



JOHNS HOPKINS  
BLOOMBERG SCHOOL  
*of* PUBLIC HEALTH

## Lecture 2

---

Basic functions for building regression  
models: splines; indicator variables;  
interactions

# Review of key concepts from Lecture 1

- ▶ Statistics: Methodology for the scientific method using quantitative evidence
- ▶ Statistical model:
  - ▶ a mathematical approximation that describes the mechanism by which the observed data might have been generated
  - ▶ Provides a precise statement of a set of hypotheses
  - ▶ Are models correct? True? Right?
  - ▶ Models can be useful by helping you operationalize the scientific method
- ▶ Regression:
  - ▶  $(Y, X)$  have a joint distribution, note  $X$  may be a series of columns of information.
  - ▶ The regression of  $Y$  on  $X$  is  $E(Y|X)$



# Two key uses / purposes for regression

1. Study the etiology of a process; how Y is caused by or associated with a set of Xs
  - ▶ Let  $X=(R,C)$ ; Study how risk factors R affect the outcome Y while controlling for potential confounders C
  - ▶ Let  $X=(R,C,E)$ ; Study how the effects of risk factors R are modified by variables E while controlling for confounders C
2. Predict Y using X

Regression fundamentals are the same for both!

Features of the regression model fit of most interest and strategy for model building can differ depending on your purpose.



# Types of regression discussed in 140.653-654

- ▶ General:  $\text{ave}(Y|X)$
- ▶ Linear model:  $\text{ave}(Y|X) = B_0 + \sum_{j=1,p} B_j X_j$
- ▶ Additive models:  $\text{ave}(Y|X) = \sum_{j=1,p} s_j(X_j)$
- ▶ Generalized linear models (GLMs):  $g(\text{ave}(Y|X)) = B_0 + \sum_{j=1,p} B_j X_j$  ; g- “link” function
  - ▶ Linear:  $g(u) = u$
  - ▶ Logistic:  $g(u) = \log(u/(1-u)) = \text{“logit”}(u)$
  - ▶ Log-linear:  $g(u) = \log(u)$
  - ▶ Probit, tobit, complementary log-log,...
- ▶ Generalized additive models (GAMs):  $g(\text{ave}(Y|X)) = B_0 + \sum_{j=1,p} s_j(X_j)$
- ▶ Classification and regression trees (CART):  $E(Y|X)$  is a “step function” in higher dimensional X-space
- ▶ Random forests:  $E(Y|X)$  is an average of a large number of “bootstrapped” trees



# Key datasets

## **Nepali Children's Anthropometry (NCA) Data**

- ▶ Cross-sectional nutrition survey of 4,000+ pre-school children
- ▶ Height, weight, arm-circumference and age on each
- ▶ Questions:
  1. How does height vary with age. What is the average “growth rate” over the first 5 years of life?
  2. How does shorter-term nutritional status vary by age; are younger children in better or worse status as measured by weight or arm-circumference controlled for height?
  3. How well can you predict a child's weight given his height and age?



# Key datasets

## **National Medical Expenditure Survey – Medical costs and smoking-caused diseases**

- ▶ Now known as Medical Expenditure Panel Survey, conducted by AHRQ
- ▶ NEMS 1987 – national survey of 20,000 non-institutionalized adults, included supplemental survey on smoking behaviors
- ▶ Key variables: total medical expenditures, presence of smoking-caused disease (Lung cancer, COPD, CHD, Stroke,...), age, gender, SES, smoking status
- ▶ Questions:
  1. How much more is spent per year on persons with smoking-caused diseases (SCDs) than on otherwise similar persons without SCDs?
  2. Does this SCD-attributable expenditure differ by current smoking status or access to health care?
  3. How does the risk of LC or COPD depend on the total pack-years of smoking and age?
  4. How does the risk of CHD/Stroke change for former smokers as a function of the time since they quit?

