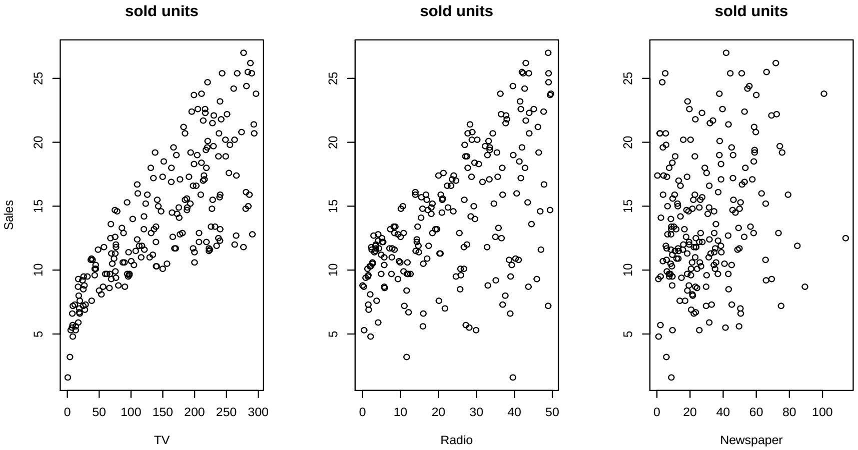
STA314, Lecture 1.

Sept 10, 2018
Stanislav Volgushev

Motivating example (regression)

You work for a marketing firm. One of your clients collected data in 200 market locations. For each market, they have the amount of money spent on TV, Radio and newspaper advertisement (in 1000 $) of a certain product and the sales (in thousands of units) of that product during a given time frame. They want you to help them predict sales and help decide how to advertise.



Some terminology

Typically, we will have a *data set (sample)* consisting of *data (observations)* $(x_1, Y_1), ..., (x_n, Y_n)$. Y_i are usually real values, x_i can be real values, vectors or other variables.

x_i are called *predictors (covariates, regressors, features, input variables)*.

Y_i are called *response (outcome, target, dependent variable)*.

What are the aims in Statistical Learning?

Prediction: Given a new predictor value x_0, predict response Y_0.

- Example: assume we spend 151 units on TV advertisement for a certain market. How many units do we expect to sell?
- Example: assume we spend 120 units on TV advertisement, 15 unites on radio advertisement and 5 units on newspaper advertisement. How many units do we expect to sell?

Inference: Understand the relationship between predictor and response in more detail and use that to make decisions. Quantify uncertainty.

- Example: does spending money of TV advertisement have any effect on sales?
- Example: assume we increase TV advertisement by 10 units. What is the effect on sales? How precisely can we quantify this effect?
- Example: assume we can spend a given budget on advertising on TV, radio and newspaper. How should we allocate the budget?

Regression models

Given observations $(x_1, Y_1), ..., (x_n, Y_n)$ try to find relationship between x_i and Y_i (example above: x_i is money spent on TV, Y_i are sales)

Machine Learning/Statistics: build a *regression model* to describe influence of x_i on Y_i.

Such models take the form

$$Y_i = f(x_i) + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i] = 0$$

where f is called *regression function* and ε_i are called *errors*.

- f describes the 'systematic' influence of x_i on Y_i.
- ε_i captures everything that f does not describe.
- $\mathbb{E}[\varepsilon_i] = 0$ implies: $f(x_i)$ is the average value of Y_i at predictor value x_i.

Remarks on ε_i

- In advertisement example: variation due unobserved factors such as size of market, local population, 'randomness' (ex.: weather for umbrella sales) etc.
- In Physics experiments: measurement error
- In Biology: genetic variation, environmental influence etc.

How do we use this model for prediction?

$$Y_i = f(x_i) + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i] = 0$$

First assume we know f. Given a new predictor x_0, what is the 'best way' to predict response Y_0?

More specifically: a typical way of measuring the quality of a prediction $\widehat{Y}_0$ at point x_0 is the *mean squared error* (short: **MSE**)

$$MSE(\widehat{Y}_0) = \mathbb{E}[(Y_0 - \widehat{Y}_0)^2].$$

Which $\widehat{Y}_0$ will minimise the MSE if f is known?

