MTH499/599 Lecture Notes 04

Donghui Yan

Department of Math, Umass Dartmouth

January 25, 2017

Outline

- Introduction
- A toy example
- Regression

Examples of models you may hear of

- Newton's law on distance and gravity

$$s = \frac{1}{2}gt^2$$

- Heat equation as model of heat transfer

$$\frac{\partial u}{\partial t} - \alpha \nabla^2 u = 0$$

- Turing machine as a model of computer
- Fractional Brownian motion as a model for stock

$$\frac{dS_t}{S_t} = adt + \sigma dZ_t$$

- Your own examples?

What is a model?

- Roughly, a device which one could use to *interpret or understand* a phenomenon
  - It may not be the same as 'truth'
  - "All models are wrong; some are useful" (George Box, 1919-2013)
- In statistics, we model the *relationship* between X and Y as

$$Y = f(X) + \epsilon,$$

and estimate f via a sample $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$.

- Unsupervised learning when Y is not present.

Techniques for modeling

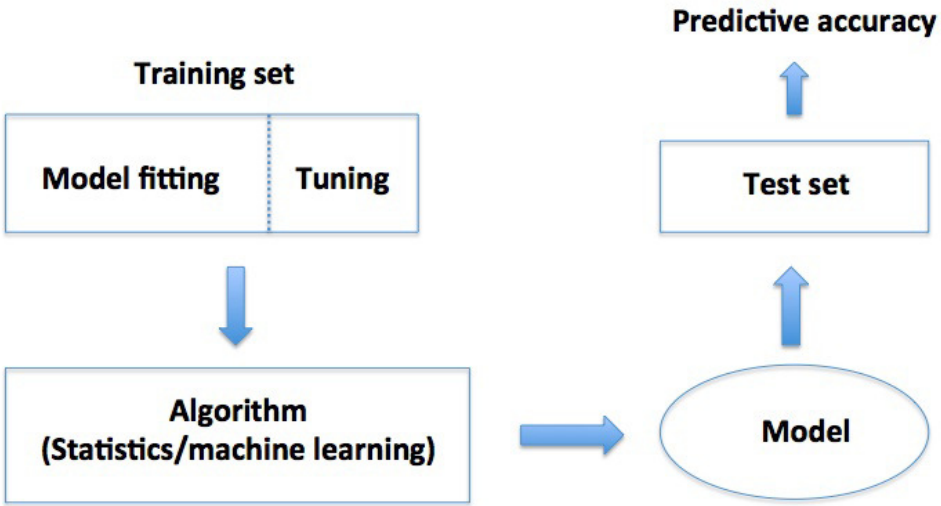
- Statistical modeling
  - Linear or generalized linear model
    - f can be expressed as a simple function of parameters θ
    - e.g., simple linear model, logistic regression
  - Nonparametric model with f as basis expansion
    - e.g., spline, kernel smoothing, trees etc
  - Semi-parametric model
    - Parametric part incorporates knowledge and residue as a nonparametric model
- Machine learning methods
  - Supervised, unsupervised, semi-supervised learning, depending on available of labels (i.e., Y)
  - Overlap between statistical and machine learning methods
- Mathematical modeling (omitted)
  - e.g., differential equations.

What makes a good model?

- Really depends on the goal
  - *Predictive accuracy* if the goal is prediction
    - The machine learning predictive culture
  - *Interpretation* if the goal is to understand a phenomenon
    - Good estimation + simple model
    - Traditional statistics ('generative culture')
  - Sometimes a combination of both
  - The general guideline is *a balance of model complexity and predictive ability*
    - More parameters, complicated form of $f \leftrightarrow$ complicated model
    - This is why linear model is used so often.

Predictive accuracy

- Needs two sets of data
  - One for modeling fitting (called *training set*)
    - May further split: part for model fitting and part for tuning
  - One for model evaluation *test set*
    - *Cross-validation* or *bootstrap* if data is scarce
- To compare two different models
  - Train on the same training set and evaluate on the same test set
  - Repeat on different datasets and *t-test* for rigorous comparison.



