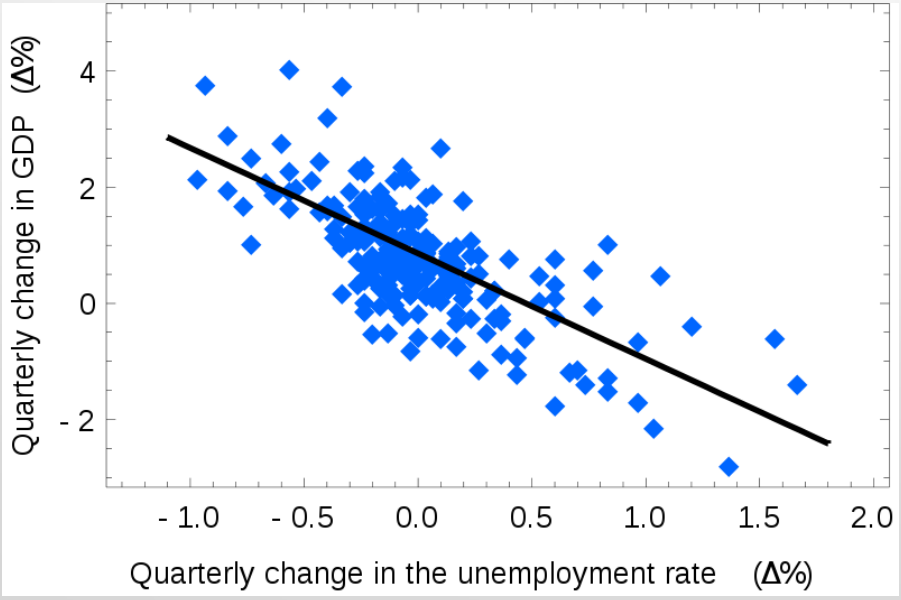
# Intro to Linear Regression And Modelling in R

# Linear Regression and Modelling

1. **Modelling**
   - **find causal relationship between dependent variables Y and independent variables X**
2. **Linear Regression**
   - **construct a model from *observations* using *linear predictor functions* - which can be nonlinear, also shows up in Neural Nets**



# Linear Regression and Modelling

1. **Theory**
   - **Find constants (or parameters) $\beta$ that minimize the sum of squared**

***residuals*** $e_i = y_i - (\beta_0 + f(x_i)_1 \beta_1 + .. + f(x_i)_n \beta_n)$

$$S(b) = \sum_{i=1}^{n} (y_i - x_i^T b)^2 = (y - Xb)^T (y - Xb),$$

$$\hat{\beta} = \arg\min_{b \in \mathbb{R}^p} S(b) = \left( \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \right)^{-1} \cdot \frac{1}{n} \sum_{i=1}^{n} x_i y_i$$

3. **Single variable simple linear regression**

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

$$\hat{\beta} = \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n} (\sum x_i)^2} = \frac{\operatorname{Cov}[x, y]}{\sigma_x^2}, \quad \hat{\alpha} = \overline{y} - \hat{\beta}\, \overline{x}\ .$$

# Linear Regression and Modelling

## 1. Assumptions – validity

**Linear regression is applicable to data satisfies the following:**

I. **Complete (no missing entries)**
II. **Variance of the underlying data does not change across observations**
III. **Observations and Residuals are uncorrelated with each other Residuals have normal distribution**
IV. **Problem is not ill posed numerically ill conditioned**
V. **Outliers dealt with**

**More advanced methods (weighted least squares, regularization, non-parametric) can handle some of these conditions**

## 2. Performance measures

I. **$R^2$ – the coefficient of determination**
II. **Statistical significance**

**Later we'll see how the lm() function in R builds linear models and their performance measures**

# Matrix Simple Linear Regression

- Nothing new-only matrix formalism for previous results
- Remember the normal error regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad i = 1, ..., n$$

- Expanded out this looks like

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 + \epsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_2 + \epsilon_2 \\ &\cdots \\ Y_n &= \beta_0 + \beta_1 X_n + \epsilon_n \end{aligned}$$

- which points towards an obvious matrix formulation.

Slides courtesy of Frank Wood

# Regression Matrices

- If we identify the following matrices

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \cdot & \\ \cdot & \\ \cdot & \\ 1 & X_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \epsilon_n \end{pmatrix}$$

- We can write the linear regression equations in a compact form $\mathbf{y} = \mathbf{X}\beta + \epsilon$

# Regression Matrices

- Of course, in the normal regression model the expected value of each of the $\epsilon$'s is zero, we can write $\mathbb{E}(\mathbf{y}) = \mathbf{X}\beta$
- This is because

$$\mathbb{E}(\epsilon) = \mathbf{0}$$

$$\begin{pmatrix} \mathbb{E}(\epsilon_1) \\ \mathbb{E}(\epsilon_2) \\ . \\ . \\ . \\ \mathbb{E}(\epsilon_n) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ . \\ . \\ . \\ 0 \end{pmatrix}$$

# Error Covariance

Because the error terms are independent and have constant variance $\sigma^2$

$$\sigma^2\{\epsilon\} = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \dots & & & \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix}$$

$$\sigma^2\{\epsilon\} = \sigma^2 \mathbf{I}$$

# Matrix Normal Regression Model

In matrix terms the normal regression model can be written as

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

where $\mathbb{E}(\epsilon) = \mathbf{0}$ and $\sigma^2\{\epsilon\} = \sigma^2\mathbf{I}$, i.e. $\epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$

