Regression

Adopted from 'Statistics for Business and Economics'

Models

- Representation of some phenomenon
- Mathematical model is a mathematical expression of some phenomenon
- Often describe relationships between variables
- Types
  - Deterministic models
  - Probabilistic models

Deterministic Models

- Hypothesize exact relationships
- Suitable when prediction error is negligible
- Example: force is exactly mass times acceleration
  - $F = m \cdot a$

Probabilistic Models

Hypothesize two components

Deterministic

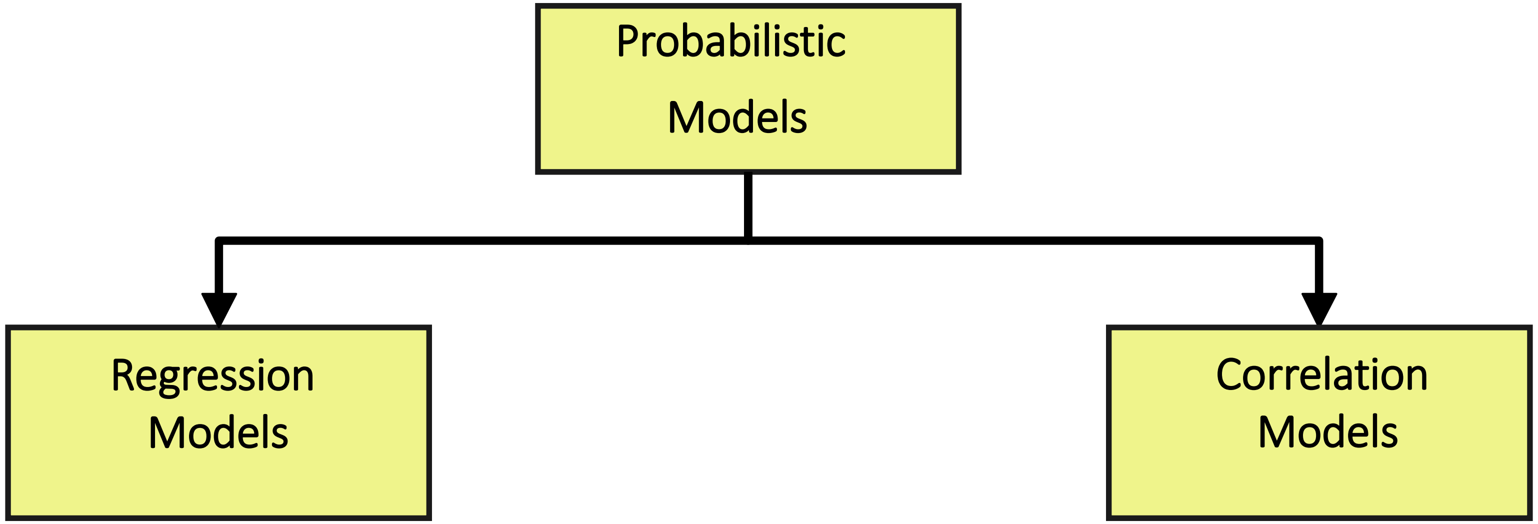
Random error

Example: sales volume (*y*) is 10 times advertising spending (*x*) + random error

$$y = 10x + \varepsilon$$

Random error may be due to factors other than advertising

Types


Probabilistic
Models
Regression
Models
Correlation
Models

Regression Models

- Answers ‘What is the relationship between the variables?’
- Equation used
  - One numerical dependent (response) variable
    - What is to be predicted
  - One or more numerical or categorical independent (explanatory) variables
- Used mainly for prediction and estimation

Steps

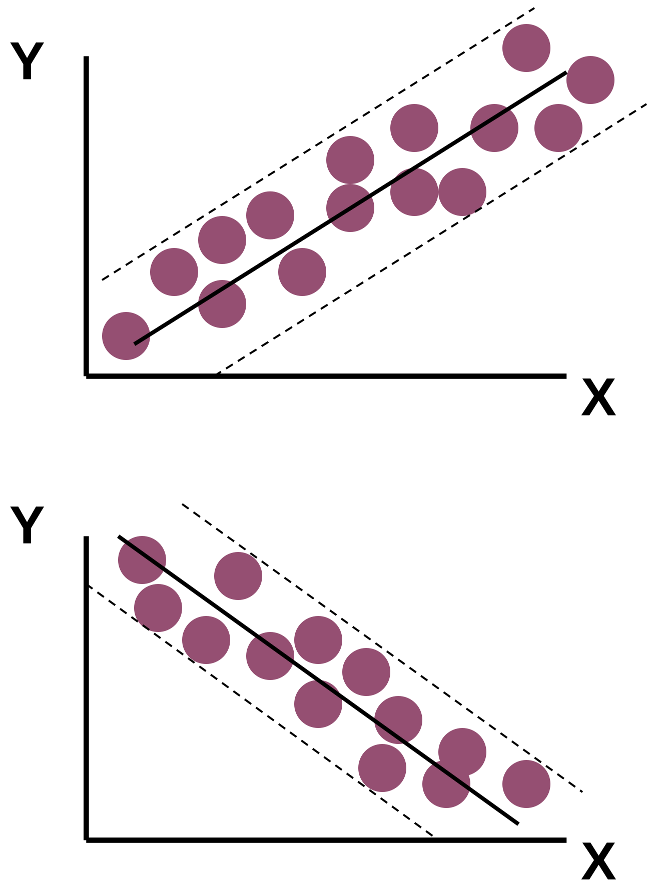
1. Hypothesize deterministic component
2. Estimate unknown model parameters
3. Specify probability distribution of random error term
    - Estimate standard deviation of error
4. Evaluate model
5. Use model for prediction and estimation