# Today's main topic

- ▶ Classical Multiple Linear Regression Model
- ▶ Basic tools for building regression models:
  - ▶ Step functions
  - ▶ Linear splines
  - ▶ Cubic splines
- ▶ Interactions



# Classical multiple linear regression (mlr) model

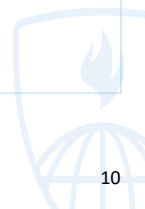




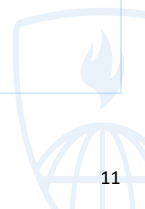
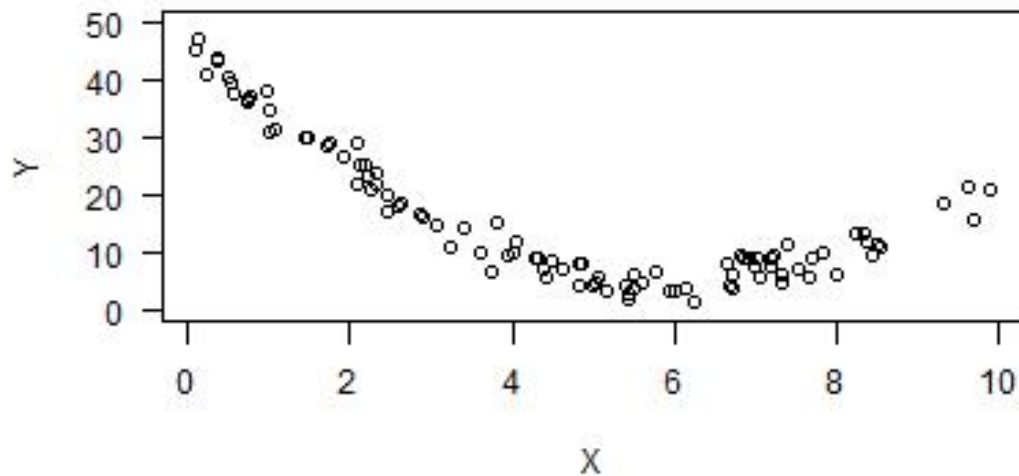
# Picture of $p = 1$ case



# What is linear about a MLR?



## Toy example for today's lecture



# Step functions

## Creating a step function to describe the relationship between average Y and X:

- ▶ X in range (a,b)
- ▶ Partition the range into p intervals:  $(a=c_0, c_1, c_2, \dots, c_p=b)$ 
  - ▶ 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  $j=1, \dots, p-1$
- ▶ One less indicator variable than interval!
- ▶ Fit MLR with intercept:  $Y_i = B_0 + B_1 X_{i1} + \dots + B_{p-1} X_{ip-1} + e_i$



# Step functions

- ▶ 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:  $Y_i = B_0 + B_1 X_{1i} + B_2 X_{2i} + e_i$
- ▶ Interpret the following:
  - ▶  $B_0$
  - ▶  $B_0 + B_1$
  - ▶  $B_1$



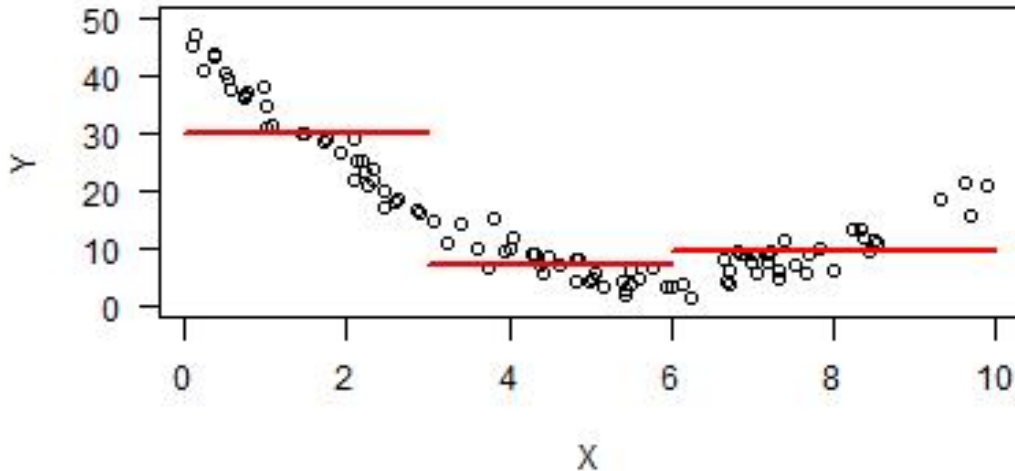
# Step functions

- ▶ 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:  $Y_i = B_0 + B_1 X_{1i} + B_2 X_{2i} + e_i$
- ▶ Interpret the following:
  - ▶  $B_0$
  - ▶  $B_0 + B_2$
  - ▶  $B_2$



# Step functions

In our toy example, the estimated values for  $B_0$ ,  $B_1$  and  $B_2$  are 30, 7 and 9, respectively.



