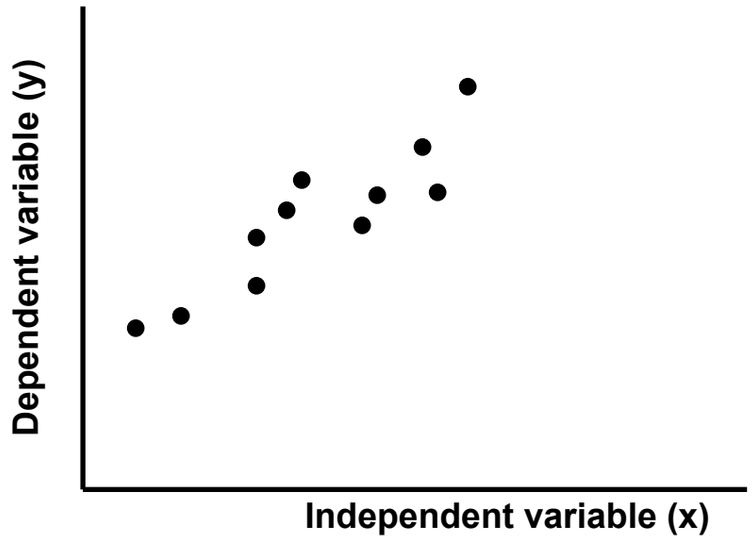
cse 891 - regression and time series prediction

Arend Hintze

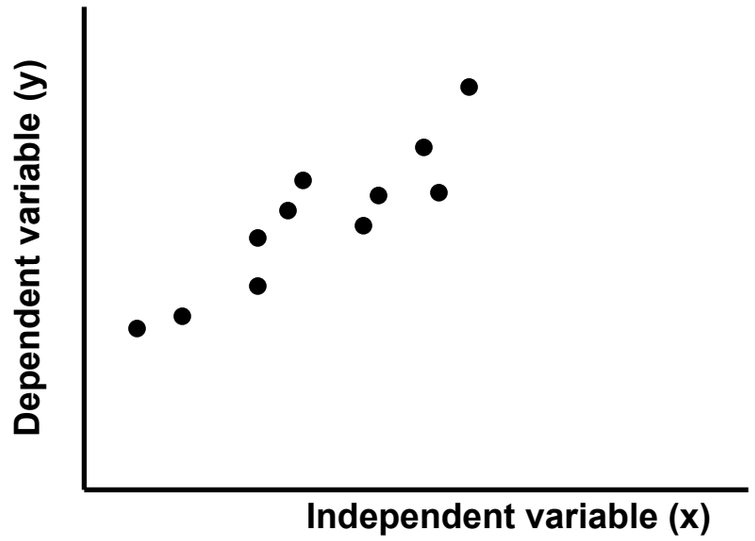
Definition

- given a collection of instances
- where each instance is a tuple (X,Y)
- where x is the attribute
- where y is the variable
- find a function f which maps each attribute x to it's corresponding value y

Regression



Regression



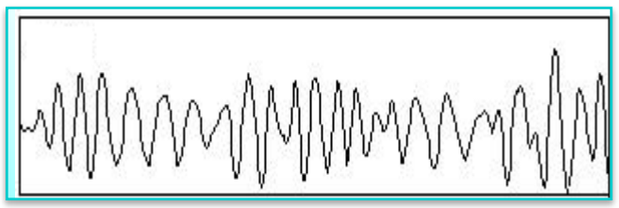
- tries to explain variability in dependent variable Y by the variability on the independent variable X

- if the model sufficiently explains X -> Y, it can be used for prediction

Time Series Forecasting

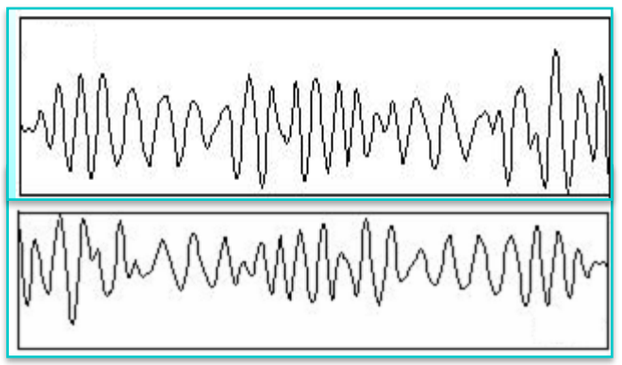
- a time series is a sequence of continuous-values observations ordered by time

