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 $y$ and one independent variable $x$ related by a regression model:

$$ y = g(x, a_1, a_2, ..., a_m) + e $$

$g(..)$ = known function (e.g. a polynomial)
$a_1, a_2, ..., a_m = m$ unknown regression parameters
$e =$ random residual, $E[e] = 0, Var[e] = \sigma_e^2$, 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:

$$ y(x) = g(x, a_1, a_2, a_3) + e = a_1 + a_2x +a_3x^2 + e $$

Mean of $y(x)$ is:

$$ E[y(x)] = a_1 + a_2x +a_3x^2 $$
Objective is to **estimate** the \( a_i \)'s from a set of \( y \) measurements \([y_1, y_2, ..., y_n]\) taken at different known \( x \) values \([x_1, x_2, ..., x_n]\). The complete set of **measurement equations** is:

\[
y_i = y(x_i) = a_1 + a_2x_i + a_3x_i^2 + e_i ; \quad i = 1, ..., n
\]

The residual errors \([e_1, e_2, ..., e_n]\) 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.

<table>
<thead>
<tr>
<th>Operation</th>
<th>Matrix</th>
<th>Indexed</th>
<th>MATLAB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matrix product of ( A ) and ( B )</td>
<td>( C = AB )</td>
<td>( C_{ik} = \sum_{j=1}^{n} A_{ij} B_{jk} ; i = 1...m, k = 1...m )</td>
<td>( C=A*B )</td>
</tr>
<tr>
<td>Matrix transpose of ( B )</td>
<td>( A = B' )</td>
<td>( A_{ij} = B_{ji} ; i = 1...m, j = 1...n )</td>
<td>( A=B' )</td>
</tr>
<tr>
<td>Matrix inverse of ( C )</td>
<td>( D = C^{-1} ) where ( D ) is defined by: ( CD=DC=I )</td>
<td>( \sum_{j=1}^{m} C_{ij} D_{jk} = \sum_{j=1}^{m} D_{ij} C_{jk} = I_{ik} )</td>
<td>( D=inv(C) )</td>
</tr>
<tr>
<td>Sum-of-squares of elements of ( V )</td>
<td>( SSV = V'V )</td>
<td>( SSV = \sum_{j=1}^{n} V_j V_j = \sum_{j=1}^{n} V_j^2 )</td>
<td>( SSV=V'*V )</td>
</tr>
</tbody>
</table>

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

\[
Y = HA + E
\]
where:
\[
Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad H = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}, \quad A = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}, \quad E = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}
\]

Least-Squares Estimates of Regression Parameters

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

\[
\hat{Y} = HA \quad ; \quad \hat{A} = \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \end{bmatrix}
\]

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

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

SSE is minimized when:

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

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

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

Note that \(H'H\) is a 3 by 3 matrix and \(H'Y\) 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:

\[
>> \text{ahat} = (H'\ast H) \backslash (H'\ast y)
\]

These equations only have a unique solution if \(n \geq m\) (i.e. if there are at least as many measurements as unknowns).
The predicted $y(x)$ is obtained by substituting the $a_i$ estimates for the true $a_i$ values in the regression function $g(x, a_1, a_2, a_3)$:

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

Estimates and prediction are random variables.

Same approach extends to any model with a $g(x, a_1, a_2, ..., a_m)$ 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 $y$ 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:

$$\begin{bmatrix} x_1, x_2, ..., x_4 \end{bmatrix} = \begin{bmatrix} 0.5 & 2.0 & 3.0 & 4.0 \end{bmatrix}$$

$$\begin{bmatrix} y_1, y_2, ..., y_4 \end{bmatrix} = \begin{bmatrix} 0.4134 & 2.1453 & 1.7466 & 3.0742 \end{bmatrix}$$

$$H = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \end{bmatrix} \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix}$$

MATLAB gives:

$$\begin{bmatrix} 1 & 0.5 & 0.25 \\ 1 & 2.0 & 4.0 \\ 1 & 3.0 & 9.0 \\ 1 & 4.0 & 16.0 \end{bmatrix}$$

$$\hat{a} = \begin{bmatrix} 0.0924 \\ 0.8829 \\ -0.0471 \end{bmatrix}$$

So prediction equation is:

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

Plot this equation on same axes as measurements.