Specifying the Model

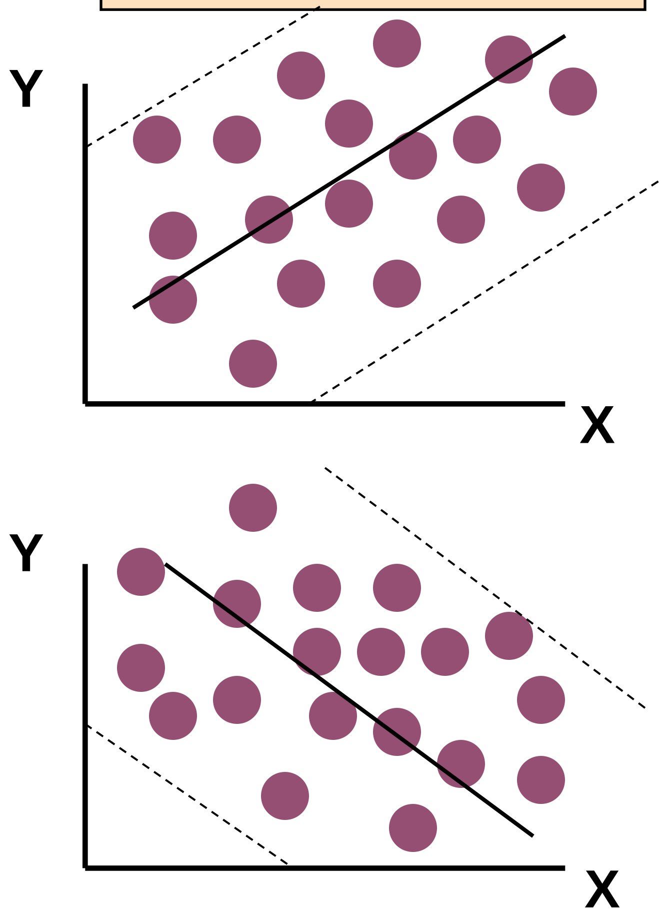
1. Define variables
    1. Conceptual (e.g., Advertising, price)
    2. Empirical (e.g., List price, regular price)
    3. Measurement (e.g., $, Units)
2. Hypothesize nature of relationship
    1. Expected effects (i.e., Coefficients' signs)
    2. Functional form (linear or non-linear)
    3. Interactions