Creating step functions is common; what do you think about this practice?



# Linear splines

- ▶ 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”
- ▶ Formula:  $E(Y|X) = B_0 + B_1 X + B_2 (X-c_1)^+ + B_3 (X-c_2)^+ + \dots B_{k+1} (X-c_k)^+$

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

- ▶ Interpretation of coefficients:
  - ▶  $B_0, B_1$  intercept and slope in the left-most interval
  - ▶  $B_j$ , for  $j \geq 2$ : change in slope from before to after associated knot





# Linear splines

- ▶ Back to our toy example, consider the multiple linear regression of  $Y$  on:

$$X_1 = X, X_2 = (X-3)^+; X_3 = (X-6)^+$$



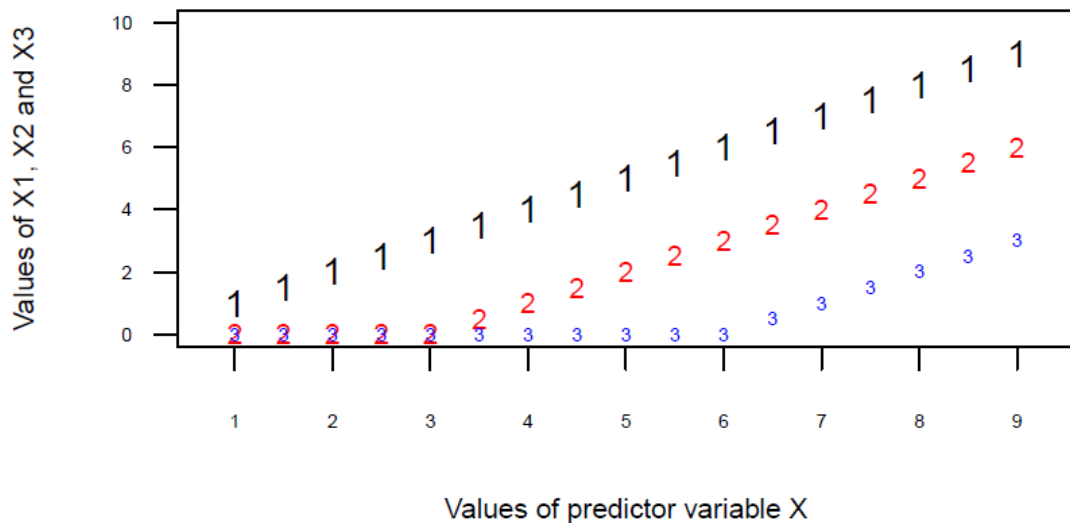
# Linear splines

- ▶ The design matrix contains information about predictor/covariate values for each observation in the data
- ▶ The columns of the matrix define each predictor variable
- ▶ The rows of the matrix provide the values of the predictor variable for the corresponding observation

Intercept	$X_1=X$	$X_2=(X-3)^+$	$X_3=(X-6)^+$
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