Linear model

- A linear model is specified by $Y = \beta_0 + \beta_1 X + \epsilon$
  - β_0 and β_1 are constants, ϵ is random error
  - Relationship linear in $\boldsymbol{\beta} = (\beta_0, \beta_1)$ $(not\ in\ X)$
  - Most used model in practice
- Terminology
  - $Y \leftrightarrow$ dependent variable, or response variable
  - $X \leftrightarrow$ independent variable, predicator variable, or regressor
  - $\boldsymbol{\beta} \leftrightarrow$ *parameters* or regression coefficients
- Why linear model?
  - Simple to express, easy for model fitting and interpretation
  - Many complicated phenomena can be well approximated by a linear model.

A toy example

- A salesperson collected some data about a used Saturn car from newspaper on advertised price and age

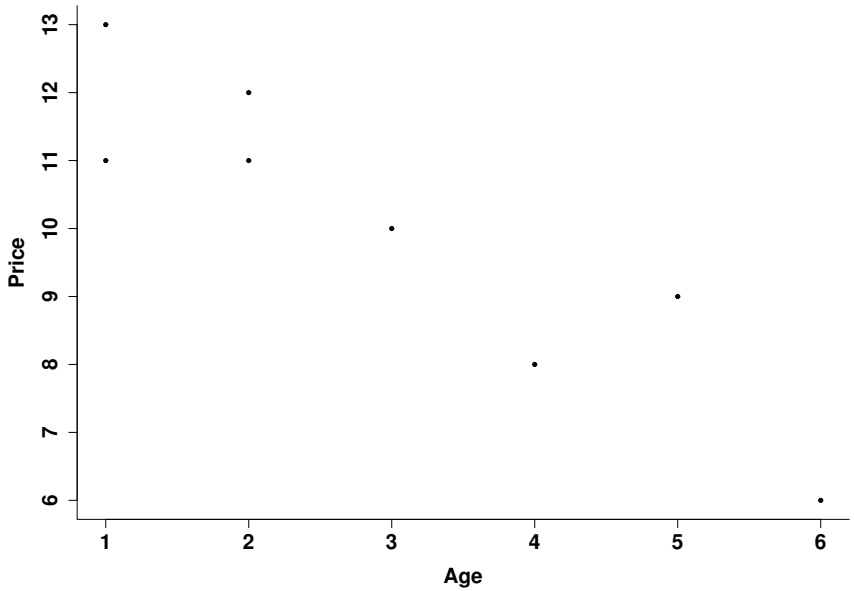
Age (years)	6	5	4	3	2	2	1	1
Price (thousands)	6	9	8	10	11	12	11	13

- He wants to find out the relationship between price and age.

A toy example (continued)

- In statistical data analysis, it often helps to plot the data (visualization)
  - *exploratory data analysis*
    (A field pioneered by John W. Tukey)
  - What can one observe from the plot?

John W. Tukey (1915-2000)