Relationships

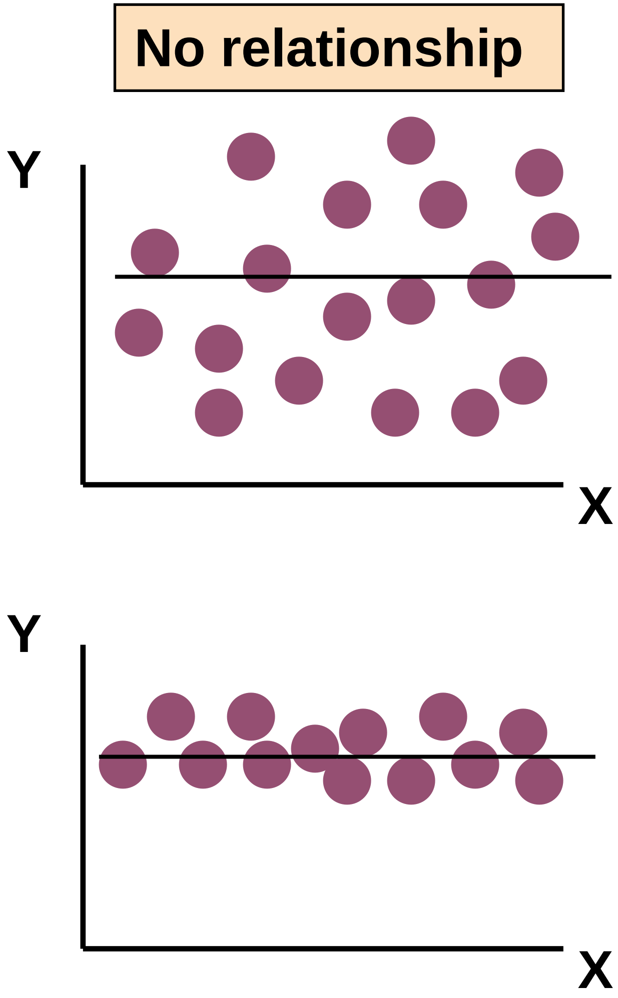
Strong relationships



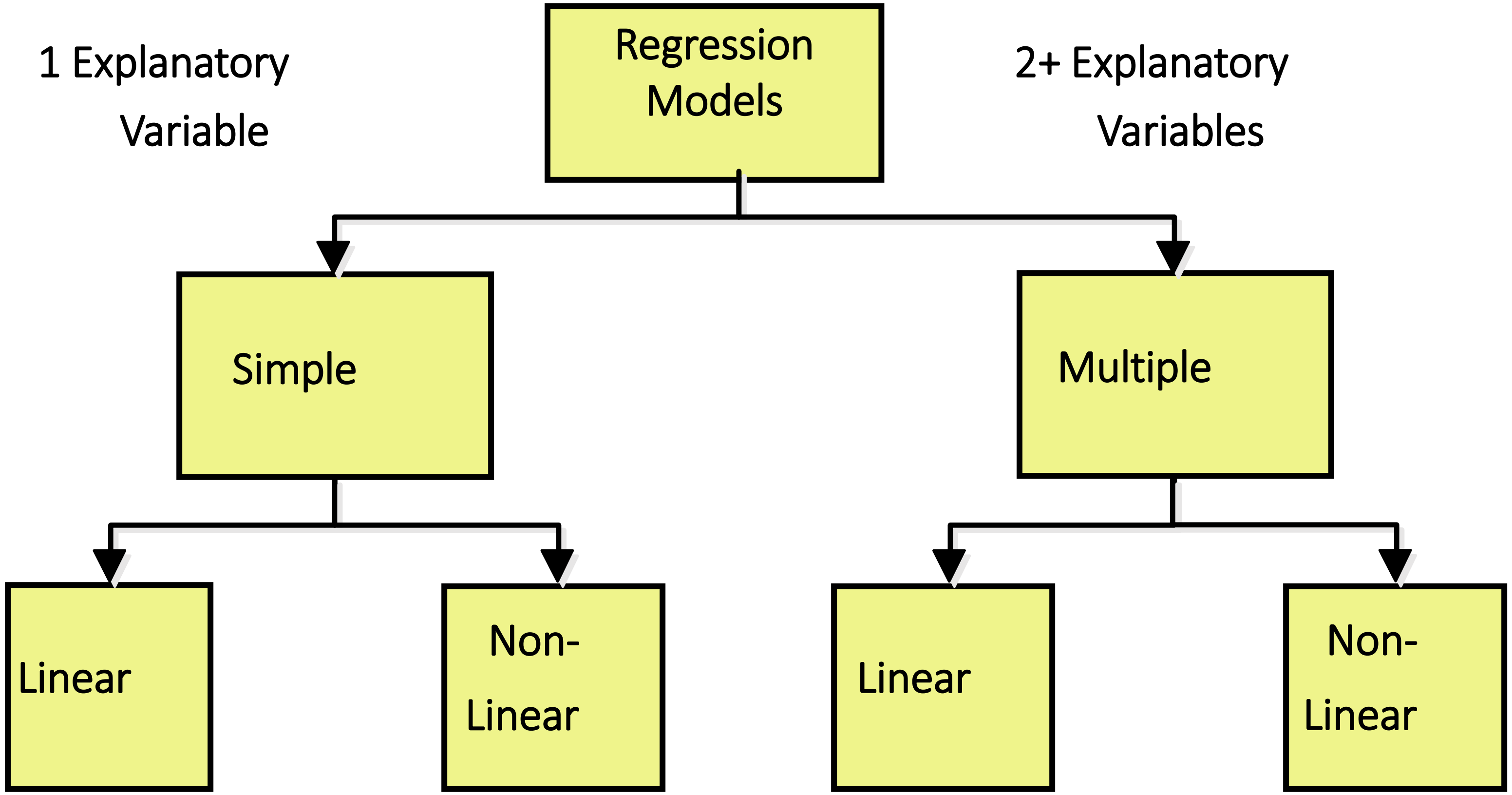
Weak relationships



Relationships



Types of Regression Models



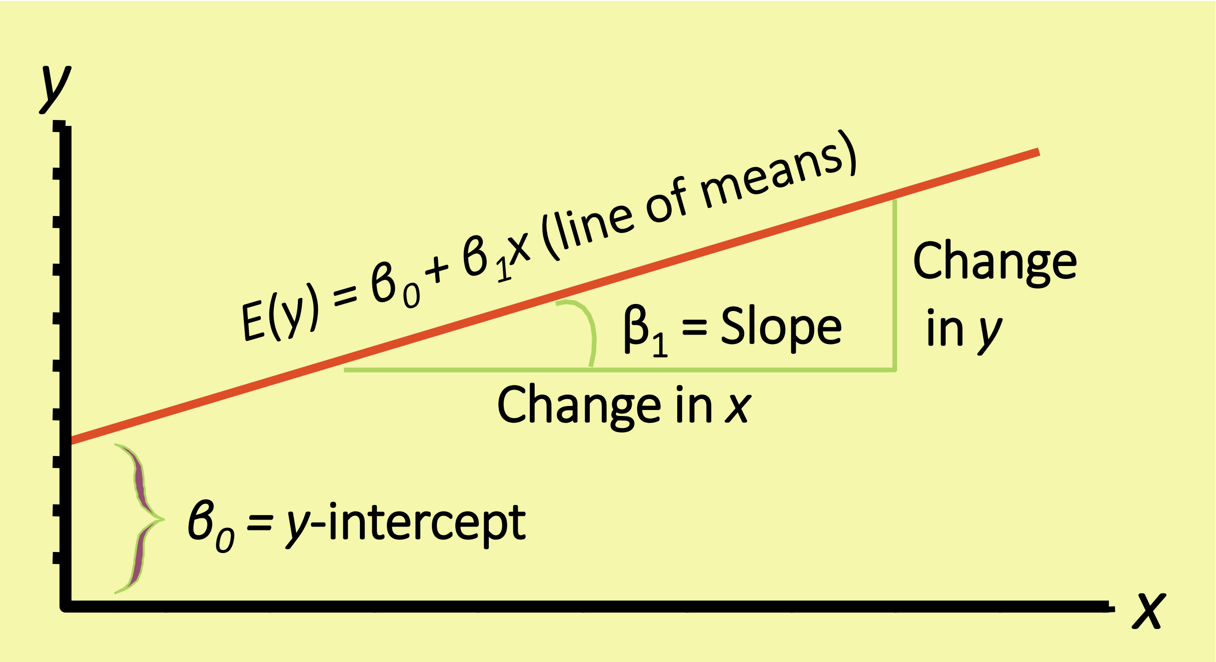
The Model

Relationship between variables is a linear function

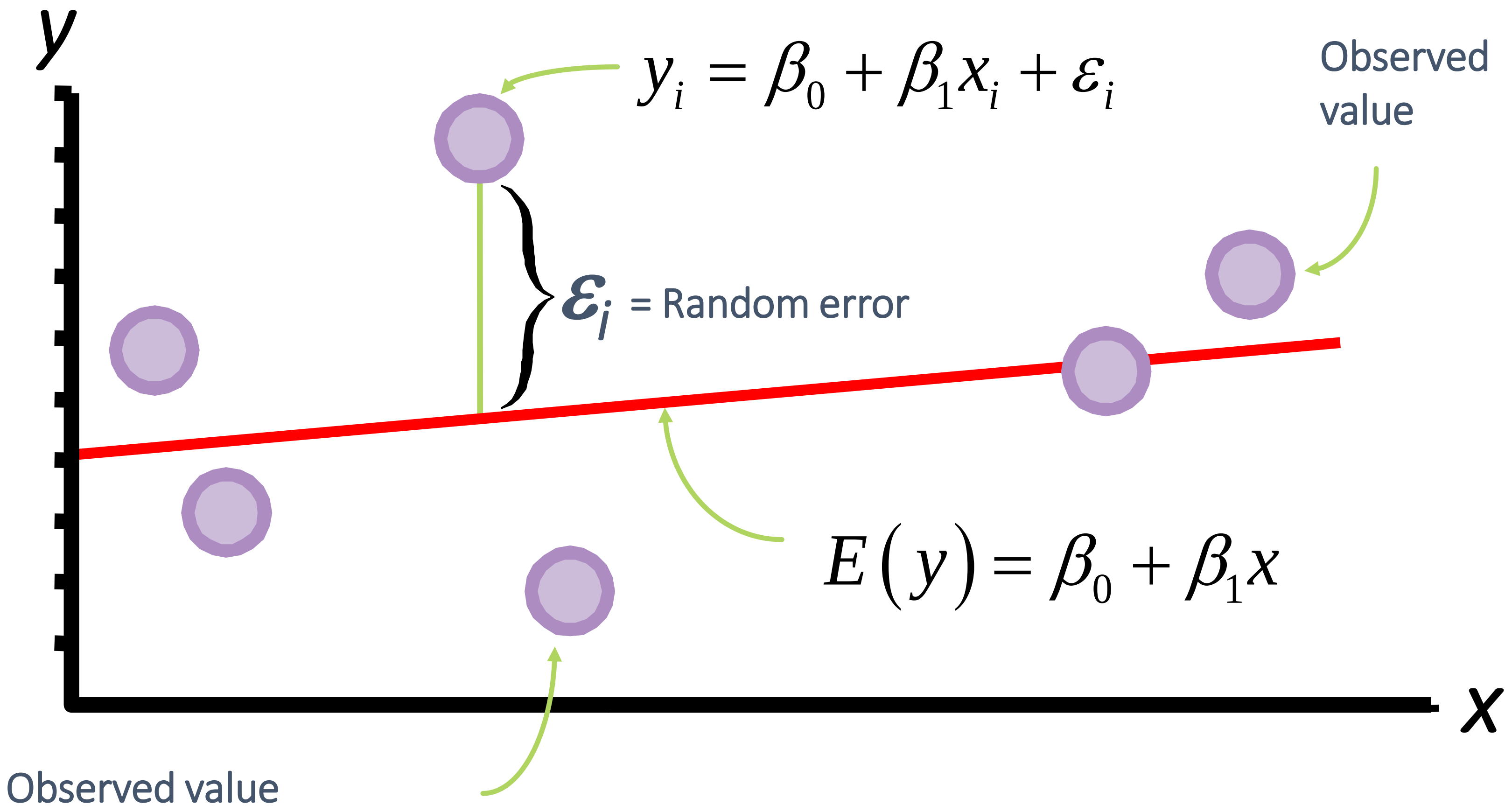
$$y = \beta_0 + \beta_1 x + \varepsilon$$