- two categories: Univariate vs. Multivariate



$t_0\, t_1\, t_2$ t_N

Univariate: given the values of the time series $y_0, y_1, y_2, \ldots, y_N$ predict the value of y_{N+1}



$t_0\, t_1\, t_2$ t_N

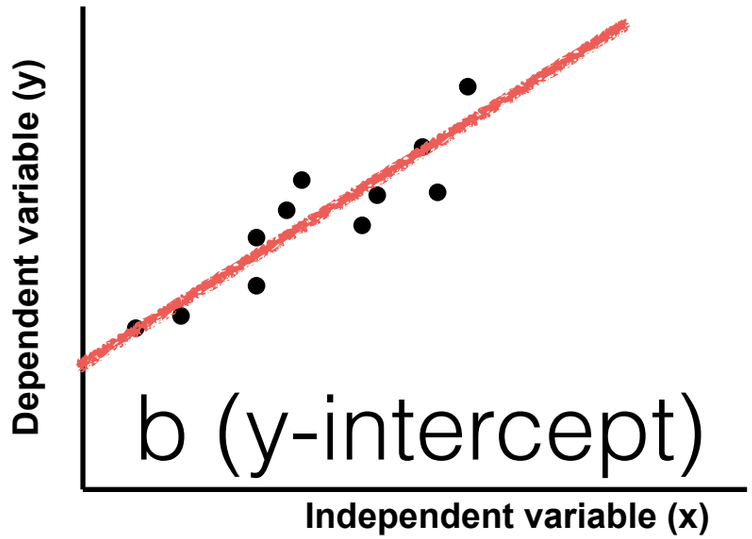
Multivariate: given the values of the target time series $y_0, y_1, y_2, \ldots, y_N$ and other auxiliary time series $x_0, x_1, x_2, \ldots, x_N, x_{N+1}$ predict the value of y_{N+1}

Application examples

Task	Attribute set, $\boldsymbol{x}$	Target variable, y
Forecasting the monthly sales	Historical monthly sales and other predictor variables (inventory, etc)	Monthly sales at time t
Predicting power consumption at data centers	Sensor measurements of temperature, fan speed, etc	Expected power consumed
Predicting crime rate	Statistics about housing, population, job/income, education, etc	Crime rate in a given city or region

linear regression

$$a = \frac{Y_2 - Y_1}{X_2 - X_1}$$



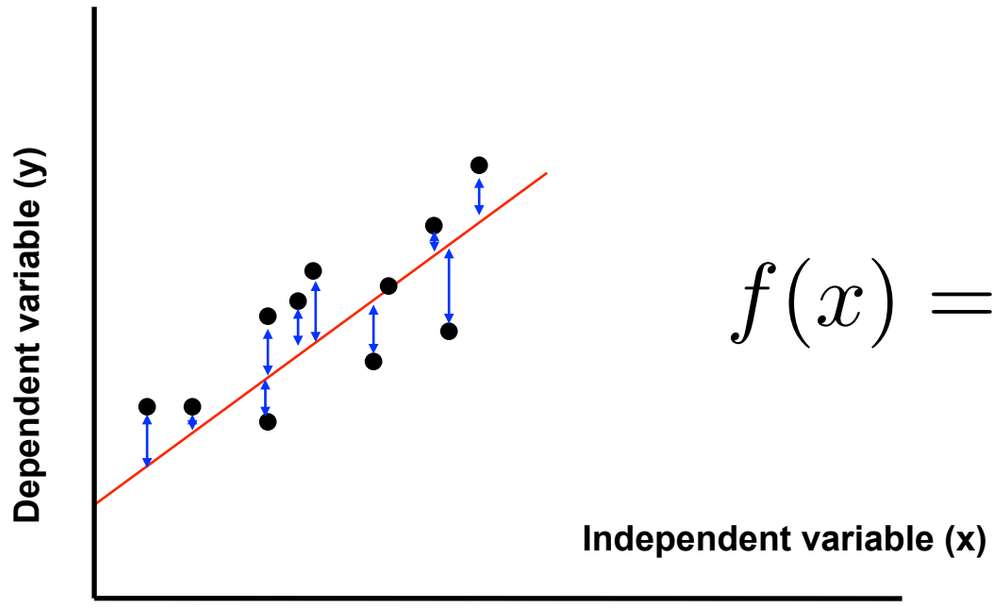
a (slope)

$$f(x) = W_0 + W_1 x$$

$$f(x) = ax + b$$

- Simple: there is only 1 independent variable
- Linear: Fit is a straight line
- Question: how to find a and b?