Some math (see blackboard for details): if $Y_0 = f(x_0) + \varepsilon_0$, x_0 fixed number and ε_0 independent of $\widehat{Y}$

$$\mathbb{E}[(\widehat{Y}_0 - Y_0)^2] = \mathbb{E}[(\widehat{Y}_0 - f(x_0))^2] + \mathbb{E}[\varepsilon_0^2].$$

- $\mathbb{E}[\varepsilon_0^2]$ irreducible part. Even if we know f_0 this part can not be improved.
- $\mathbb{E}[(\widehat{Y}_0 - f(x_0))^2] \geq 0$ depends on $\widehat{Y}_0$.
- MSE minimized at $\widehat{Y}_0 = f(x_0)$!

A first estimator for f: the K-nn method

The best possible value for $\widehat{Y}_0$ is $\widehat{Y}_0 = \mathbb{E}[Y_0] = f(x_0)$. Given data $(x_1, Y_1), ..., (x_n, Y_n)$, how do we *learn/estimate* $f(x_0)$?

First: assume $x_1, ..., x_n$ take only values 0 or 1 and $x_0 = 0$. A natural approach:

$$\widehat{f}(0) = Average(Y_i : x_i = 0) = \frac{\sum_{i=1}^n Y_i I\{x_i = 0\}}{\sum_{i=1}^n I\{x_i = 0\}}$$

Here

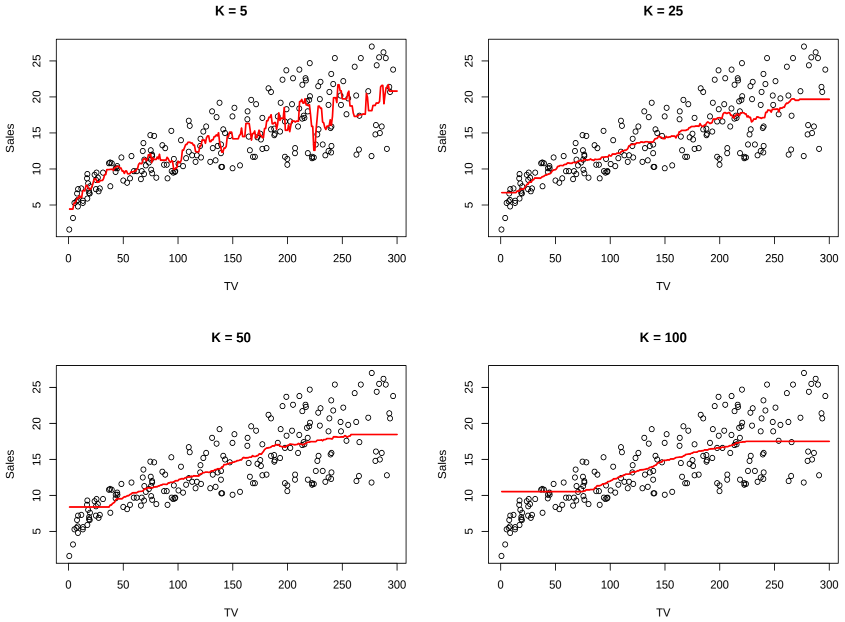
$$I\{x_i = 0\} = \begin{cases} 1, & \text{if } x_i = 0 \\ 0, & \text{otherwise} \end{cases}$$

Problem: this works in data sets where $x_1, ..., x_n$ take only few distinct values and we are interested in predicting outcome for one of those values. Example: advertisement data set has no point with $x = 152$.