- y: Dependent (Response) Variable
- β_0: Population *y*-intercept
- β_1: Population Slope
- x: Independent (Explanatory) Variable
- ε: Random Error

The Model

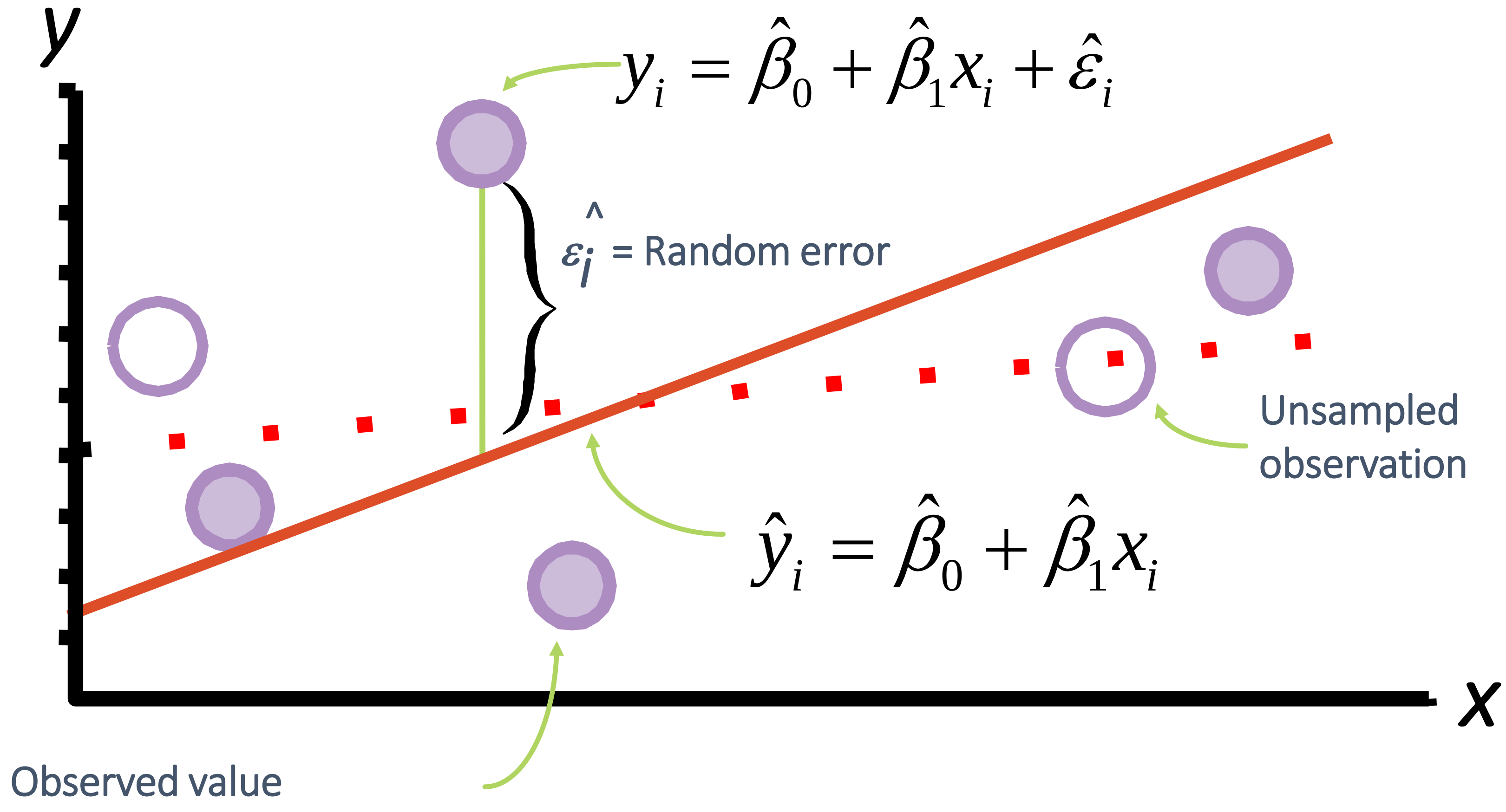

y
E(y) = β0 + β1x (line of means)
β1 = Slope
Change
in y
Change in x
β0 = y-intercept
x

The Model



y
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
Observed value
ε_i = Random error
$E(y) = \beta_0 + \beta_1 x$
x
Observed value

