1.017/1.010 Class 22
Linear Regression

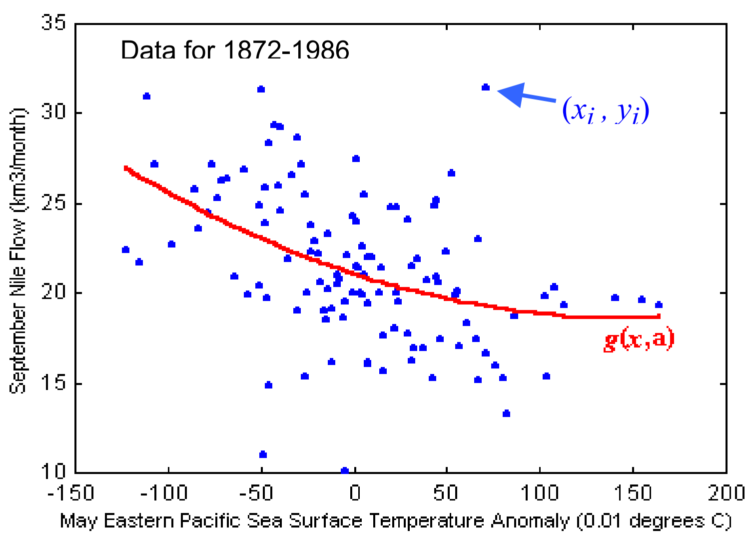
Regression Models

Fluctuations in measured (**dependent**) variables can often be attributed (in part) to other (**independent**) variables. ANOVA identifies likely independent variables. Regression methods quantify relationship between dependent and independent variables.

Consider problem with one random dependent variable $\boldsymbol{y}$ and one independent variable x related by a **regression model**:

$$\boldsymbol{y} = g(x, a_1, a_2, \ldots, a_m) + \boldsymbol{e}$$

$g(..)$ = known function (e.g. a polynomial)
$a_1, a_2, \ldots, a_m$ = m unknown regression parameters
$\boldsymbol{e}$ = **random residual ,** $E[\boldsymbol{e}] = 0$, $Var[\boldsymbol{e}] = \sigma_e^2$, CDF = $F_e(e)$.



Illustrate basic concepts with the following special case, where $g(..)$ is quadratic in x and linear in the a_i's:

$$\boldsymbol{y}(x) = g(x, a_1, a_2, a_3) + \boldsymbol{e} = a_1 + a_2 x + a_3 x^2 + \boldsymbol{e}$$

Mean of $\boldsymbol{y}(x)$ is:

$$E[\boldsymbol{y}(x)] = a_1 + a_2 x + a_3 x^2$$

Objective is to **estimate** the a_i's from a set of $\boldsymbol{y}$ measurements [$\boldsymbol{y}_1\ \boldsymbol{y}_2\ ....\ \boldsymbol{y}_n$] taken at different **known** x values [$x_1\ x_2\ ....\ x_n$]. The complete set of **measurement equations** is:

$$\boldsymbol{y}_i = \boldsymbol{y}(x_i) = a_1 + a_2 x_i + a_3 x_i^2 + \boldsymbol{e}_i \ ; \quad i = 1, ..., n$$

The residual errors [$\boldsymbol{e}_1 \boldsymbol{e}_2\ ....\ \boldsymbol{e}_n$] are all assumed to be **independent** with **identical distributions**.

Matrix Notation

Regression models and calculations are most easily expressed in terms of matrix operations. Suppose:

A = matrix (MATLAB array) with m rows and n columns
B = matrix with n rows and m columns
C, D= matrices with m rows and m columns
V = vector with n rows and 1 column

Vectors are special cases with only 1 row or column.

Operation	Matrix	Indexed	MATLAB
Matrix product of *A* and *B*	$C = AB$	$C_{ik} = \sum_{j=1}^{n} A_{ij} B_{jk}\ ;\ i = 1...m,\ k = 1...m$	`C=A*B`
Matrix transpose of *B*	$A = B'$	$A_{ij} = B_{ji}\ \ i = 1...m, j = 1...n$	`A=B'`
Matrix inverse of *C*	$D = C^{-1}$ where D is defined by: $CD=DC=I$	$\sum_{j=1}^{m} C_{ij} D_{jk} = \sum_{j=1}^{m} D_{ij} C_{jk} = I_{ik}$ $i, k = 1...m$ $I_{ik} = 1 \quad i = k$ $I_{ik} = 0 \quad i \neq k$	`D=inv(C)`
Sum-of-squares of elements of *V*	$SSV = V'V$	$SSV = \sum_{j=1}^{n} V'_j V_j = \sum_{j=1}^{n} V_j^2$	`SSV=V'*V`