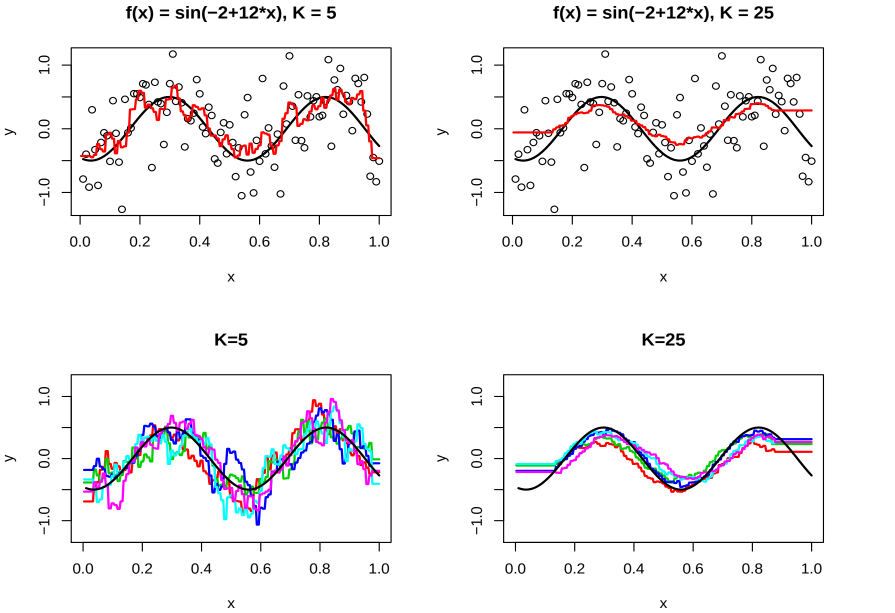
Idea: instead of requiring $x_i = x_0$ take K of the 'closest' x_i. **K-nn** (K *nearest neighbours*) method.

$$\widehat{f}(x_0) = \frac{\sum_{i=1}^n Y_i I\{x_i \text{ among closest K to } x_0\}}{K}$$

Examples of K-nn regression for regressing Sales on TV advertisement



Examples of K-nn with simulated data



Bias-variance decomposition

- For smaller K, regression function is very 'wiggly'. A lot of variation between samples generated from the same model.
- If K is very large, data are not described well. Less variation between samples generated from the same model.
- Intermediate values of K seem to be sensible.

This can be formalized through the concepts of *bias* and *variance*. Recall that the irreducible part of the MSE takes the form $\mathbb{E}[(\hat{Y}_0 - f(x_0))^2]$. After some calculations (see blackboard)

$$\mathbb{E}[(\hat{Y}_0 - f(x_0))^2] = \mathrm{Var}\,(\hat{Y}_0) + \{\mathbb{E}[\hat{Y}_0] - f(x_0)\}^2.$$

- $\mathbb{E}[\hat{Y}_0] - f(x_0)$ is called **bias**. It describes how far from the truth the prediction $\hat{Y}_0$ is *on average*.
- $\mathrm{Var}\,(\hat{Y}_0)$ describes how much variation the estimator $\hat{Y}_0$ has.
- Ideal estimator would have small bias and small variance, but usually that is impossible.

The bias and variance of K-nn

$Y_i = f(x_i) + \varepsilon_i$, ε_i i.i.d. with $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}\,(\varepsilon_i) = \sigma^2$. Then (see bb)

$$\mathrm{Var}\,[\hat{f}(x_0)] = \sigma^2/K,$$

Bias: more complicated. For simplicity: $x_i = i/n, i = 1, ..., n$, $f : [0,1] \to \mathbb{R}$ two times continuously differentiable, $K = 2\ell + 1$, $x_0 = j/n$ with $\ell < j < n - \ell$. Then (see bb)

$$\mathbb{E}[\hat{f}(x_0)] = f(x_0) + \frac{1}{K}\sum_{u=-\ell}^{\ell} f\Big(\frac{j+u}{n}\Big) = f(x_0) + \frac{1}{3}f''(x_0)(K/n)^2 + r_{K,n}$$

where $r_{K,n}$ remainder term, 'small'.

Variance decreases as K increases, bias increases as K increases. To get a good MSE we have to balance bias and variance. This is called *bias-variance trade-off*.

Analysis exercise: if $\ell = \ell_n$ with $\ell_n \to 0, n\ell_n \to \infty$ as $n \to \infty$ then remainder term above is $o(\ell_n^2/n^2)$.