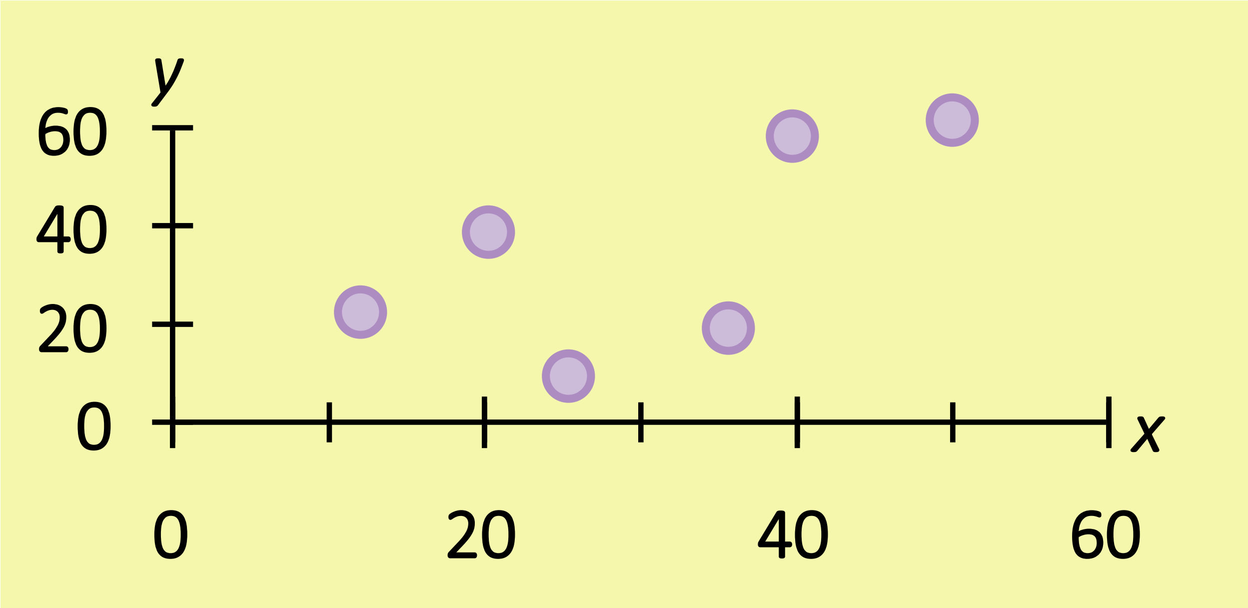
The Model



y
$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\varepsilon}_i$
$\hat{\varepsilon}_i$ = Random error
Unsampled observation
$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
x
Observed value

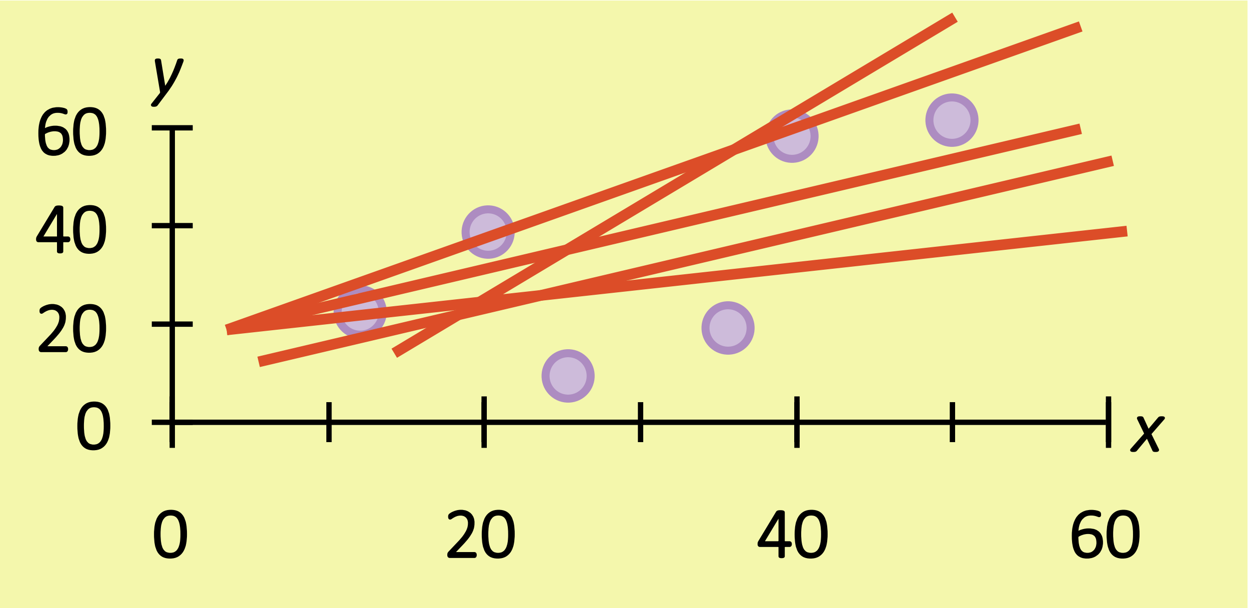
Estimating Parameters

1. Plot of all (x_i, y_i) pairs
2. Suggests how well model will fit



Estimating Parameters

- How would you draw a line through the points?
- How do you determine which line 'fits best'?



Estimating Parameters

'Best fit' means difference between actual *y* values and predicted *y* values are a minimum