It is convenient for MATLAB computations to write the set of **measurement equations** in matrix form:

$$\boldsymbol{Y} = HA + \boldsymbol{E}$$

where:

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_n \end{bmatrix} \quad H = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix} \quad A = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \quad \boldsymbol{E} = \begin{bmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \\ \vdots \\ \boldsymbol{e}_n \end{bmatrix}$$

Least-Squares Estimates of Regression Parameters

Estimated a_i's are selected to give **best fit** between measurements and predictions . The predicted $\boldsymbol{Y}$ is computed from the a_i estimates (predictions and estimates are indicated by ^ symbols):

$$\hat{\boldsymbol{Y}} = H\hat{\boldsymbol{A}} \quad ; \quad \hat{\boldsymbol{A}} = \begin{bmatrix} \hat{\boldsymbol{a}}_1 \\ \hat{\boldsymbol{a}}_2 \\ \hat{\boldsymbol{a}}_3 \end{bmatrix}$$

Measurement / prediction fit is described by **sum-of-squared prediction errors** (estimates indicated by ^ symbols) :

$$SSE(\hat{\boldsymbol{A}}) = [\boldsymbol{Y} - H\hat{\boldsymbol{A}}]'[\boldsymbol{Y} - H\hat{\boldsymbol{A}}] = \sum_{i=1}^{n} \left[\boldsymbol{y}_i - \left(\hat{\boldsymbol{a}}_1 + \hat{\boldsymbol{a}}_2 x_i + \hat{\boldsymbol{a}}_3 x_i^2\right)\right]^2$$

SSE is **minimized** when:

$$[H'H]\hat{\boldsymbol{A}} = H'\boldsymbol{Y}$$

This matrix equation is a concise way to represent three simultaneous equations in the three unknown a_i estimates. The formal solution is:

$$\hat{\boldsymbol{A}} = [H'H]^{-1} H'\boldsymbol{Y}$$

Note that $H'H$ is a 3 by 3 matrix and $H'\,\boldsymbol{Y}$ is a 3 by 1 vector for the example.

The estimation equations can be solved (for any particular set of measurements Y) with the MATLAB backslash \ operator:

```
>>  ahat = (H'*H)\(H'*y)
```

These equations only have a unique solution if $n >= m$ (i.e. if there are at least as many measurements as unknowns).

The **predicted** $\boldsymbol{y}(x)$ is obtained by substituting the a_i estimates for the true a_i values in the regression function $g(x, a_1, a_2, a_3)$:

$$\hat{\boldsymbol{y}}(x) = h(x)\hat{\boldsymbol{A}} = \hat{\boldsymbol{a}}_1 + \hat{\boldsymbol{a}}_2 x + \hat{\boldsymbol{a}}_3 x^2 \;\; ; \;\; h(x) = \begin{bmatrix} 1 & x & x^2 \end{bmatrix}$$

Estimates and prediction are **random variables.**

Same approach extends to any model with a $g(x, a_1, a_2, \ldots, a_m)$ that depends linearly on a_i's. Simply redefine H and $h(x)$.

Example -- Regression Model of Soil Sorption

A laboratory experiment provides measurements of organic solvent $\boldsymbol{y}$ sorbed onto soil particles (in mg. of solvent sorbed/kg. of soil) for different aqueous concentrations x of the solvent (in mg dissolved solvent/liter of water). Assume that the regression model proposed above applies.

Suppose specified (controlled) x values and corresponding y values are:

$$[x_1\, x_2\, \ldots .\, x_4] = [0.5 \quad 2.0 \quad 3.0 \quad 4.0 \quad]$$

$$[y_1\, y_2\, \ldots .\, y_4] = [0.4134 \quad 2.1453 \quad 1.7466 \quad 3.0742]$$

$$H = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \end{bmatrix} = \begin{bmatrix} 1 & 0.5 & 0.25 \\ 1 & 2.0 & 4.0 \\ 1 & 3.0 & 9.0 \\ 1 & 4.0 & 16.0 \end{bmatrix} \qquad Y = \begin{bmatrix} 0.4134 \\ 2.1453 \\ 1.7466 \\ 3.0742 \end{bmatrix}$$

MATLAB gives:

```
>> ahat = (H'*H)\(H'*y)
ahat =
    0.0924    0.8829   -0.0471
```

So prediction equation is:

$$\hat{y}(x) = 0.0924 + 0.8829x - 0.0471x^2$$

Plot this equation on same axes as measurements.