Name:.....

Data Mining and Visualization

Assignment 2

Fitting a line by splines

Assignments constitute part of the examination and must be handed in time. You are asked to hand in the solutions during a week following the week on which the assignment has been discussed in classes. You submit an electronic copy of your work to CANVAS.

Problem 1 – Splines and Bootstrap In the lecture, bootstrapping is discussed in the context of fitting splines in a non-linear regression problem. The bootstrapping of residuals has been suggested although no formal or even intuitive arguments were given to support the claim that this kind of bootstrap will actually work in this case. The following discussion is suppose to provide some rationale for a choice of bootstrap for the regression case. We need only a heuristic understanding of splines in context of regression. Please, refer to the lecture notes on splines for more details.

1. For the sake of our discussion, please, sketch on paper an example of data that would be suitable for applying a spline method of regression fitting. Describe the coordinates on the graph, marking which of them corresponds to the independent (explanatory) variable and which to the dependent (response or outcome) variable. Give a hypothetical real life situation in which your data are corresponding to actual data.
2. For the sketched data, propose the knot points. We consider cubic splines. For your choice of the knots sketch your best guess of the spline fit to the data. (By a guess we mean a visual choice of a regression spline fit without making any numerical calculations.) Mark on the graph the residuals for the proposed fit.
3. In reality, the following two scenarios can corresponding to this type of the data.
   - Values of the explanatory variable are non-random (fixed) like, for example, in equal spacing of abscissae. In such a case the only randomness can be attributed to the response variable.
   - Values of the explanatory variable are random so that they can be also considered as a random sample.

   Having these scenarios in mind discuss in which situation the following resampling methods would constitute a valuable bootstrap method of mining your data:
   - Resampling randomly only values of the response variable (Y).
   - Resampling randomly matched pairs of the response and explanatory variables (X, Y).
   - Resampling residuals as described in the lecture.

   Provide one/two sentence argument for your choices.

Problem 2 – Linear Regression Here we review some basic notation of the regression model. We have response variables $\mathbf{Y} = (Y_1, \ldots, Y_n)$ and a matrix of regressor (explanatory variables) $\mathbf{X} = [x_{ij}]_{i=1,j=1}^{n,r}$. The assumed model is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Here $\boldsymbol{\beta}$ is an unknown vector of the parameters and $\boldsymbol{\epsilon}$ is a random noise. The goal is to estimate $\boldsymbol{\beta}$ given the set of observations $(\mathbf{Y}, \mathbf{X})$.

In the least square approach, one estimates $\boldsymbol{\beta}$ by minimizing the Eucledean distance between the observations and the linear model

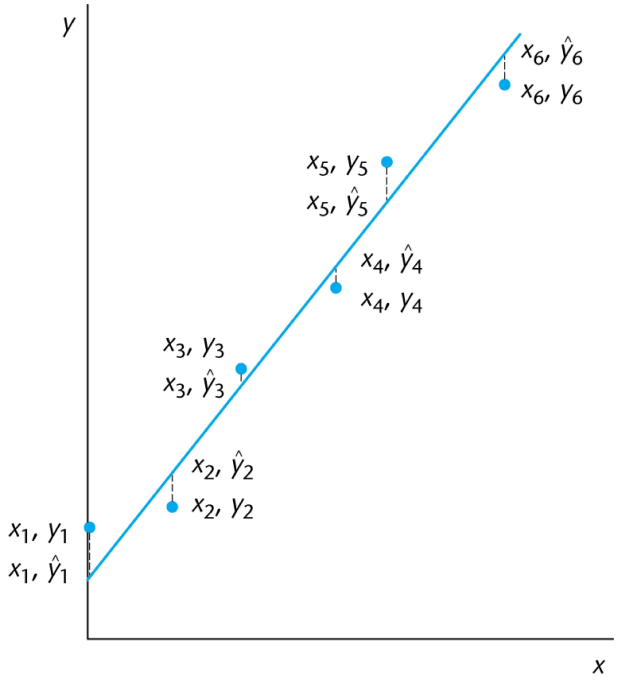
$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

In other words, we want to find a point $\hat{\boldsymbol{\beta}}$ such that $\mathbf{X}\hat{\boldsymbol{\beta}}$ is the closet point to the observations $\mathbf{y}$ among all $\mathbf{X}\boldsymbol{\beta}$'s.