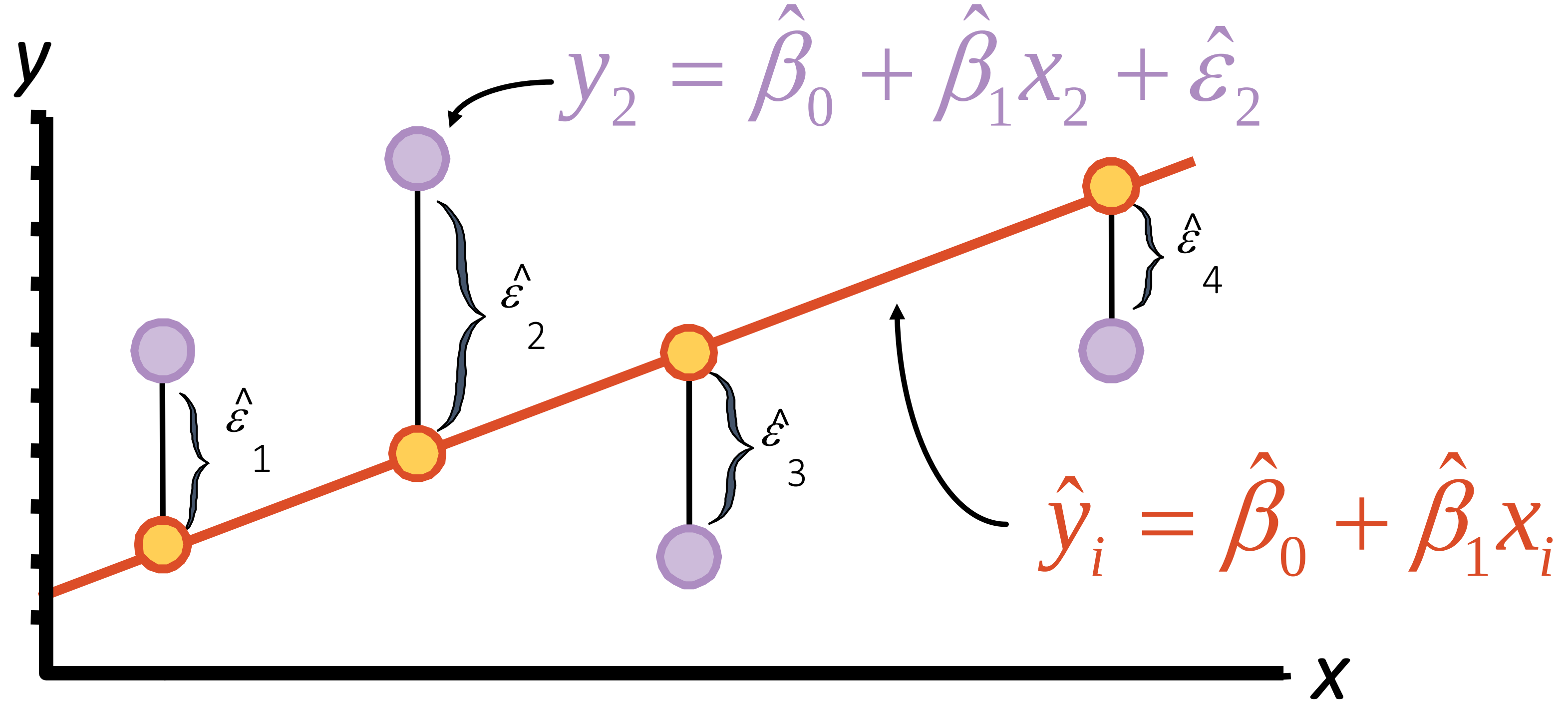
But positive differences off-set negative

$$\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

Least Squares minimizes the Sum of the Squared Differences (SSE)

Estimating Parameters

$$\text{LS minimizes } \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2 + \hat{\varepsilon}_4^2$$



Estimating Parameters

Prediction Equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Slope $\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$

y-intercept $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Calculations

$$f\,(a,\,b) = a + b\,x,$$

$$R^2 \equiv \sum [y_i - f\,(x_i,\,a_1,\,a_2,\,\ldots,\,a_n)]^2$$

$$R^2\,(a,\,b) \equiv \sum_{i=1}^{n} [y_i - (a + b\,x_i)]^2$$

$$\frac{\partial (R^2)}{\partial a_i} = 0$$

$$\frac{\partial (R^2)}{\partial a} = -2 \sum_{i=1}^{n} [y_i - (a + b\,x_i)] = 0$$

$$\frac{\partial (R^2)}{\partial b} = -2 \sum_{i=1}^{n} [y_i - (a + b\,x_i)]\,x_i = 0.$$

Calculations

$$\frac{\partial(R^2)}{\partial a} = -2\sum_{i=1}^{n}[y_i - (a + b\,x_i)] = 0$$

$$\frac{\partial(R^2)}{\partial b} = -2\sum_{i=1}^{n}[y_i - (a + b\,x_i)]\,x_i = 0.$$

$$n\,a + b\sum_{i=1}^{n} x_i$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$$

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\,y_i \end{bmatrix},$$

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\,y_i \end{bmatrix}.$$

Calculations

$$\mathsf{A} \equiv \begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

$$\mathsf{A}^{-1} = \frac{1}{|\mathsf{A}|} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

$$= \frac{1}{a\,d - b\,c} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\, y_i \end{bmatrix}.$$

$$\begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{bmatrix} \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i\, y_i \\ n \sum_{i=1}^{n} x_i\, y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \end{bmatrix},$$

Calculations

$$\begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{bmatrix} \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i\, y_i \\ n\sum_{i=1}^{n} x_i\, y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \end{bmatrix},$$

$$\begin{aligned} a &= \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i\, y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\ &= \frac{\bar{y}\left(\sum_{i=1}^{n} x_i^2\right) - \bar{x}\sum_{i=1}^{n} x_i\, y_i}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2} \\ b &= \frac{n\sum_{i=1}^{n} x_i\, y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\ &= \frac{\left(\sum_{i=1}^{n} x_i\, y_i\right) - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2} \end{aligned}$$

Calculations

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
x_1	y_1	x_1^2	y_1^2	$x_1 y_1$
x_2	y_2	x_2^2	y_2^2	$x_2 y_2$
:	:	:	:	:
x_n	y_n	x_n^2	y_n^2	$x_n y_n$
Σx_i	Σy_i	Σx_i^2	Σy_i^2	$\Sigma x_i y_i$

