# MA2252 Introduction to computing

## lectures 21-22

### Least squares approximation

Matias Ruiz

November 2023

# Regression

**Regression** is a statistical technique used to find the 'best-fit curve' that describes a scatter plot.

Depending on the data trend in a scatter plot, one may use

- Linear Regression curve
- Non-linear Regression curve

## Linear Regression curve example

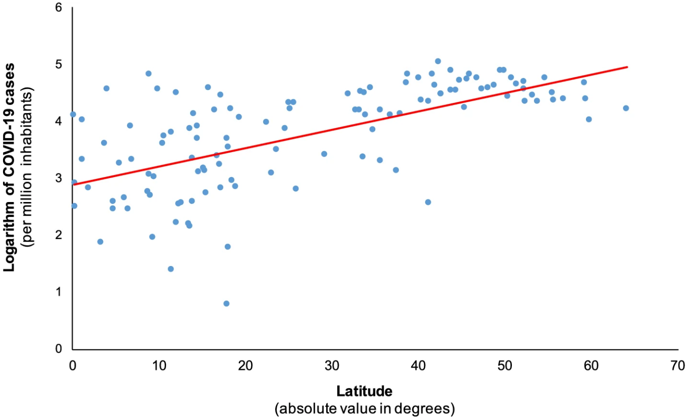


Figure: Scatter plot showing linear trend [1]

## Non-linear Regression curve example

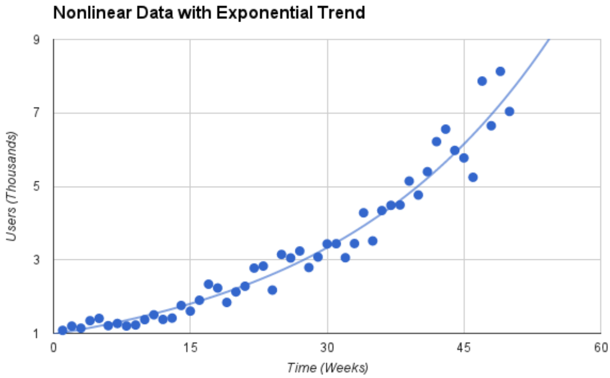


Figure: Data showing number of active users on a website with time [2]

[1] Chen, S., Prettner, K., Kuhn, M. et al. Climate and the spread of COVID-19. Sci Rep 11, 9042 (2021). https://doi.org/10.1038/s41598-021-87692-z

[2] http://sam-koblenski.blogspot.com

# Regression model

A regression model provides a function to describe the relationship between one (or more) independent variables and a dependent variable.

A basic regression model is the 'Least Squares Regression model'.

# Least Squares Regression

Here, the relationship between dependent data points $y_i (i = 1, 2, ...m)$ and independent data points $x_i$ is modelled as

$$\hat{y}(x) = \sum_{i=1}^{n} \alpha_i f_i(x) \tag{1}$$

where

- $\hat{y}(x)$ is an estimation function
- $\alpha_i$ are parameters of estimation function
- $f_i(x)$ are linearly independent basis functions

# Least Squares Regression (contd.)

The parameters are then found by minimising the total squared error E.

$$E = \sum_{i=1}^{m} (\hat{y} - y_i)^2 \tag{2}$$

Substituting (1) in (2) gives

$$E = \sum_{i=1}^{m} \left( \sum_{j=1}^{n} \alpha_j f_j(x_i) - y_i \right)^2 \tag{3}$$

E is a function of $n$ variables namely $\alpha_j (j = 1, 2, \cdots, n)$.

# Least Squares Regression (contd.)

The solution for $n$ parameters $\alpha_j$ which minimise the total squared error E is given as

$$\beta = pinv(A) * Y \tag{4}$$

Here,

- $\beta$ is a column vector with $n$ entries $\alpha_j$
- $A$ is a $m$x$n$ matrix with entries $A(i, j) = f_j(x_i)$
- $pinv(A)$ is the pseudo-inverse of $A$
- $Y$ is a column vector with $m$ entries $y_i$

# Nonlinear Estimation Functions

Sometimes, a nonlinear estimation function provides the best fit for a scatter plot. This means we require

$$\hat{y}(x) = g(\alpha_1, \alpha_2, \cdots, \alpha_n, x) \tag{5}$$

where $g$ is some nonlinear function.

In some special cases, a transformation such as

$$\tilde{y}(x) = h(\hat{y}(x)) \tag{6}$$

can linearise the equation (5) into (1).

# Nonlinear Estimation Functions (contd.)

**Example:** Consider the estimation function

$$\hat{y}(x) = \alpha_1 e^{\alpha_2 x} \tag{7}$$

Applying the transformation

$$\tilde{y}(x) = log(\hat{y}(x)) \tag{8}$$

converts (7) into

$$\tilde{y}(x) = \tilde{\alpha_1} + \alpha_2 x \tag{9}$$

where we define $\tilde{\alpha_1} = log(\alpha_1)$. Now, least squares regression can be applied to equation (9).The parameter $\alpha_1$ can be found using $\alpha_1 = e^{\tilde{\alpha_1}}$.