Lecture2 Handout

Elizabeth Colantuoni

1/25/2021

I. Objectives

Upon completion of this session, you will be able to do the following:

- Use indicator variables, linear and cubic splines and their interactions to represent scientific questions
- Given a regression model formulated in terms of simple predictor varialbles, give a valid scientific interpretation of each term in the model

In this session, we will discussing the basic tools for building regression models: indicator variables; linear and cubic splines; and their interactions

II. Classical Multiple Linear Regression (MLR) Model

Consider a population of interest where we define an outcome of interest (Y) and a set of explanatory variables ($X_1, X_2, ..., X_p$).

For a sample of size n, we observe ($Y_i, X_{1i}, X_{2i}, ..., X_{pi}$) for each $i = 1, 2, ..., n$.

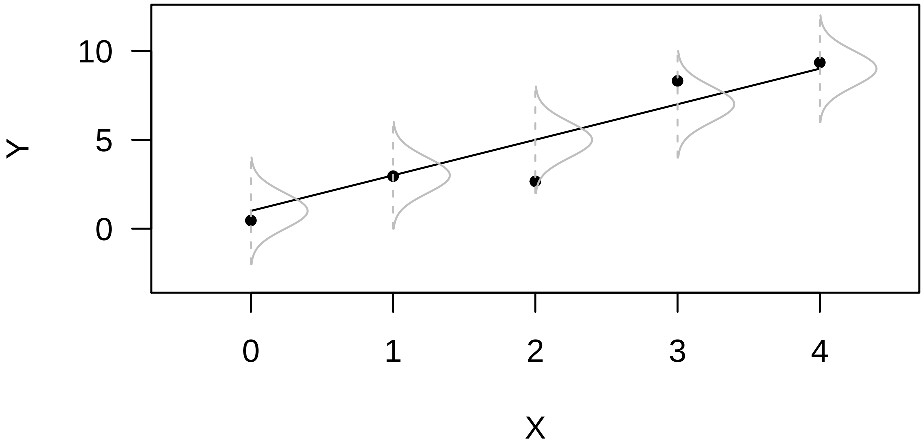
The classical multiple linear regression (MLR) model assumes:

$$Y_i = \mu_i + \epsilon_i$$

where

- μ_i is the systematic component and ϵ_i is the random component
- $\epsilon_i \sim N(0, \sigma^2), Cov(\epsilon_i, \epsilon_j) = 0$
- $\mu_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ...\beta_p X_{pi}$

A. Picture of model in p=1 case



B. What is linear in the MLR?

The systematic component is linear in the association parameters, i.e. β_i for $i = 1, 2, ..., p$.

Examples of linear models

- Simple linear regression: $Y = \beta_0 + \beta_1 X_1 + \epsilon$
- Quadratic model for X: $Y = \beta_0 + \beta_2 X_1 + \beta_3 X_1^2 + \epsilon$

Examples of non-linear models

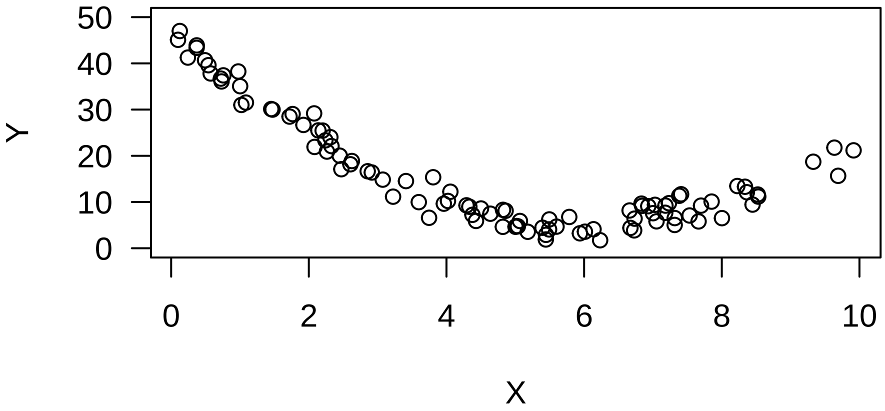
- Weibull growth model: $Y = \beta_0 + (\beta_1 - \beta_0) exp(-\beta_2 X^{\beta_3}) + \epsilon$
- Fourier model: $Y = \beta_0 + \beta_1 cos(X + \beta_2) + \beta_3 sin(X + \beta_2) + \epsilon$

II. Modeling non-linear trends in linear regression

There are three simple functions of a single X to allow for non-linear relationships using linear regression:

- Step function
- Linear spline
- Cubic spline

Throughout this section, we will consider the following toy example:



A. Step Function (Indicator or dummy variables)

Consider a step function for this data; although this would not be a great fit. Just an example!