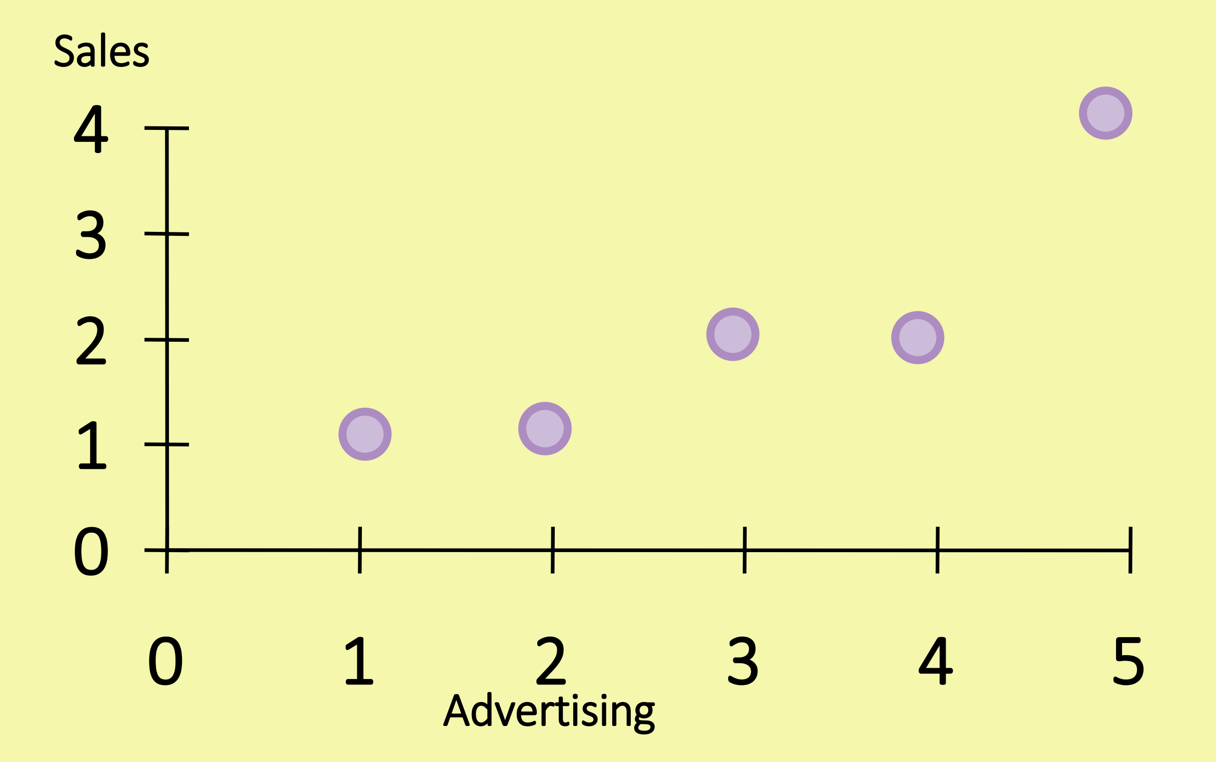
Example

You gather the following data:

Ad $	Sales (Units)
1	1
2	1
3	2
4	2
5	4

Find the least squares line relating sales and advertising

Example



Example

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	1	1	1	1
2	1	4	1	2
3	2	9	4	6
4	2	16	4	8
5	4	25	16	20
15	10	55	26	37

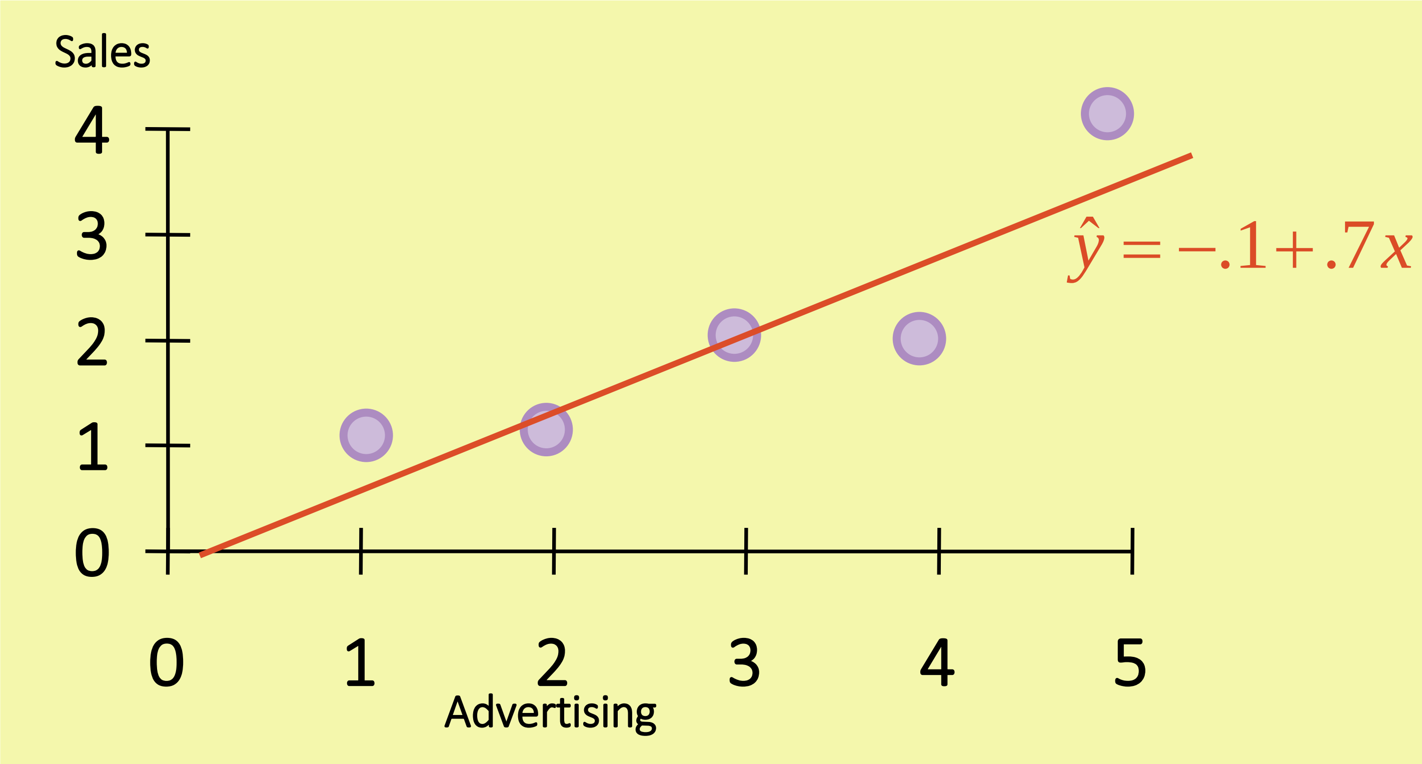
Example

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{37 - \frac{(15)(10)}{5}}{55 - \frac{(15)^2}{5}} = .70$$

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x} = 2 - (.70)(3) = -.10$$

$$\hat{y} = -.1 + .7x$$

Example



Bayesian Learning

Adopted from 'Data Mining Concepts and Techniques'

Introduction

- A statistical classifier: performs *probabilistic prediction, i.e.,* predicts class membership probabilities
- Foundation: Based on Bayes' Theorem.
- Performance: A simple Bayesian classifier, *naïve Bayesian classifier*, has comparable performance with decision tree and selected neural network classifiers
- Incremental: Each training example can incrementally increase/decrease the probability that a hypothesis is correct — prior knowledge can be combined with observed data
- Standard: Even when Bayesian methods are computationally intractable, they can provide a standard of optimal decision making against which other methods can be measured

The Model

A good strategy is to predict:

$$\arg\max_{Y} P(Y|X_1, \ldots, X_n)$$

(for example: what is the probability that the image represents a 5 given its pixels?)

