Multivariant Regression

- A multi-variant relation between a dependent variable $y$ and several independent variables $x_0, x_1, x_2, \ldots, x_m$ can be represented by a table.
  - Each column of the table represents a different variable and each row a different record of instance of the relation between the various variables.
  - A row is often called a *tuple* meaning a set of related values at one instant in time or circumstance.
  - The variable $y$ is dependent on the values of the other variables $x_i$.
  - Note there is no indication or requirement that this relation be time ordered.
  - The variable $y$ might be body weight and the variables $x_i$ might be age, race, sex, etc.
  - A time series is a special case of a general relation in which the order of the tuples is important and affects the properties of the variables.
  - A relation is sometimes called a *flat file* since a paper table may be lain flat on a table.
<table>
<thead>
<tr>
<th></th>
<th>$y$</th>
<th>$x_0$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>…..</th>
<th>$x_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y(1)$</td>
<td>1</td>
<td>$x_1(1)$</td>
<td>$x_2(1)$</td>
<td>…..</td>
<td>$x_m(1)$</td>
<td></td>
</tr>
<tr>
<td>$y(2)$</td>
<td>1</td>
<td>$x_1(2)$</td>
<td>$x_2(2)$</td>
<td>…..</td>
<td>$x_m(2)$</td>
<td></td>
</tr>
<tr>
<td>…..</td>
<td>…..</td>
<td>…..</td>
<td>…..</td>
<td>…..</td>
<td>…..</td>
<td>…..</td>
</tr>
<tr>
<td>$y(N)$</td>
<td>1</td>
<td>$x_1(N)$</td>
<td>$x_2(N)$</td>
<td>…..</td>
<td>$x_m(N)$</td>
<td></td>
</tr>
</tbody>
</table>

- Note that the second column contains all 1s. This was included to allow for a constant term in any linear model.
Multivariant Regression 2

- Under special circumstances we can model this relation with a linear model of the form \( y(j) = c_0 x_0(j) + c_1 x_1(j) + c_2 x_2(j) + \ldots + c_m x_m(j) \).
  - This expression states that the \( j \)th value of \( y \) is a linear combination of a constant term \( c0 \) and a sum of constant multiples of the remaining terms.

- Write this model as a sum \( \hat{y}(j) = \sum_{i=0}^{m} c_i x_i(j) \), \( 1 \leq j \leq n \) Note that if \( i = 0 \) then \( x0 = 1 \) for all \( j \).

- Consider the variables \( xi \) and \( y \) as functions of the observation index \( j \), so that each column becomes an \( n \) dimensional column vector.

- This set of equations can then be written in matrix form as \( \hat{y} = \hat{X}c \)
  - The overhead arrow implies a column vector corresponding to the first column of the table.
  - The hat above the arrow indicates that this is an estimate of this dependent variable provided by the model.
  - The capital letter inside parentheses \( (X) \) implies a matrix.
  - The elements of this matrix are obtained from the rows and columns of the independent variables \( xi \) in the table.
Multivariant Regression 3

- The relation can be written as $\hat{y} = (X)c$.
  - The product on the right hand side of the equation is a matrix product.
  - The $j$th row of this matrix equation corresponds to the sums given in the preceding equations.
  - The vector $y$ is a single column with $n$ rows and contains the output predicted by the model for each record.
  - The vector $c$ is a column vector with $m+1$ rows that contains the unknown parameters of the model.
  - The matrix $(X)$ is a rectangular matrix with $n$ rows and $m+1$ columns. This matrix is often called the "design matrix". It is made up of all of the observations of the various variables.
  - If $n > m+1$ the $n$ equations are an over-determined set for the $m+1$ unknowns $c_i$.
  - The formal solution to this set of equations can be written $c = \text{Inverse}(X)\hat{y}$.
  - The RHS contains only data.
Least Squares Optimization

- What is meant by the inverse of a rectangular matrix like $X$?

- The method of least squares provides a technique for "best fitting" a dependent variable $y(j)$ in terms of the set of independent variables $x_i(j)$.
  - The name for this technique comes from the fact that we choose the coefficients to minimize the variance of the model residuals.
  - The residual for the jth record is $\text{Res}(j) = y(j) - \hat{y}(j)$.
  - The variance of the residuals is the mean of the square of the residuals
    \[ \text{Var}(\varepsilon) = \frac{1}{N} \sum_{j=1}^{N} \left( y(j) - \sum_{i=0}^{M} c_i x_i(j) \right)^2 \]
    - Actually, an unbiased estimate of the "residual variance" is properly normalized by a factor smaller than $N$ by the total number of coefficients in the model. Each coefficient corresponds to a degree of freedom lost from the original data. When $N$ is large this difference is negligible.

- The residual variance will be a minimum (or maximum) if its total derivative with respect to the independent free parameters is zero.
  \[ d(\text{Var}) = \sum_{i=0}^{M} \frac{\partial \text{Var}}{\partial c_i} dc_i = 0 \]
Least Square Optimization: Normal Equations

- Because the unknown coefficients of the model are independent this condition can be satisfied only if

\[
\frac{\partial \text{Var}}{\partial c_k} = 0 \quad \text{for all } c_k
\]

- Thus we must carry out this differentiation to find the condition that will guarantee minimum variance.