Least Square Method

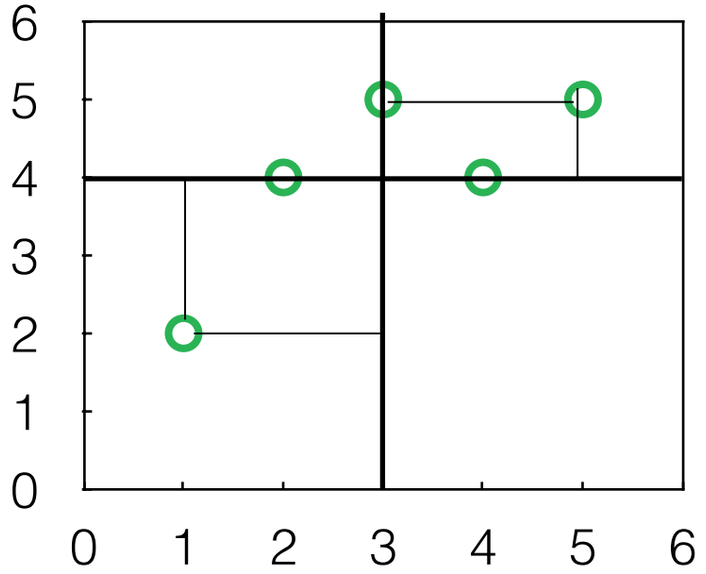


$$f(x) = ax + b$$

$$err = \sum_{i=0}^{N} (y_i - ax_i - b)^2$$

$$err = \sum_{i=0}^{N} (y_i - f(x_i))^2$$

step by step



X	Y
1	2
2	4
3	5
4	4
5	5

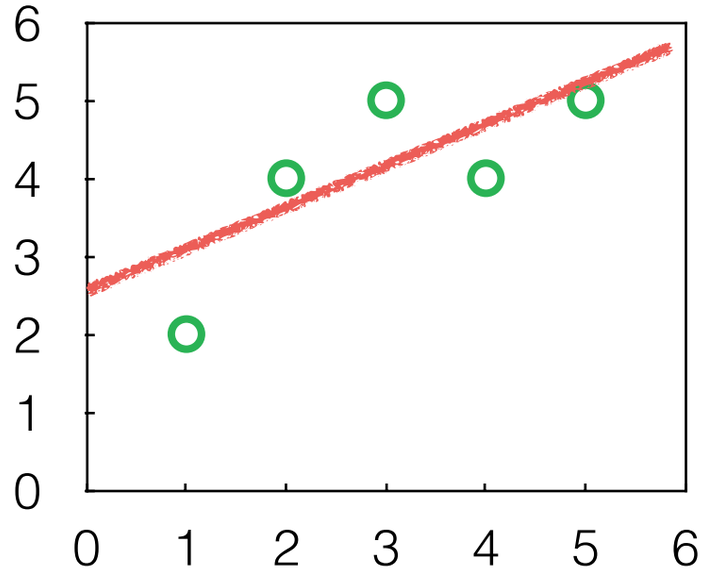
$$f(x) = ax + b$$

$$a = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})} = \frac{6}{10}$$

$$b = \bar{Y} - (a\bar{X})$$

http://www.youtube.com/watch?v=JvS2triCgOY

how good is the fit?



Regression sum of squares is the amount of variability explained by the model

- just guessing -> $f(x) = \bar{Y}$
- total sum of squares = $TSS = \sum_{i=1}^{N} (y_i - \bar{y})^2$
- sum of squares errors = $SSE = \sum_{i=1}^{N} (y_i - f(x_i))^2$
- SSE <= TSS
- regression sum of squares TSS=SSE +SSR $SSR = \sum_{i=1}^{N} (f(x_i) - \bar{y})^2$

goodness of fit

$$SSE = \sum_{i=1}^{N} \left(y_i - f(x_i)\right)^2$$ Residual sum of squared errors

$$SST = \sum_{i=1}^{N} \left(y_i - \bar{y}\right)^2$$ Sample variance

$$SSR = \sum_{i=1}^{N} \left(f(x_i) - \bar{y}\right)^2$$ Variability explained by the model

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

The higher R^2, the better the model fits the data

http://zunzun.com/