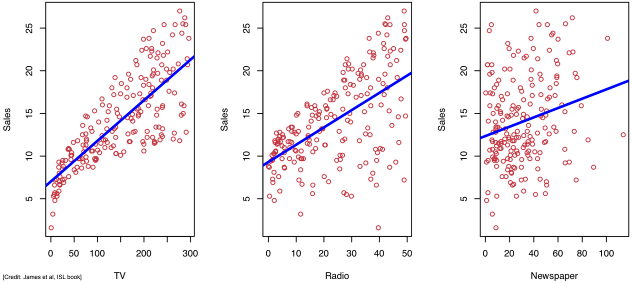
Section 2A. Intro to Statistical Learning

Statistics for Data Science

Victor M. Preciado, PhD MIT EECS
Dept of Electrical & Systems Engineering
University of Pennsylvania
preciado@seas.upenn.edu

What is Statistical Learning?

- **Example**: Consider data collected from 200 marketing campaigns
  - For each campaign, we know amounts invested in ***radio***, ***TV***, and ***newspaper*** advertisement
  - For each campaign, we also know the total number of ***sales***
  - Using this data, can we predict **Sales** (our output) of a campaign from the **invested amounts** (our inputs)? Sales $\approx f\,(\text{Radio}, \text{TV}, \text{Newspaper})$



[Credit: James et al, ISL book]

Notation

- We generically denote the input vector by $X = (X_1, \ldots, X_p)^\intercal$ and a particular value of the input vector by $\mathbf{x} = (x_1, \ldots, x_p)^\intercal$. In the previous example, the *input* vector has three entries ($p = 3$):
$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} \textbf{TV} \\ \textbf{Radio} \\ \textbf{Newspaper} \end{pmatrix}.$$
Particular values of the input vector X will be denoted by a bold font letter $\mathbf{x} = (x_1, \ldots, x_p)^\intercal \in \mathbb{R}^p$. For example, define $\mathbf{x} = (1000, 800, 900)^\intercal$; hence, $X = \mathbf{x}$ indicates that the investments are **TV** $= 1000$, **Radio** $= 800$, and **Newspaper** $= 900$.

- We generically denote the output variable by Y and a particular value of the output by y. For example, define $Y =$ **Sales** and $y = 2500$; hence, $Y = y$ indicates that **Sales** $= 2500$.

Statistical Learning problem

Theoretical setup:

- In this course, we assume that X is a vector of r.v.'s and Y is a scalar r.v.
- In the field of statistical learning, we commonly assume that our outputs are generated by an *additive model* of the form

  $$Y = f(X) + \varepsilon$$

  where f is an *unknown* function called the *regression function* and ε represents the *measurement noise* (which we assume to be a r.v. with zero mean and known variance, $\mathbb{E}[\varepsilon] = 0$ and $\mathrm{Var}[\varepsilon]=\sigma^2$).

Problem: The main problem in statistical learning is to estimate f (the unknown regression function) from random samples drawn from the additive model.

Penn Engineering