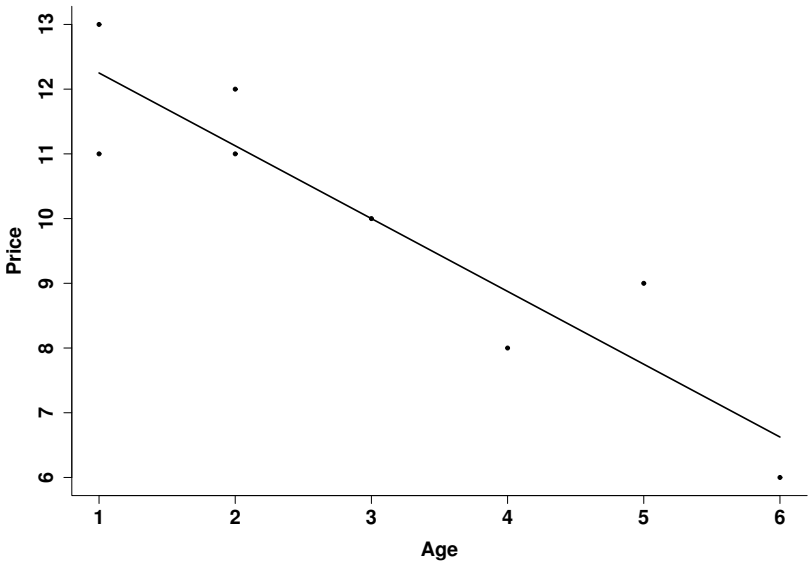


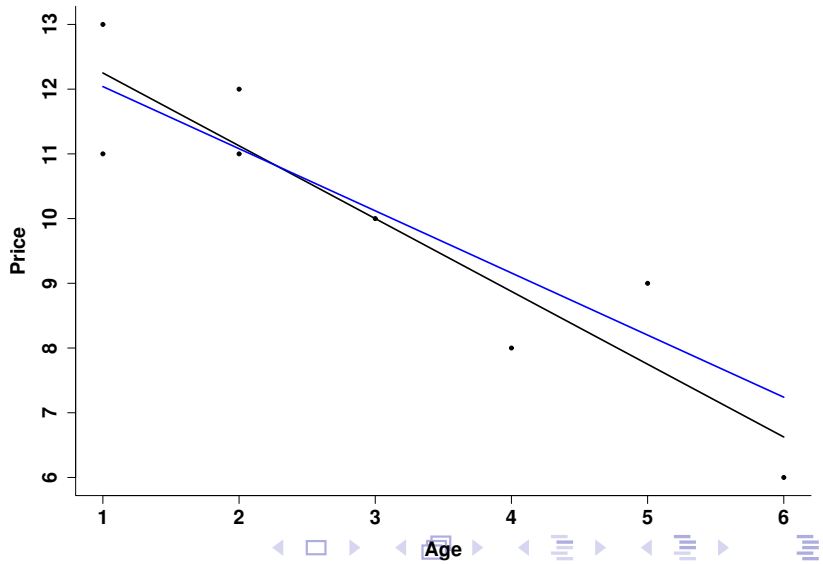
A toy example (continued)

- It seems that a straight line could describe $price \sim age$

$$\text{Price} = \beta_0 + \beta_1 * Age$$

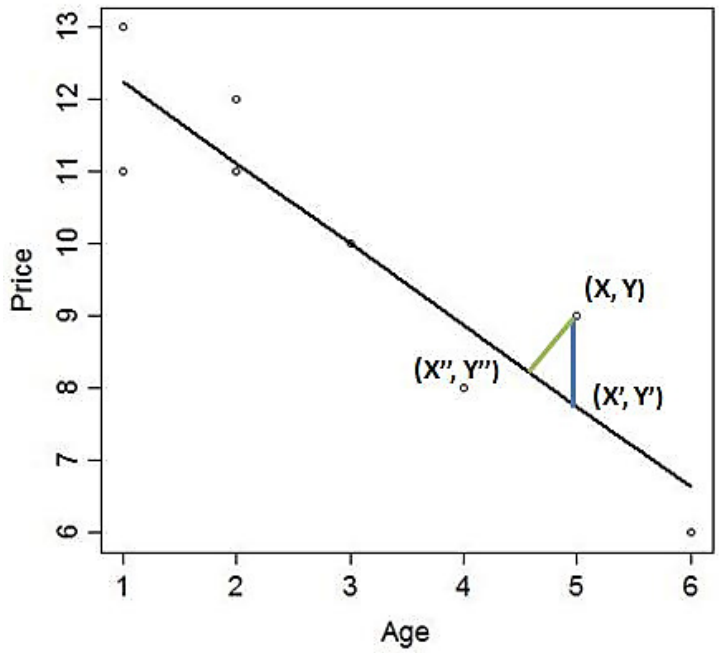
  - Indeed there are multiple such lines.





A toy example (continued)

- Our goal is to find the 'best' line
  - Only makes sense when a metric of 'best' is defined
  - Possibilities
    - Least square of mis-fit $\sum_{i=1}^{n}(Y_i - Y_i')^2$
    - Least absolute mis-fit $\sum_{i=1}^{n} \mid Y_i - Y_i' \mid$
    - Least distance $\sum_{i=1}^{n} \mid\mid x_i - x_i'' \mid\mid$



A toy example (continued)

- The least square fit $Y = \beta_0 + \beta_1 X$ was chosen
  - Simple and can lead to an analytical solution
  - Let $\hat{Y}_i$ denote fitted value. Then $\hat{Y}_i = \beta_0 + \beta_1 X_i$ and

$$\begin{aligned} \mathcal{L}(\beta_0, \beta_1) &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \\ &= \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \end{aligned}$$

  - Thus LS fit amounts to solve

$$\arg \min_{\beta_0, \beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

for a given data sample $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_N)$.