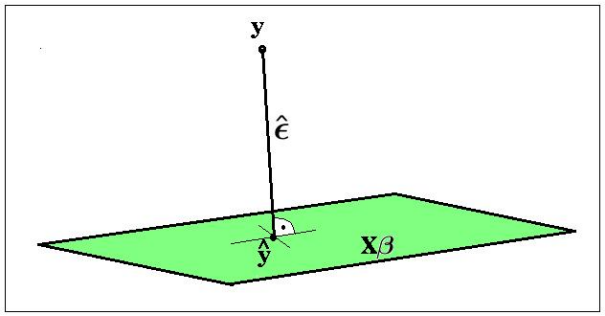
It follows from the principles of linear algebra that $\mathbf{X}\hat{\boldsymbol{\beta}}$ is the orthogonal projection $\hat{\mathbf{y}}$ of the observation vector $\mathbf{y}$ to the space of spanned by the columns of the design matrix $\mathbf{X}$. This projection can be expressed as the matrix $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, so that

$$\hat{\mathbf{y}} = \mathbf{P}\mathbf{y}$$
$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

The vector of residuals $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to the prediction $\hat{\mathbf{y}}$.



Schematic representation of the regression fit



Sixteen observations on the viscosity of a polymer (y) and two process variables – reaction temperature (x_1) and catalyst feed rate (x_2) are shown below:

$$\mathbf{X} = \begin{bmatrix} 1 & 80 & 8 \\ 1 & 93 & 9 \\ 1 & 100 & 10 \\ 1 & 82 & 12 \\ 1 & 90 & 11 \\ 1 & 99 & 8 \\ 1 & 81 & 8 \\ 1 & 96 & 10 \\ 1 & 94 & 12 \\ 1 & 93 & 11 \\ 1 & 97 & 13 \\ 1 & 95 & 11 \\ 1 & 100 & 8 \\ 1 & 85 & 12 \\ 1 & 86 & 9 \\ 1 & 87 & 12 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} 2256 \\ 2340 \\ 2426 \\ 2293 \\ 2330 \\ 2368 \\ 2250 \\ 2409 \\ 2364 \\ 2379 \\ 2440 \\ 2364 \\ 2404 \\ 2317 \\ 2309 \\ 2328 \end{bmatrix}$$

Compute the following

- Projection matrix $\mathbf{P}$.
- The least square fit $\hat{\boldsymbol{\beta}}$.
- Residual to the least square fit.
- Find the prediction for the temperature 80 and catalyst feed rate 14.

In your computation you can assist yourself with a computer program but all important steps of computations have to be shown.

Problem 3 – Simple Linear Regression In the simple linear regression, the model for the data is assumed to be $Y_i = a + bx_i + \epsilon_i$, $i = 1, \ldots, n$, with parameters a, b being real numbers and ϵ_i's are independent with $N(0, \sigma^2)$ distribution. This is a special case of $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

- Identify $\mathbf{X}$, $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$ for this special case.

- In a laboratory containing polarographic equipment six samples of dust were taken at various distances from the polarograph and the mercury content of each sample was determined. The following results were obtained:

Distance from polarograph,m	`1.4,3.8,7.5,10.2,11.7,15.0`
Mercury concentration, ng/g	`2.4,2.5,1.3,1.3,0.7,1.2`

  Identify, $\mathbf{Y}$, $\mathbf{X}$, that correspond to the simple linear regression for the above data, compute $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, and find the least square estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ using

  $$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

- Plot the data and the regression fit.

Problem 4 – Fitting non-linear function by linear regression Consider an interval $[0, 1]$ and the following values of the response to certain values of x given in pairs of (x, y)'s:

$$(0.2, 5), (0.3, 3), (0.6, 6), (0.7, 2), (0.9, 7).$$

Answer the following questions

- If you would like to approximate the dependence between y and x by a constant function, how would you approach to the problem. What would be the least square fit to the data?
- Explain the same for linear, quadratic, cubic fits. Show how the fits can be obtained by using the linear regression.
- Which of the cases discussed above corresponds to the classical simple regression?
- Sketch the obtained fits. Comment what you observe on the graphs.