To create a step function to describe the relationship between average Y and X in range (a,b), you would do the following:

- Partition the range into p intervals: $(a = c_0, c_1, c_2, \ldots, c_p = b)$; e.g. p=4 for quartiles or p=10 for deciles of X
- Define (p-1) indicator variables: $X_j = 1$ if $c_j \leq X < c_{j+1}$; 0 otherwise for $j = 1, ..., p-1$. NOTE: you are creating one less indicator variable than interval!
- Fit MLR with intercept: $Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_{p-1} X_{p-1i} + \epsilon_i$

B. Application of step function to Toy Example

1. Model set up:

Consider a step function with partition (0,3,6,10), i.e. p = 3.

We will need to define two indicator variables:

- $X_1 = 1$ if $3 \leq X < 6$; 0 otherwise
- $X_2 = 1$ if $6 \leq X < 10$; 0 otherwise

The model is: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$.

2. Model Interpretation:

Write down the interpretation of the following parameters and combinations of parameters. See Lecture 2 slides and recording for the solution!

- β_0
- $\beta_0 + \beta_1$

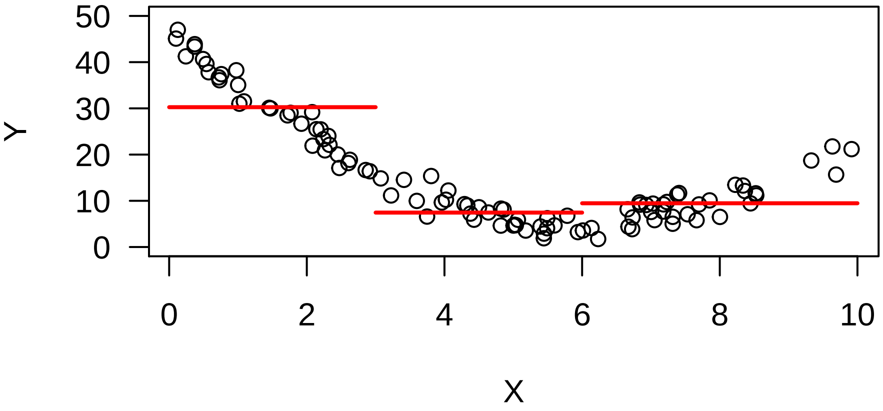
- β_1

- $\beta_0 + \beta_2$

- β_2

3. Model fit

A visualization of the model fit.



Comment: This approach can often be found in medical/epidemiological studies. Based on the figure above, what do you think are drawbacks to using this approach?

C. Linear Spline

Idea: linear spline (aka "broken arrow", "hockey stick"; "intervention" model) assumes that there is a linear relationship between Y and X with slope that can change at pre-specified locations called "knots".

The model specification is:

$$Y = \beta_0 + \beta_1 X + \beta_2 (X - c_1)^+ + \beta_3 (X - c_2)^+ + ...\beta_{k+1} (X - c_k)^+$$

where

- $u^+ = u$ if u $>$ 0 and 0 otherwise

Interpretation of coefficients:

- β_0, β_1 are the intercept and linear slope in the left-most interval
- β_j, $j \geq 2$ are the change in linear slope from before to after associated knot

D. Application of linear spline to toy example

Consider the following multiple linear regression of Y on: $X_1 = X$, $X_2 = (X-3)^+$, $X_3 = (X-6)^+$:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

1. Interpretation of the model and parameters

Please answer the following questions on your own and then see Lecture 2 slides and recording for the solution!

- What is the regression model for values of X < 3?
- What is the interpretation of β_0 and β_1
- What is the regression model for values of X between 3 and 6?
- What is the interpretation of $\beta_1 + \beta_2$ and β_2?
- What is the regression model for values of X between 6 and 10?
- What is the interpretation of $\beta_1 + \beta_2 + \beta_3$ and β_3?