Least square fit

- Commonly referred to as OLS and credited to C.F. Gauss (1795)
- To solve the LS fit problem

$$\mathcal{L}(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

- Taking partial derivatives of $\mathcal{L}$ w.r.t. β_0 and β_1 and setting to 0

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = 0 \\ \frac{\partial \mathcal{L}}{\partial \beta_1} &= -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) X_i = 0 \end{aligned}$$

Least square fit (continued)

We arrive at a system of linear equations with two unknowns

$$\begin{cases} \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) & = 0 \\ \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)X_i & = 0 \end{cases}$$

$$\Longrightarrow \begin{cases} \sum Y_i - n\beta_0 - \beta_1 \sum X_i & = 0 \\ \sum X_i Y_i - \beta_0 \sum X_i - \beta_1 \sum X_i^2 & = 0 \end{cases}$$

$$\Longrightarrow \begin{cases} \sum X_i \sum Y_i - n \sum X_i \beta_0 - \left(\sum X_i\right)^2 \beta_1 & = 0 \\ n \sum X_i Y_i - n \left(\sum X_i\right) \beta_0 - n \left(\sum X_i^2\right) \beta_1 & = 0. \end{cases}$$

Least square fit (continued)

- Solving a system of linear equations and we get

$$\hat{\beta}_1 = \frac{\sum X_i Y_i - \sum X_i \sum Y_i/n}{\sum X_i^2 - (\sum X_i)^2/n} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}$$
$$= SS_{xy}/SS_{xx} = r \cdot SD_x \cdot SD_y/(SD_x)^2 = r \cdot SD_y/SD_x$$

$$\hat{\beta}_0 = \sum Y_i/n - \hat{\beta}_1 \sum X_i/n = \overline{Y} - \hat{\beta}_1 \overline{X}$$

- Interpretation
  - $\hat{\beta}_1$ is normalized correlation of variable X and Y
  - The center of regression, $(\overline{X}, \overline{Y})$, passes the regression line.

Least square fit of the toy example

- By hand calculation, we get

$$\overline{X} = 3,\ \overline{Y} = 10,$$
$$S_x = 1.851640,\ S_y = 2.267787,$$
$$r = -0.9185587$$

- So $\hat{\beta}_0 = 13.375,\ \hat{\beta}_1 = -1.125$, and

$$Price = 13.375 - 1.125 \cdot Age$$

  - Same as what we would get by R function $lm(Y \sim X)$
  - Interpretation: the price of used Saturn car decreases by 1.125K dollars each year.

Some quantities from the least square fit

- Two basic quantities

$$\begin{aligned} SS_{xx} &= \sum (x-\bar{x})^2, \\ SS_{xy} &= \sum (x-\bar{x})(y-\bar{y}) \end{aligned}$$

- Quantities related to sum of squares

$$\begin{aligned} \text{Total sum of squares} \quad & SST = \sum (y-\bar{y})^2 \\ \text{Sum of squared errors} \quad & SSE = \sum e^2 = \sum (y-\hat{y})^2 \\ \text{Sum of squared regression} \quad & SSR = SST - SSE. \end{aligned}$$

The R^2 statistic

$$R^2 = SSR/SST = 1 - SSE/SST$$

- Is a crude measure of *goodness-of-fit*
- Also called *amount of variance explained* (in percentage)
  - Need to be cautious (back to this later)
    - As it does not consider the number of parameters
  - Adjusted R^2 is defined as

$$R^2_{adj} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

    - n is the number of observations
    - p is the number of parameters.