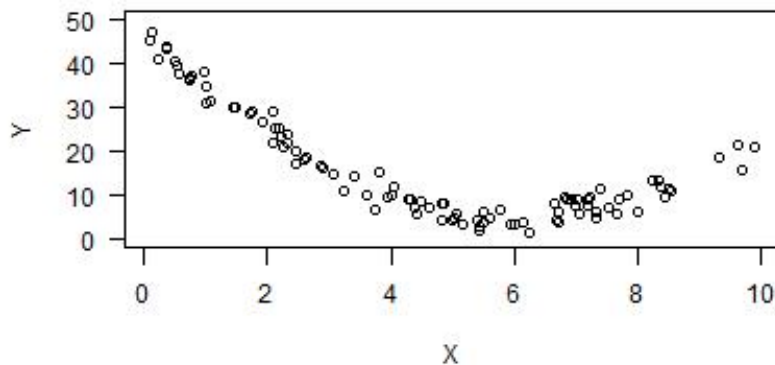
# Linear splines

- ▶ We can make a visualization of the design matrix for our linear spline model



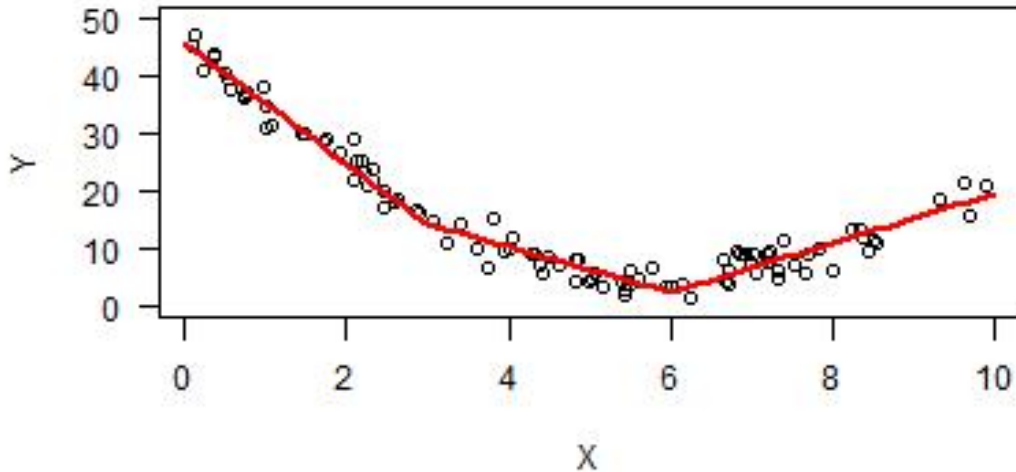
# Linear splines

- ▶ Estimate the values of  $B_0$ ,  $B_1$ ,  $B_2$ , and  $B_3$



# Linear splines

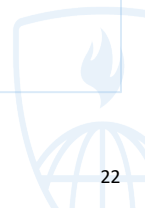
- ▶ I estimated the linear spline model parameters (we will discuss how later) and plotted the estimated mean of  $Y$  for each value of  $X$ .
- ▶ Do you prefer the linear spline model to the step function approach? Why?



# Cubic splines

- ▶ Idea: linear splines are nice, but they have ugly elbows (discontinuities in their first derivative); 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, with jumps in its third derivative at selected “knots”
- ▶ Formula:  $E(Y|X) = B_0 + B_1 X + B_2 X^2 + B_3 X^3 + \text{Sum}(k=4,p) \{ B_k [(X-c_{k-3})^+]^3 \}$

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



# Cubic Splines

► Formula:

$$E(Y|X) = B_0 + B_1 X + B_2 X^2 + B_3 X^3 + \text{Sum}(k=4,p) \{ B_k [(X-c_{k-3})^+]^3 \}$$

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

► Interpretation of coefficients:

- $B_0, B_1, B_2, B_3$  – coefficients of a cubic function for left most interval
- $B_j, j>3$  – change in cubic coefficient slope from  $j-3^{\text{rd}}$  to  $j-2^{\text{nd}}$  interval (not very useful on its own)

# Cubic splines

- ▶ Within the toy example with knots of  $c_1 = 3$ ;  $c_2 = 6$ , list (with definitions) the variables that will define your “Design Matrix”. HINT: you need to define 5 variables.