# Least Square Estimation

If we remember both the starting normal equations that we derived

$$nb_0 + b_1 \sum X_i = \sum Y_i$$
$$b_0 \sum X_i + b_1 \sum X_i^2 = \sum X_i Y_i$$

and the fact that

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ X_1 & X_1 & \dots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ . & \\ . & \\ . & \\ 1 & X_n \end{bmatrix} = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ X_1 & X_1 & \dots & X_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix}$$

# Least Square Estimation

Then we can see that these equations are equivalent to the following matrix operations

$$\mathbf{X}'\mathbf{X}\ \mathbf{b} = \mathbf{X}'\mathbf{y}$$

with

$$\mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$$

with the solution to this equation given by

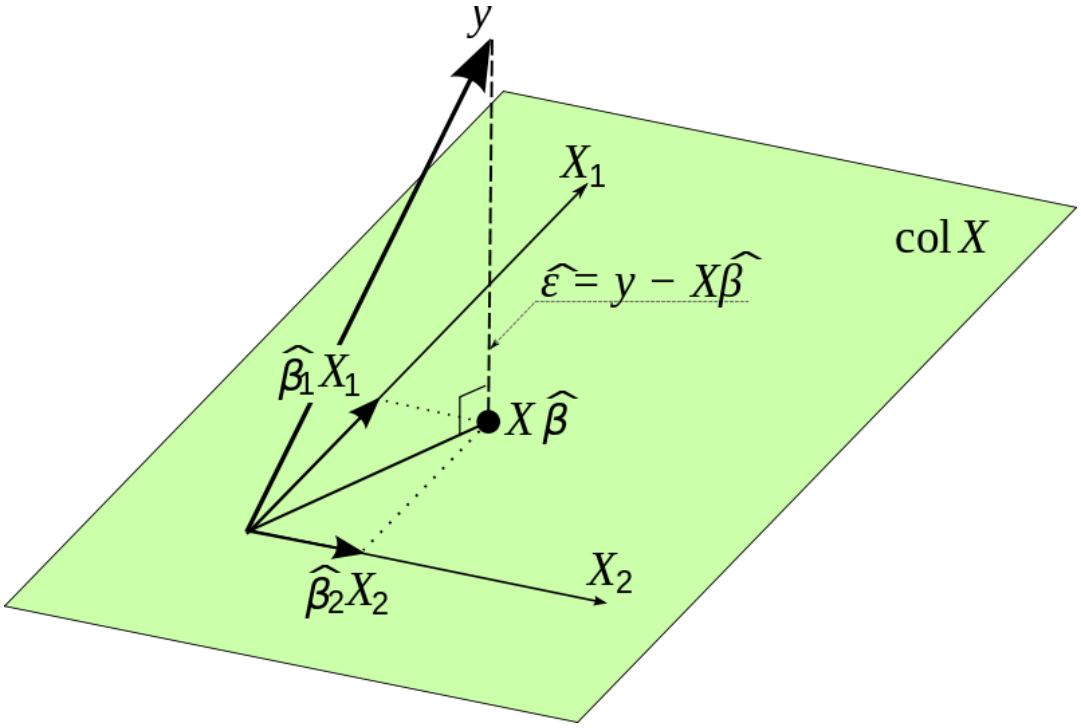
$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

when $(\mathbf{X}'\mathbf{X})^{-1}$ exists.

# When does $(\mathbf{X}'\mathbf{X})^{-1}$ exist?

$\mathbf{X}$ is an $n \times p$ (or $p+1$ depending on how you define $p$) design matrix.

$\mathbf{X}$ must have full column rank in order for the inverse to exist, i.e. $rank(\mathbf{X}) = p \implies (\mathbf{X}'\mathbf{X})^{-1}$ exists.



# Polynomial Regression

$$y_i = a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_m x_i^m + \varepsilon_i \ (i = 1, 2, \ldots, n)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^m \\ 1 & x_2 & x_2^2 & \ldots & x_2^m \\ 1 & x_3 & x_3^2 & \ldots & x_3^m \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^m \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\vec{y} = \mathbf{X}\vec{a} + \vec{\varepsilon}.$$

$$\widehat{\vec{a}} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\vec{y}.$$

## Example

Find a linear trend in a stock time series (requires package quantmod )

```
library(quantmod)
getSymbols("YHOO",src="google")

closeprice.vector <- as.vector(YHOO$YHOO.Close)
dates.vector <- sequence(length(closeprice.vector))

X = matrix( c(t(rep(1, length(closeprice.vector))) , t(dates.vector)),
        + length(closeprice.vector),
        + 2,
        + FALSE)

a = solve(t(X)%*%X)%*%t(X)%*%closeprice.vector
```

What does vector *a* contain? Can you predict the value of the stock one day, one week and one month from today? Do you think long term estimates are reliable? Does modelling time series violate any assumptions? Plot the stock with its line of best fit.

In general the *lm* function in R is used for least squares regression of linear models (linear as in it can be written in Matrix form, not as in whether you assume the regression function is a line, polynomial, exponential, log, etc.)

# The R lm() function

R has a function that will construct a linear model to specification and display the performance measures and statistical significance levels

tutorial is on: http://data.princeton.edu/R/readingData.html

another one: http://blog.yhathq.com/posts/r-lm-summary.html