6.867 Machine Learning Fall 2017

Lecture 2. Linear Regression

Advertisement Campaign

- Planning Marketing Budget Across Channels: TV, Radio and NewsPaper

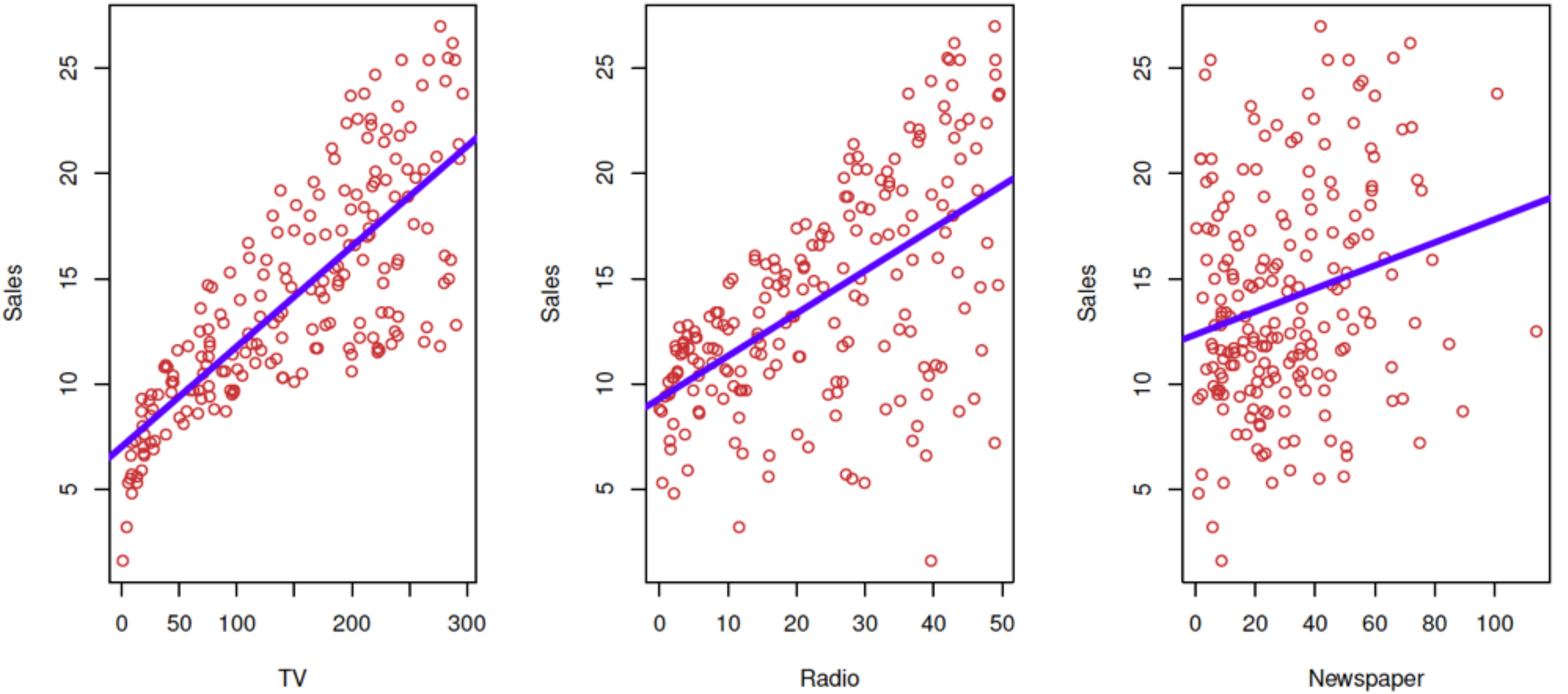
- Data across 200 Markets
    - Spending for TV, Radio, NewsPaper
    - Resulting Sales

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
6	8.7	48.9	75	7.2
7	57.5	32.8	23.5	11.8
8	120.2	19.6	11.6	13.2
9	8.6	2.1	1	4.8
10	199.8	2.6	21.2	10.6

Sample Data

- Questions
    - Is there a relationship between Sales and Marketing Budget?
    - If yes, can we "predict" Sales given Marketing Budget across Channels?
    - And, how "important" are each of the channels? do they interact?