The Model

Total probability Theorem: $P(B) = \sum_{i=1}^{M} P(B|A_i)P(A_i)$

Bayes' Theorem: $P(H|\mathbf{X}) = \frac{P(\mathbf{X}|H)P(H)}{P(\mathbf{X})} = P(\mathbf{X}|H) \times P(H) / P(\mathbf{X})$

- Let X be a data sample ("*evidence*"): class label is unknown
- Let H be a *hypothesis* that X belongs to class C
- Classification is to determine P(H|X), (i.e., *posteriori probability):* the probability that the hypothesis holds given the observed data sample X
- P(H) (*prior probability*): the initial probability
  - E.g., X will buy computer, regardless of age, income, ...
- P(X): probability that sample data is observed
- P(X|H) (likelihood): the probability of observing the sample X, given that the hypothesis holds

  E.g., Given that X will buy computer, the prob. that X is 31..40, medium income

The Model

- Given training data **X**, *posteriori probability of a hypothesis* H, P(H|**X**), follows the Bayes' theorem

$$P(H|\mathbf{X}) = \frac{P(\mathbf{X}|H)P(H)}{P(\mathbf{X})} = P(\mathbf{X}|H) \times P(H) / P(\mathbf{X})$$

- Informally, this can be viewed as

  posteriori = likelihood x prior/evidence

- Predicts **X** belongs to C_i iff the probability $P(C_i|\mathbf{X})$ is the highest among all the $P(C_k|X)$ for all the *k* classes
- Practical difficulty: It requires initial knowledge of many probabilities, involving significant computational cost

The Model

- Let D be a training set of tuples and their associated class labels, and each tuple is represented by an n-D attribute vector **X** = (x_1, x_2, …, x_n)
- Suppose there are *m* classes C_1, C_2, …, C_m.
- Classification is to derive the maximum posteriori, i.e., the maximal $P(C_i|\mathbf{X})$
- This can be derived from Bayes' theorem

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})}$$

- Since P(X) is constant for all classes, only
    - needs to be maximized $P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i)$

The Model

- A simplified assumption: attributes are conditionally independent (i.e., no dependence relation between attributes):

$$P(\mathbf{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times ... \times P(x_n|C_i)$$

- This greatly reduces the computation cost: Only counts the class distribution
- If A_k is categorical, $P(x_k|C_i)$ is the # of tuples in C_i having value x_k for A_k divided by $|C_{i, D}|$ (# of tuples of C_i in D)
- If A_k is continous-valued, $P(x_k|C_i)$ is usually computed based on Gaussian distribution with a mean μ and standard deviation σ

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

and $P(x_k|C_i)$ is

$$P(\mathbf{X}|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i})$$

Example

Class:

C1:buys_computer = 'yes'

C2:buys_computer = 'no'

Data to be classified:

X = (age <=30,

Income = medium,

Student = yes

Credit_rating = Fair)

age	income	student	credit_rating	comp
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Example

age	income	student	credit_rating	_comp
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

$P(C_i)$: P(buys_computer = “yes”) = 9/14 = 0.643

P(buys_computer = “no”) = 5/14= 0.357

Compute $P(X|C_i)$ for each class

P(age = “<=30” | buys_computer = “yes”) = 2/9 = 0.222

P(age = “<= 30” | buys_computer = “no”) = 3/5 = 0.6

P(income = “medium” | buys_computer = “yes”) = 4/9 = 0.444

P(income = “medium” | buys_computer = “no”) = 2/5 = 0.4

P(student = “yes” | buys_computer = “yes) = 6/9 = 0.667

P(student = “yes” | buys_computer = “no”) = 1/5 = 0.2

P(credit_rating = “fair” | buys_computer = “yes”) = 6/9 = 0.667

P(credit_rating = “fair” | buys_computer = “no”) = 2/5 = 0.4

X = (age <= 30 , income = medium, student = yes, credit_rating = fair)

$P(X|C_i)$: P(X|buys_computer = “yes”) = 0.222 x 0.444 x 0.667 x 0.667 = 0.044

P(X|buys_computer = “no”) = 0.6 x 0.4 x 0.2 x 0.4 = 0.019

$P(X|C_i)*P(C_i)$: P(X|buys_computer = “yes”) * P(buys_computer = “yes”) = 0.028

P(X|buys_computer = “no”) * P(buys_computer = “no”) = 0.007

Therefore, X belongs to class (“buys_computer = yes”)

Avoiding Zero Probability

- Naïve Bayesian prediction requires each conditional prob. be **non-zero**. Otherwise, the predicted prob. will be zero
- $$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i)$$
- Ex. Suppose a dataset with 1000 tuples, income=low (0), income= medium (990), and income = high (10)
- Use Laplacian correction (or Laplacian estimator)
  - *Adding 1 to each case*
    - Prob(income = low) = 1/1003
    - Prob(income = medium) = 991/1003
    - Prob(income = high) = 11/1003
  - The "corrected" prob. estimates are close to their "uncorrected" counterparts

Comments

- Advantages
  - Easy to implement
  - Good results obtained in most of the cases
- Disadvantages
  - Assumption: class conditional independence, therefore loss of accuracy
  - Practically, dependencies exist among variables
    - E.g., hospitals: patients: Profile: age, family history, etc.
      - Symptoms: fever, cough etc., Disease: lung cancer, diabetes, etc.
    - Dependencies among these cannot be modeled by Naïve Bayes Classifier