Note: Although the model based on

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

is **not** linear in x, it **is** linear in the parameters. Because of this, we still term this a *linear model*. It is this fact that makes the least-squares solutions easy to find.

This model is no more difficult to fit than the model

$$y = \beta_0 + \beta_1 \frac{x}{1+x} + \beta_2 (1 - e^{-x})$$

say - it is still a *linear in the parameters model*. It is in the general class of models

$$y = \beta_0 + \beta_1 g_1(x) + \beta_2 g_2(x)$$

where $g_1(x)$ and $g_2(x)$ are general functions of x.

In fact, any model of the form

$$y = \sum_{j=0}^{k} \beta_j g_j(x) + \epsilon \tag{1}$$

can be fitted routinely using least-squares; if we know x, then we can compute

$$g_0(x), g_1(x), \ldots, g_k(x)$$

and plug those values into the formula (1).

Example: Harmonic Regression.

Let

$$
\begin{aligned}
g_0(x) &= 1 \\
g_1(x) &= \begin{cases} \cos(\lambda_j x) & j \text{ odd} \\ \sin(\lambda_j x) & j \text{ even} \end{cases}
\end{aligned}
$$

where k is an even number, $k = 2p$ say.

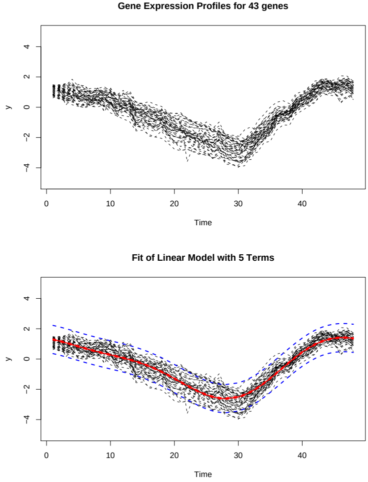
$\lambda_j, j = 1, 2, \ldots, p$ are constants

$$\lambda_1 < \lambda_2 < \cdots < \lambda_p$$

For fixed x, $\cos(\lambda_j x)$ and $\sin(\lambda_j x)$ are also fixed, known values.

Gene Expression Data Example

Harmonic Regression Fit with $p = 2$.



Why are things so straightforward ?

- because the system of equations based on the derivatives

$$\frac{\partial}{\partial \beta_j}\left\{SSE(\underset{\sim}{\beta})\right\} = 0 \qquad j = 0, 1, \ldots, k$$

can always be solved routinely, so we can always find $\widehat{\underset{\sim}{\beta}}$.

In the general model (1), simple formulae for

- $\widehat{\underset{\sim}{\beta}}$
- $s.e.(\widehat{\underset{\sim}{\beta}})$
- $\widehat{\sigma}^2$

can be found using a matrix formulation.

SEE HANDOUT - NOT EXAMINABLE !

Note: One-way ANOVA can be formulated in the form of model (1). Recall

- k independent groups
- means $\mu_1, \ldots, \mu_k$
- y_{ij} - jth observation in the ith group

Let

$$\begin{aligned} \beta_0 &= \mu_k \\ \beta_t &= \mu_t - \mu_k \qquad t = 1, 2, \ldots, k-1. \end{aligned}$$

Define new data $x_{ij}(t)$ where

$$x_{ij}(t) = \begin{cases} 1 & \text{if } t = i \\ 0 & \text{if } t \neq i \end{cases}$$

Then, using the linear regression formulation

$$y_{ij} = \beta_0 + \sum_{t=1}^{k-1} \beta_t x_{ij}(t) + \epsilon_{ij}.$$

For any i, j, $x_{ij}(t)$ is non-zero for only one value of t, when $t = i$.

We term this a regression on a *factor predictor*; it is clear that $\beta_0, \beta_1, \ldots, \beta_{k-1}$ can be estimated using least-squares.

This defines the link between

ANOVA

and

Linear Modelling

- they are essentially the SAME MODEL formulation.

This link extends to **ALL ANOVA** models; recall that we used the **General Linear Model** option in SPSS to fit two-way ANOVA.

2.2 Multiple Linear Regression

Multiple linear regression models model the variation in response y as a function of **more than one** independent variable.

Suppose we have variables

$$X_1, X_2, \ldots, X_k$$

recording different features of the experimental units. We wish to model response Y as a function of $X_1, X_2, \ldots, X_k$.

2.2.1 Multiple Linear Regression Models

Consider the model for datum i

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

where x_{ij} is the measured value of *covariate* j on experimental unit i. That is

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i$$

where the first two terms on the right hand side are the *systematic* or *deterministic* components, and the final term ϵ_i is the *random* component.

Example: $k = 2$.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

A three parameter model.

Note: We can also include *interaction* terms

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12}(x_{i1} . x_{i2}) + \epsilon_i$$

where

- The first two terms in x_{i1} and x_{i2} are **main effects**
- The third term in $(x_{i1} . x_{i2})$ is an **interaction**

This is a four parameter model.