results within the application context using concise and clear language, with focus on interesting findings. Additional supporting computer output can be moved to an Appendix. Also, include your R/Python codes in another Appendix for reproducibility purposes, though you don't have to do so if you use Markdown. Please name your file as YourLastName-Project05.PDF.

We consider a data set `jaws.txt`, which is concerned about the association between jaw bone length ($y =$ `bone`) and age in deer ($x =$ `age`). We are going try out several parametric/nonparametric nonlinear regression models in this low-dimensional ($p = 1$) setting.

1. Bring in the data $\mathcal{D}$ and make a scatterplot of `bone` vs. `age`. Optionally, add a linear fit and a nonlinear fit (with, e.g., `lowess` or `loess`) and inspect their discrepancy. Does their association look linear?

2. Data partitioning.

   (a) Randomly partition the data $\mathcal{D}$ into the training set $\mathcal{D}_1$ and the test set $\mathcal{D}_2$ with a ratio of approximately 2:1 on the sample size.

   (b) To prevent extrapolation when it comes to prediction, the range of `age` in the test set $\mathcal{D}_2$ should not exceed that in the training set $\mathcal{D}_1$. Find the (two) observations with minimum and maximum `age` in data $\mathcal{D}$ and force them to go to the training set $\mathcal{D}_1$ if they are not in $\mathcal{D}_1$.

3. First consider parametric nonlinear models.

   (a) Fit an asymptotic exponential model of the following form

   $$y = \beta_1 - \beta_2 e^{-\beta_3 x} + \varepsilon \tag{1}$$

   with the training set $\mathcal{D}_1$. Provide a summary of the fitted model and interpret the results.

   (b) To test $H_0 : \beta_1 = \beta_2$ in Model (1), fit the reduced model, i.e., the model by plugging in the condition under H_0 and use the `anova` function. Also, compare two `nls` models with AIC/BIC. Then conclude on which model is better.

   (c) Based on the better model in 2(b), add the fitted curve to the scatterplot.

   (d) Apply the better model in 2(b) to the test set $\mathcal{D}_2$. Plot the observed y_i values in $\mathcal{D}_2$ versus their predicted values $\hat{y}_i$, together with the reference line $y = x$, to check if the prediction seems reasonable. And computer the prediction mean square error (MSE)

   $$MSE = \frac{1}{|\mathcal{D}_2|} \sum_{i \in \mathcal{D}_2} (y_i - \hat{y}_i)^2.$$

4. Next consider local regression methods.

   (a) On basis of $\mathcal{D}_1$, obtain a KNN regression model with your choice of K. Plot the fitted curve together with the scatterplot of the data. Apply the fitted model to $\mathcal{D}_2$. Plot the observed and predicted response values with reference line $y = x$ and obtain the prediction MSE.

   (b) Apply kernel regression to obtain a nonlinear fit. State your choice of the kernel function and the choice of your bandwidth. Explain how you decide on the choices of kernel and bandwidth. Apply the fitted kernel regression model to the test data $\mathcal{D}_2$. Plot the observed and predicted response values with reference line $y = x$ and obtain the prediction MSE.

   (c) Apply local (cubic) polynomial regression to the training data $\mathcal{D}_1$. Again, state your choice of the kernel function and the bandwidth used. Apply the local cubic regression model to the test data $\mathcal{D}_2$. Plot the observed and predicted response values with reference line $y = x$ and obtain the prediction MSE.

5. Finally, regression/smoothing splines are applied.

   (a) Apply regression splines (e.g., natural cubic splines) to model the training data $\mathcal{D}_1$. Plot the resultant curve. Then use the fitted model to predict the test data $\mathcal{D}_2$. Plot the observed and predicted response values and obtain the prediction MSE on $\mathcal{D}_2$.

   (b) Apply smoothing splines to $\mathcal{D}_1$. Always specify the choice of your kernel function and comment on how you determine the tuning parameter. Add the resultant curve to the scatterplot. Then apply the fitted model to the test data $\mathcal{D}_2$. Plot the observed and predicted response values and obtain the prediction MSE.

6. Tabulate all the prediction MSE measures. Which methods give favorable results?