Advertisement Data



Advertisement Campaign and Regression

- Data: in market $n, 1 \leq n \leq 200$
  - Sales y_n
  - Channel spending: $\mathbf{x}_n = (x_n^{\text{TV}}, x_n^{\text{Radio}}, x_n^{\text{Newspaper}})$
- Regression: fit a function or model $f : \mathbf{x} \rightarrow y$
  - so as to minimize *(squared) loss*

$$\sum_{n=1}^{200} (y_n - f(\mathbf{x}_n))^2$$

- Ideal solution: (if we know joint distribution of $Y, \mathbf{X}$)

$$f(\mathbf{x}) = \mathbb{E}[Y | \mathbf{X} = \mathbf{x}]$$

Linear Regression

- Approximate $f(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ as

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + \text{constant}$$
$$= w_{\text{TV}}x^{\text{TV}} + w_{\text{Radio}}x^{\text{Radio}} + w_{\text{NewsPaper}}x^{\text{NewsPaper}} + w_0$$

- Or, more generally

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} = \sum_{i=0}^{p} w_i x_i, \;\; \text{with } x_0 = 1$$

- Regression: find $\mathbf{w}$ that minimizes

$$\sum_{n=1}^{N} (y_n - \mathbf{w}^T\mathbf{x}_n)^2$$

Linear Regression

- Let $Y^N = [y_n]^T$ and $X^N = [\mathbf{x}_n]^T$
- Then squared loss with parameter $\mathbf{w}$ can be written as

$$L(\mathbf{w}) = (Y^N - X^N\mathbf{w})^T(Y^N - X^N\mathbf{w})$$

- To find minimizing $\mathbf{w}$, we compute gradient of (removing ref to N)

$$\nabla L(\mathbf{w}) = -2X^TY + 2(X^TX)\mathbf{w}$$

- By setting gradient to 0, we obtain

$$\mathbf{w}^\star = (X^TX)^{-1}X^TY$$

Linear Regression

- Advertisement linear regression leads to

$$
\begin{aligned}
w_0 &= 2.939 \\
w_{\text{TV}} &= 0.046 \\
w_{\text{Radio}} &= 0.189 \\
w_{\text{NewsPaper}} &= -0.001
\end{aligned}
$$

$$\text{Residual-Sum-of-Square (RSS) or Fit Error} = 556.825$$

- Is this *good, bad, ugly* ?
  - Squared error / Residual Sum of Squares (RSS): not informative
  - How much should we trust model coefficients ?

Linear Regression: Evaluating Model

- An informative *relative* metric "R square"
  - In words: fraction of "variance" in the data explained by model
    - perfect fit will explain it fully, i.e. it will be 1
    - no fit will explain none, i.e. it will be 0

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}}, \text{ where } \mathrm{TSS} = \sum_n (y_n - \bar{y})^2, \ \bar{y} = \frac{1}{N} \sum_n y_n$$

- For Advertisement data:

$$R^2 = 0.897$$

Linear Regression: Uncertainty of Model Params

- Importance of channel: how much should we trust parameter values?
  - Find confidence intervals by evaluating their variances

$$\begin{aligned}\mathrm{Cov}[\mathbf{w}] &= \mathrm{Cov}[AY], \quad \text{where } A = (X^T X)^{-1} X^T \\ &= A\mathrm{Cov}[Y]A^T\end{aligned}$$

  - Recall

$$\begin{aligned}Y &= \mathbb{E}[Y|\mathbf{X}] + (Y - \mathbb{E}[Y|\mathbf{X}]) \\ &= f(\mathbf{X}) + \varepsilon, \quad \text{where } \mathbb{E}[\varepsilon] = \mathbf{0}\end{aligned}$$

  - We shall assume noise is Gaussian with zero mean

$$\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \Rightarrow \mathrm{Cov}[Y] = \sigma^2 \mathbf{I}$$

  - That is,

$$\mathrm{Cov}[\mathbf{w}] = \sigma^2 (X^T X)^{-1}$$

Linear Regression

- The standard deviations for parameters (after estimating σ)

$$
\begin{aligned}
w_0 &= 2.939 & \quad & 0.311 \\
w_{\text{TV}} &= 0.046 & & 0.0014 \\
w_{\text{Radio}} &= 0.189 & & 0.0086 \\
w_{\text{NewsPaper}} &= -0.001 & & 0.0059
\end{aligned}
$$

- Clearly suggests that
  - NewsPaper is not so effective (at least no confidence)
  - TV and Radio are effective with Radio more effective than TV
- Question: should we simply invest ALL Marketing budget in Radio?
  - Do investments in TV and Radio help each other?