Quantities calculated on the toy example

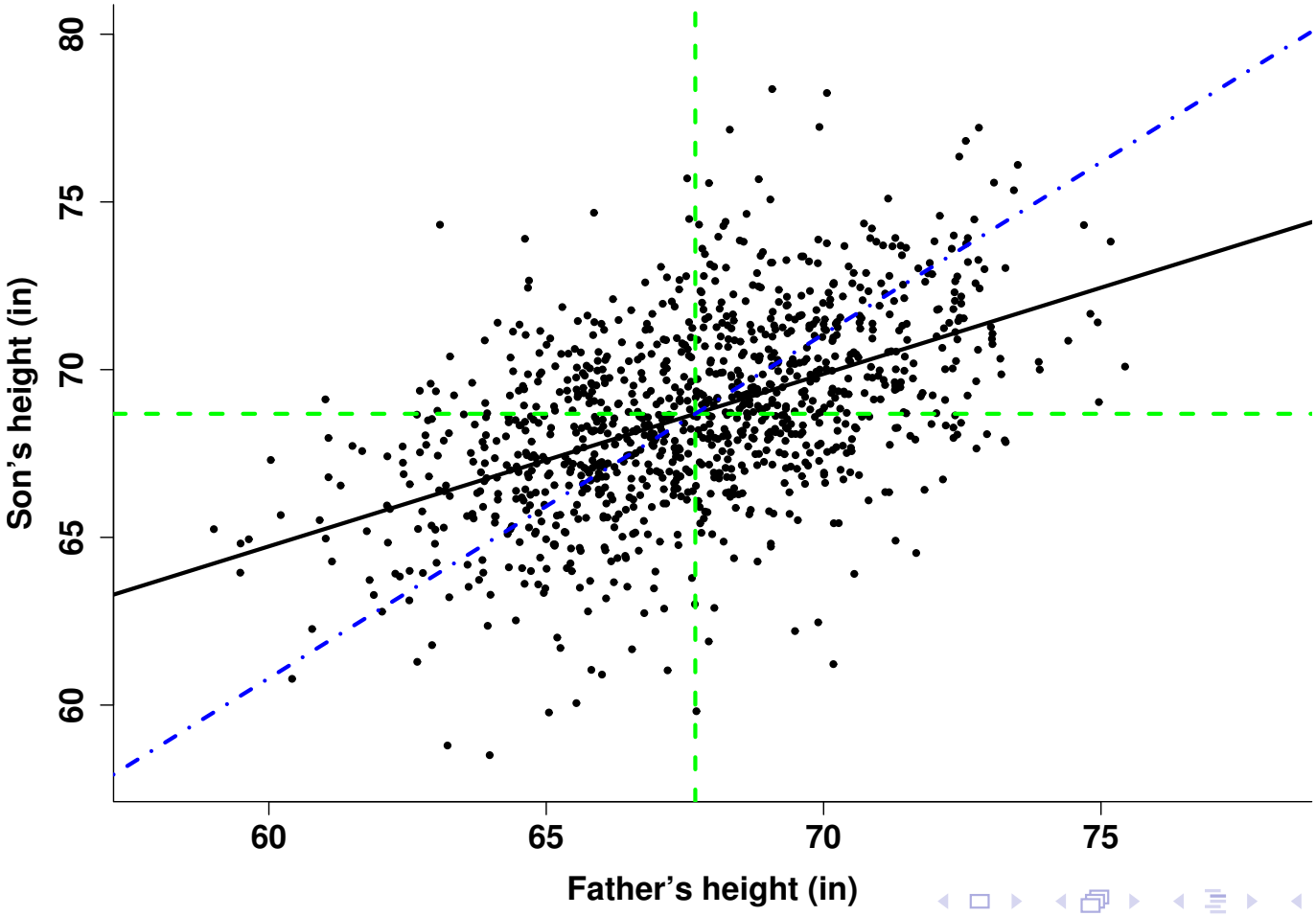
- $SS_{xx} = 24,\ SS_{xy} = -3.857$
- $SST = 36,\ SSE = 5.624,\ SSR = 30.376$
- $R^2 = 0.8437,\ R^2_{adj} = 0.8177$

Linear regression

- Linear least square fit also called *linear regression*
- *Regression* is a statistical phenomenon
  - Observed in the classical father-son data by Pearson (1857-1936)
  - 1078 measurements of a father's height and his son's height

Karl Pearson, one of the founders of modern statistics.

The father-son data and the regression line



Regression towards to the mean

- Regression towards towards the center of regression
  - *Regression towards mediocrity in hereditary stature*, Galton (1886)
  - One SD_x of change in $x \Longrightarrow r \cdot SD_y$ of change in y (why?)
    - When $x > Ave_x$, the increase in y is slower than in x
    - When $x < Ave_x$, the decrease in y is slower than in x.