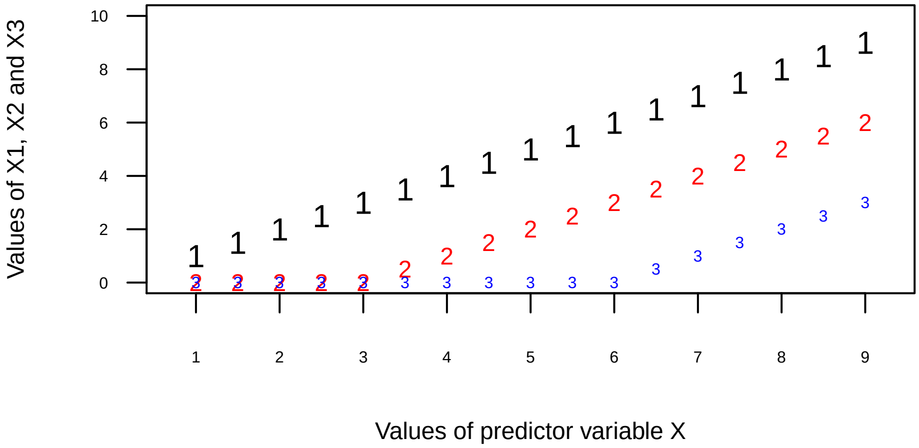
2. Design matrix for linear spline model

The "design matrix", also known as the "model matrix" or "regressor matrix" is often denoted by X. Each row of the matrix represents the values of the predictor/explanatory variables for a particular observation in the data. Each column represents one predictor/explanatory variable.

The "design matrix" for our example is given by:

intercept	x1	x2	x3
1	1	0	0
1	2	0	0
1	3	0	0
1	4	1	0
1	5	2	0
1	6	3	0
1	7	4	1
1	8	5	2
1	9	6	3

We can create an illustration of the design matrix. In the figure below, the numbers indicate the predictor variable X_1, X_2 and X_3 respectively.



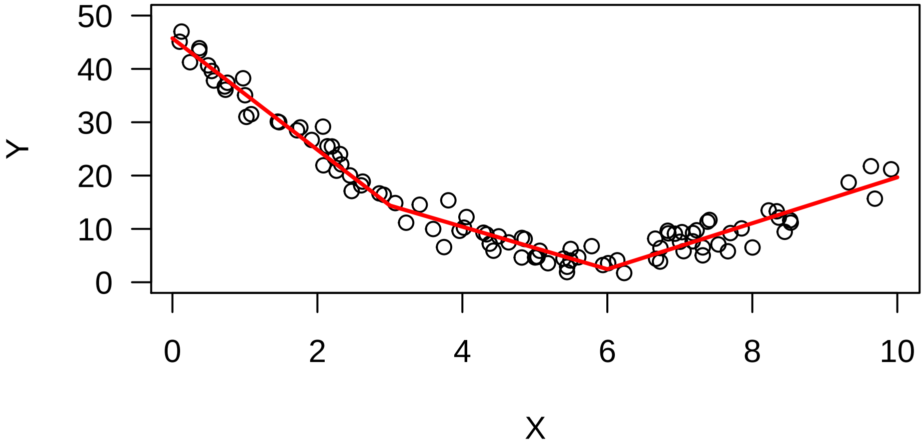
3. Estimation

Use the figure displaying the toy example data and estimate the parameters β_0, β_1, β_2 and β_3.

4. Assessment of quality of the model

I fit the multiple linear regression we specified above; obtained the "best" estimates of β_0, β_1, β_2 and β_3 based on our data (more to come on this).

Below is a plot of estimated average Y as a function of X.



Do you prefer the linear spline model to the step function approach? Why?

E. Cubic spline

Idea: linear splines are nice, but they have ugly elbows (discontinuities in their first derivative)! There are alternatives to the linear splines that make the functions join together smoothly at the boundaries and allow some more bend in each interval.

Express E(Y|X) as a "locally cubic" function of X that is continuous and has continuous first and second derivatives, but jumps in its third derivative at selected "knots".

$$E(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \sum_{k=4}^{p} \beta_k [(X - c_{k-3})^+]^3$$

where u^+ = u if u > 0 and 0 otherwise.