Problem 5 – B-splines In the lecture we have introduced B-splines as a basis for cubic splines. They were defined as follows

- Assume $\xi_1, \ldots, \xi_K$ internal knots and two endpoints ξ_0 and ξ_{K+1}.
- Add three more knots that are equal to ξ_0 and additional three knots that are equal to ξ_{K+1} for the total of $K+8$ knots that from now are denoted by τ_i, $i = 1, \ldots, K+8$.
- Define recursively functions $B_{i,m}$, that are splines of the $m-1$th order of smoothness (0 smoothness is discontinuity), $i = 1, \ldots, K+8$, $m = 1, \ldots, 4$
- For the knots τ_i, $i = 1, \ldots, K+8$ we define $B_{i,m}$, $i = 1, \ldots, K+8$, $m = 1, \ldots, 4$
- The piecewise constant (0-smooth), $i = 1, \ldots, K+7$,

$$B_{i,1}(x) = \begin{cases} 1 & \text{if } \tau_i \leq x < \tau_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

- Higher order of smoothness , $i = 1, \ldots, K+8-m$,

$$B_{i,m}(x) = \frac{x - \tau_i}{\tau_{i+m-1} - \tau_i} B_{i,m-1}(x) + \frac{\tau_{i+m} - x}{\tau_{i+m} - \tau_{i+1}} B_{i+1,m-1}(x).$$

- $B_{i,4}$ are cubic order splines that constitutes basis for all cubic splines.

Consider an interval $[0, 1]$ and only one internal point $\xi_1 = 1/2$.

1. How many total knots for the cubic B-splines are considered according to the above definition?
2. Write down explicitly all knots τ_i =’s.
3. Write down explicitly all functions for each recursion step.
4. Sketch the obtained functions.
5. Which of them constitutes the basis for all cubic splines with the given initial internal knot $\xi_1 = 1/2$.
6. How the computations change if $\xi_1 = \xi$ is another internal point, i.e. it is not equal to $1/2$?
7. [1] Argue that cubic B-splines are still piecewise cubic polynomials on $[0, 1]$ with continuous second derivative, i.e. that they are indeed splines and that they are linearly independent, i.e. one cannot be expressed by a linear combination of others. Consequently, any piecewise cubic spline can be expressed by a linear combination of the B-splines.

[1] This is a more challenging problem for PhD and more mathematically inclined students, for the rest it is an extra (not required) problem.

Problem 6 – Smoothing splines The following is a simplified account of using smoothing splines to provide generalized additive fit to the regression problem with one predictor

$$y = \alpha + f(x) + \epsilon.$$

- A spline basis method that avoids any knot selection
- It is using the maximal set of knots (knot is located at each location that is given in the data)
- It is not overfitting because irregularity is penalized
- It is estimated by a linear function outside the range of predictors (smoothing on the boundaries)
- It minimizes the penalized residual sum of squares

$$PRSS(f, \lambda) = \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda \int f''(t)^2 dt$$

- $\lambda = 0$: any fit that interpolates data exactly.
- $\lambda = \infty$: the least square fit (second derivative is zero)
- We fit by the cubic splines with the maximal number of knots equal to the values of x's and

$$f(x) = \sum_{j=1}^{N+4} \gamma_j B_j(x) \tag{1}$$

  $B_j(x)$ are natural splines: linear outside the data range and the cubic polynomial inside of it.

- The solution has the form

$$\hat{\boldsymbol{\gamma}} = \left(\mathbf{B}^T\mathbf{B} + \lambda\boldsymbol{\Omega}_B\right)^{-1}\mathbf{B}^T\mathbf{y},$$

  where

$$\boldsymbol{\Omega}_B = \left[\int B_i''(t) B_j''(t)\ dt\right]$$

- To see this substitute (1) to the PRSS – it becomes a regular least squares problem

Discuss the following properties.

1. Explain why if $\lambda = 0$ the optimal fit will interpolate data exactly.
2. Explain why if $\lambda = \infty$ the optimal fit will be the regular least squares fit of the regression line.

3. Substitute (1) to the PRSS and express the latter using vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{N+4})$ and the matrices

$$\mathbf{B} = [B_{ij}] = [B_j(x_i)]\,,$$
$$\boldsymbol{\Omega}_B = \left[\int B_i''(t) B_j''(t)\; dt\right].$$

4.[2] Compute the derivative of the PRSS with respect to $\boldsymbol{\gamma}$ and check that it is equal to zero if

$$\boldsymbol{\gamma} = \left(\mathbf{B}^T\mathbf{B} + \lambda\boldsymbol{\Omega}_B\right)^{-1}\mathbf{B}^T\mathbf{y}.$$

[2]This is a more challenging problem for PhD and more mathematically inclined students, for the rest it is an extra (not required) problem.