- Why linear regression? · Basic technique w/ many extensions
  · Widely used

- Consider the following dataset which is the sales of a particular product in 200 markets w/ advertising budgets for three different media : television, newspaper and radio.

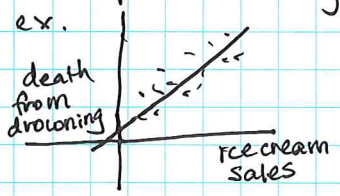
Possible questions :

1. Are there linear relationships btwn budget vs. sales? media vs. sales? How strong are these relationships?
2. Which media contribute to sales?
3. How accurately can we estimate the effect of media on sales?
4. How accurately can we predict future sales from media?

We cannot answer 2 or 3 with linear regression.

(2) implies causality, and correlation / association $\not\Rightarrow$ causality.

ex.



Obviously icecream sales don't cause death by drowning. They may correlate ~~as~~ if they are both affected by the same (third) variable.

ex. (good) Weather → death by drowning ↑
→ ice cream sales ↑

- Simple linear regression : we model linear dependence of a response variable / output / dependent variable Y on a predictor / input / feature / independent variable X.

- Stochastic linear model:

$$\boxed{Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}$$

ε : statistical error ; considered random

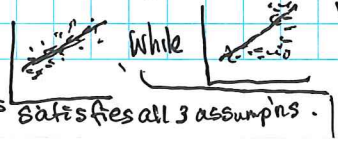
- Some notes and assumptions.

  - (ε (statistical error) can represent ex measurement error ; unknown, non-systemic effects)
  - We assume x_i are fixed ie non-random for simplicity.
  - β_0 and β_1 are parameters of the model ; they are unknown and non-random
  - Since we consider ε to be random $\Rightarrow$ Y is also random $\Rightarrow$ model is stochastic

Assumptions on ε_i's ① $\varepsilon_i, \dots, \varepsilon_n$ are independent and uncorrelated.

$\forall i = 1, \dots, n$ ② $E(\varepsilon_i) = 0$ $\Rightarrow$ no systemic error
③ $Var(\varepsilon_i) = \sigma^2$

ex.



violates ② and

violates ③

This satisfies all 3 assumptions.

Lecture 7: Introduction; Linear Regression, cont'd

22 Feb 2018

(from prev page) Simple linear regression: stochastic model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad , i = 1, \dots, n$$

- The data are represented as realizations $(x_1, y_1), \dots, (x_n, y_n)$ where y_i is a realization of Y_i.
- The linear model, like the one above, is linear in the parameters β_0, β_1

  ie. x_i may be non-linear in a linear model, ex. $Y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$

* pay attention to wording and symbols!

Note that:
- The unknown parameters are β_0, β_1 and σ^2.
- We cannot observe ε_i's since the true line (set of parameters) is unknown

Goal:
- We want to estimate β_0, β_1, σ^2 (true values)

So ...
- we want to find $\hat{\beta}_0$, $\hat{\beta}_1$ so that ("^" estimates") the line $\hat{\beta}_0 + \hat{\beta}_1 x$ is "close" to the data points.

That is ... the Game of Statistics:

- We see only one dataset
- We obtain the estimates $\hat{\beta}_0$, $\hat{\beta}_1$ by understanding the precision/accuracy of these estimates.

In the end
- We want to say something about the unknown β_0, β_1 (true vals)

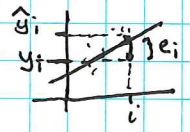
  ie. We want to infer something about the true distribution from estimates $(\hat{\beta}_0, \hat{\beta}_1)$

- How to measure "closeness"?

- often we find the params that give us the smallest residual sum of squares, (RSS). ie least squares (LSQ)

$$\underset{\text{(residual sum of squares)}}{RSS} = \sum_{i=1}^{n} e_i^2 \quad \text{where} \quad e_i = y_i - \hat{y}_i \text{ is the residual, and } \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \text{ is the fitted value}$$

- Graphically we see



Repetitions of regression give us

Note that regressions tend to be more accurate in the middle of the data than ends since the latter are more affected by (small) changes in slope

- For any combination (a, b) we can compute
  - $(\hat{y}_i)$ the fitted values $a + bx_i = \hat{y}_i$;
  - (e_i) the corresponding residuals $e_i = y_i - \hat{y}_i$; and
  - (RSS) the RSS $\sum_{i=1}^{n} e_i^2$

s.t. $$(\hat{\beta}_0, \hat{\beta}_1) = \underset{a,b}{\operatorname{argmin}} \sum_{i=1}^{n} (y - (a + bx))^2$$

Note the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are random!

Lecture 1 : Introduction; Linear Regression, cont'd

22 Feb 2018

Note / Tip : when examining LSQ regression lines, look at the data / line / graph interval by interval. The line segment should lie in the middle of the data in that interval

Multiple Linear Regression

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i$$

Here our unknown parameters are $\beta_1, \dots, \beta_p, \sigma^2$

- We have p predictor variables.
- We set $x_{i1} = 1$ s.t. $\beta_1 x_{i1} = \beta_1$ is the y-intercept of the model
- As before, $\varepsilon_1, \dots, \varepsilon_n$ are uncorrelated (and independent)
  - $E(\varepsilon_i) = 0$
  - $Var(\varepsilon_i) = \sigma^2$ $\Big\}\ \forall i = 1, \dots, n.$

We can write the above equation for our multiple linear model in

- vector form

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i \quad , \quad \mathbf{x}_i^T = (x_{i1} \cdot x_{i2} \cdots x_{ip})$$

or

and $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$

- matrix form

$$\underset{(n\times 1)}{\mathbf{Y}} = \underset{(n\times p)}{\mathbf{X}}\ \underset{(p\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}}$$

usually we assume $n > p$.

$$\underset{n\times 1}{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}} = \underset{n\times p}{\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & (x_{jk}) & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}} \underset{p\times 1}{\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}} + \underset{n\times 1}{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}$$

$\mathbf{X}$ is called a design matrix :

the jth row of $\mathbf{X}$ is denoted by $\mathbf{x}_j^T$, and

the kth column of $\mathbf{X}$ is denoted by $\mathbf{x}^{(k)}$

ie $\mathbf{X} = (x_j^{(k)})$, $j \in \{1, \dots, n\}$ and $k \in \{1, \dots, p\}$