Interpretation of coefficients:

- β_0, β_1, β_2, β_3: coefficients of a cubic function for left most interval
- β_j, $j > 3$: change in cubic coefficient slope from j-3rd to j-2nd interval (not very useful on its own)

F. Application of cubic spline to toy example:

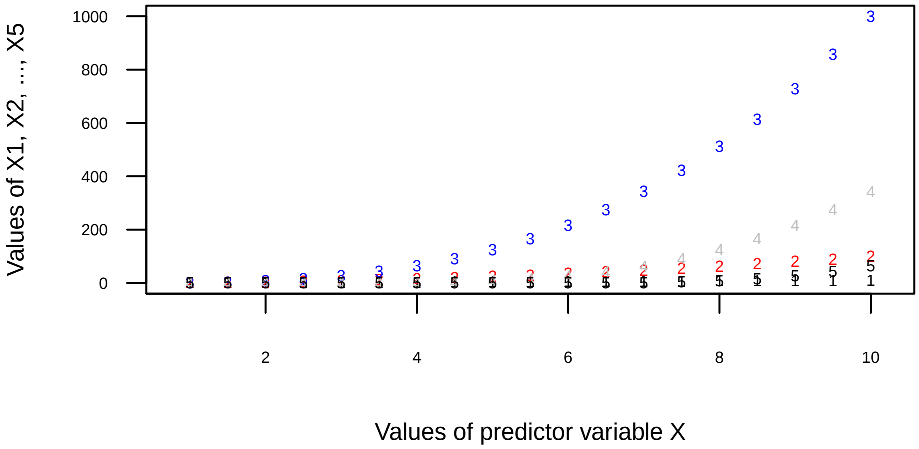
Given knots of $c_1 = 3$ and $c_2 = 6$, list (with definitions) the variables that will define your "Design Matrix". HINT: you need to define 5 variables.

1. Design matrix for cubic spline model

The table below displays the design matrix for the cubic spline model for values of X = 1, 2, . . . , 9 with knots $c_1 = 3$ and $c_2 = 6$.

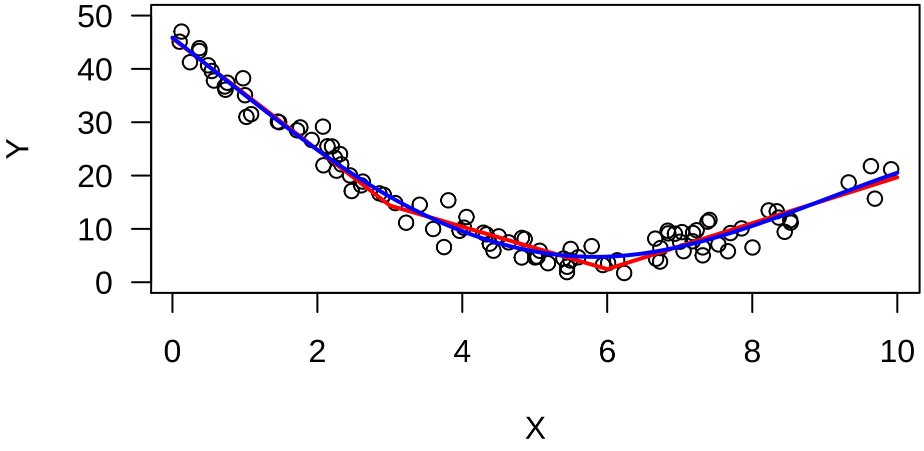
intercept	x1	x2	x3	x4	x5
1	1	1	1	0	0
1	2	4	8	0	0
1	3	9	27	0	0
1	4	16	64	1	0
1	5	25	125	8	0
1	6	36	216	27	0
1	7	49	343	64	1
1	8	64	512	125	8
1	9	81	729	216	27

Again, we create an illustration of the design matrix but for the cubic spline model. In the figure below, the numbers indicate the predictor variables X_1, X_2, . . . , X_5 respectively.



2. Assessment of quality of the model

Below is a figure comparing the model fit using the linear spline (red line) and cubic spline (blue line) models.



Which of the two models do you prefer and why?

In Lab 2, you will discuss an approach for evaluating which of these two models fits the data “best”.

III. Interactions of Simple Functions

Interactions of simple functions allow us to explore if E(Y|X) = f(X) is different for defined subsets of the population.

Here are some common questions that involve interactions.

Solutions can be found in the Lecture 2 slides and recordings.

1. During the first year of life, is the average “growth rate” in weight for male infants the same as for female infants?

- Draw a figure representing this question

- Translate the question and visualization into a regression model. Define appropriate variables and interpret the regression coefficients

2. Same as 1, but “growth curve”, rather than “growth rate”

- Draw a figure representing this question

- Translate the question and visualization into a regression model. Define appropriate variables and interpret the regression coefficients

3. Is the effect on average medical expenditures of being both poor and older greater than would be expected given the independent effects of poverty and old age alone.

- Draw a figure representing this question

- Translate the question and visualization into a regression model. Define appropriate variables and interpret the regression coefficients