Linear Regression

- Modified regression
  - Use variables $x_{\text{TV}}, x_{\text{Radio}}$ and $x_{\text{TV}} \times x_{\text{Radio}}$
  - It's again Linear Regression with different "data" matrix
  - Resulting $R^2 = 0.967!$
- In summary
  - There is a relationship between Marketing Budget and Sales
  - TV and Radio are primary channel affecting Sales
  - The investment in TV and Radio help each other
  - And resulting model is very good in its ability to predict

Linear Regression

- Using generic feature function: target Y, features $\mathbf{x}$
  - map: $\mathbf{x} \rightarrow [\phi_1(\mathbf{x}), \ldots, \phi_p(\mathbf{x})] \equiv \Phi(\mathbf{x})$
  - data: $(y_n, \Phi(\mathbf{x}_n)), \; 1 \leq n \leq N$
  - model: $y = w_0 + \sum_{i=1}^{p} w_i \phi_i(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x})$

- Least-squares solution

$$\mathbf{w}^\star = (\Phi^T \Phi)^{-1} \Phi^T Y$$

Maximum Likelihood

- Recall

$$Y = f(\mathbf{X}) + \varepsilon, \text{ where } \varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$

- Therefore, maximum-likelihood suggests selection of ***f*** so that

$$\mathbb{P}(\text{data}|f) \text{ (or its logarithm) is maximized}$$

- Now

$$\log \mathbb{P}((y_n, \mathbf{x}_n), n \geq 1 | f) = -\sum_n \frac{(y_n - f(\mathbf{x}_n))^2}{2\sigma^2}$$

$$- N \log \sigma - N \log \sqrt{2\pi}$$

Maximum Likelihood

- Therefore, maximum likelihood boils down to

$$\text{minimize } \sum_n (y_n - f(\mathbf{x}_n))^2 \text{ over } f$$

- When restricted to linear function class
  - This is precisely linear regression!

Role of Optimization

- The model selection boils down to solving optimization

$$\text{minimize } \sum_n (y_n - f(\mathbf{x}_n))^2 \text{ over } f$$

- For linear function $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$
  - We solved for optimal value in closed-form
  - But in general it may not be feasible
  - Or, matrix inversion may be prohibitive in memory consumption
  - Is there an *incremental* algorithm?

Gradient Descent

- Optimization:

$$\text{minimize } g(\mathbf{w}) \text{ over } \mathbf{w} \in \mathbb{R}^d$$

- Iterative algorithm: in iteration t+1

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \alpha_t \nabla g(\mathbf{w}^t)$$

  - where

$$\alpha_t \geq 0, \quad \lim_{t \to \infty} \alpha_t = 0, \quad \sum_t \alpha_t = \infty$$

Projected Gradient Descent

- Optimization:

$$\text{minimize } g(\mathbf{w}) \text{ over } \mathbf{w} \in \mathcal{C}, \text{ where } \mathcal{C} \text{ is a convex set}$$

- Iterative algorithm: in iteration $t+1$

$$\mathbf{v}^{t+1} = \mathbf{w}^t - \alpha_t \nabla g(\mathbf{w}^t)$$
$$\mathbf{w}^{t+1} = \mathrm{Proj}_{\mathcal{C}}(\mathbf{v}^{t+1})$$

$$\alpha_t \geq 0, \ \lim_{t \to \infty} \alpha_t = 0, \ \sum_t \alpha_t = \infty$$

Stochastic Gradient Descent

- Optimization for model learning:

$$\text{minimize } g(\mathbf{w}) \text{ over } \mathbf{w} \in \mathbb{R}^d$$

$$g(\mathbf{w}) = \sum_n (y_n - \mathbf{w}^T x_n) = \sum_n L(\mathbf{w}; x_n, y_n)$$

- Gradient has form

$$\nabla g(\mathbf{w}) = \sum_n \nabla L(\mathbf{w}; x_n, y_n)$$

- *Poor man*'s gradient descent

$$\mathbf{w}^{n+1} = \mathbf{w}^n - \alpha_n \nabla L(\mathbf{w}^n; x_n, y_n)$$

  - and potentially do this by passing over the dataset multiple times