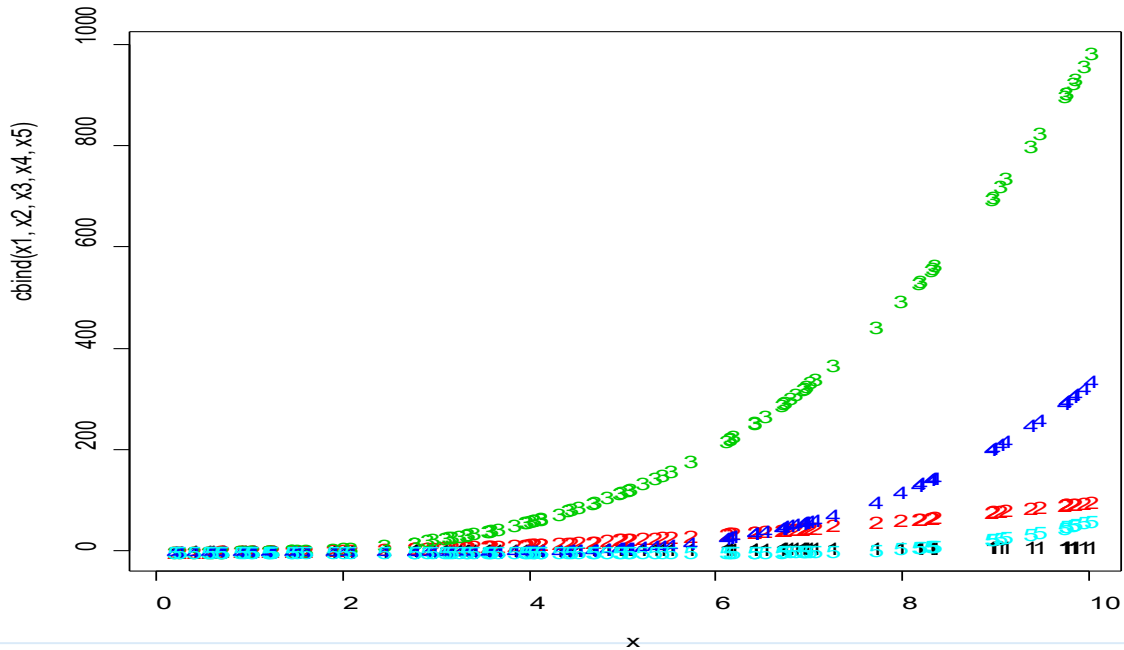
# Cubic splines

- Example of rows of the design matrix for the cubic spline model for values of  $X = 1, 2, \dots, 9$

Intercept	$X_1=X$	$X_2=X^2$	$X_3=X^3$	$X_4=[(X-3)^+]^3$	$X_5=[(X-6)^+]^3$
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

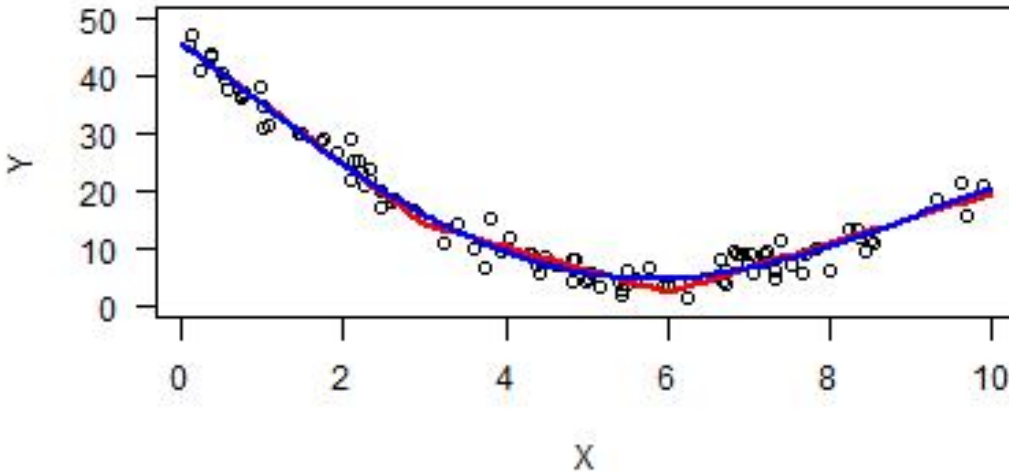
# Cubic splines

## ► Visualization of the design matrix



# Cubic splines

- ▶ I fit the cubic spline model and overlaid the estimated average  $Y$  vs.  $X$  based on the cubic spline model and the linear spline model.
- ▶ Which model do you prefer?



# Evaluation of model fit

- ▶ You will review this idea in Lab 2
- ▶ Idea:
  - ▶ Which of these models “fits” the data best?
  - ▶ Which of the models minimizes the error or residual, i.e. the distance between the observed  $y$  and the average  $y$  given  $x$ .
- ▶ If you evaluate the model fit using the same data that you used to fit the model, you will be overly optimistic in your assessment
- ▶ Cross-validation!



# Interactions of simple functions

- ▶ Interactions allow for  $E(Y|X) = f(x)$  to vary across subsets of the population of interest
- ▶ Effect modification
- ▶ During the first year of life, is the average “growth rate” in weight for male infants the same as for female infants?



# Interactions of simple functions

- ▶ During the first year of life, is the average “growth curve” in weight for male infants the same as for female infants?



# Interactions of simple functions

- ▶ 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

## In the next class session....

- ▶ We will work together to apply the concepts we learned in this lecture to the Nepal dataset