- Differentiation of the residual variance with respect to \( c_k \) gives the \( k \)th parameter

\[
\frac{\partial \text{Var}}{\partial c_k} = -\frac{2}{N} \sum_{j=1}^{N} \left( y(j) - \sum_{i=0}^{M} c_i x_i(j) \right) \left( \sum_{i=0}^{M} \delta_{ik} x_i(j) \right) = 0
\]

- Divide both sides of the equation by 2 but keep the \( 1/N \).

- Evaluate the last term (containing the delta function) to obtain \( x_k(j) \).

- Expand the binomial obtaining two terms containing sums on only \( j \)

\[
\frac{\partial \text{Var}}{\partial c_k} = \frac{1}{N} \sum_{j=1}^{N} x_k(j) \sum_{i=0}^{M} c_i x_i(j) - \frac{1}{N} \sum_{j=1}^{N} x_k(j) y(j) = 0
\]
Least Square Optimization: Normal Equations 2

\[ \frac{\partial \text{Var}}{\partial c_k} = \frac{1}{N} \sum_{j=1}^{N} x_k(j) \sum_{i=0}^{M} c_i x_i(j) - \frac{1}{N} \sum_{j=1}^{N} x_k(j) y(j) = 0 \]

- Determine the equation that the \( k^{th} \) model coefficient must satisfy to obtain minimum variance by reversing the order of summation and moving the term with \( y(j) \) to the RHS.

\[ \sum_{i=0}^{M} \left( \frac{1}{N} \sum_{j=1}^{N} x_i(j)x_k(j) \right) c_i = \frac{1}{N} \sum_{j=1}^{N} x_k(j) y(j) \]

- Taking the complete set of \( (M+1) \) such equations produced by considering all coefficients we have a set of \( (M+1) \) equations for \( (M+1) \) unknowns \( c_k \).

- The sums on \( j \) may be thought of as scalar products between vectors of length \( N \). Alternatively, they are cross correlations between the basis functions and the data to be fitted.
Least Square Optimization: Solving the Normal Equations

- In matrix form the set of M+1 normal equations
  \[
  \sum_{i=0}^{M} \left( \frac{1}{N} \sum_{j=1}^{N} x_i(j)x_k(j) \right) \hat{c}_i = \frac{1}{N} \sum_{j=1}^{N} x_k(j)y(j)
  \]
  become \((X)^T(X)\hat{c} = (X)^T \hat{y}\)

- The matrix \((X)\) is rectangular with dimension (N by M+1).
- The product \((X)^T(X)\) is a square matrix of dimension (M+1 by M+1).
  - Matrix \((X)^T\) is the transpose of matrix \((X)\).
- If the determinant of this square matrix is non-zero we can solve the set of equations by multiplying by the inverse of the square matrix on the LHS. The solution is:
  \[
  \hat{c} = [(X)^T(X)]^{-1}(X)^T \hat{y}
  \]
  - The inverse of the non-square matrix \((X)\)
    \[
    (X)^{-1} = [(X)^T(X)]^{-1}(X)^T
    \]
An Example

• Example data are shown in the second ($x_1$) and third ($y$) columns of table (1).

• The independent variable $x_1$ increases linearly from 1 to 10.

• The dependent variable $y$ increases more or less linearly, but with added noise.

• We wish to make a linear fit to the data $y(x_1)$.
  – First use the MATLAB polyfit and polyval functions
    
    ```matlab
    cof  = polyfit(x, y, 1);
    fit  = polyval(cof, x);
    ```
    
    The result is $y = 1.2182 \times x_1 -0.2000$.

• The first two columns of the table give the matrix $(X)$ with $x_0$ being a column of all 1's and $x_1$ being the independent variable.

<table>
<thead>
<tr>
<th></th>
<th>$x_0$</th>
<th>$x_1$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>10</td>
<td>12</td>
</tr>
</tbody>
</table>
Example Continued

- The matrix \((X)^T(X)\) is given by the 2 by 2 matrix displayed in table (2).

- The inverse of this square matrix is shown in table (3).

- The product \((X)^T*\mathbf{y} = [65, 458]^T\). Multiplying this vector by the inverse matrix gives \([-0.2000, 1.2181]\). Note the coefficients are produced in reverse order from those of polyfit, but are identical to five decimal places.
Minimum Variance and Prediction Efficiency

- The definition of the residual variance as a function of the model coefficients is
  \[ \text{Var}(\hat{\varepsilon}) = \frac{1}{N} \sum_{j=1}^{N} \left( y(j) - \sum_{i=0}^{M} c_i x_i(j) \right)^2 \]

- Square the terms in parentheses obtaining
  \[ \text{Var} = \frac{1}{N} \sum_{j=1}^{N} \left[ y(j)^2 - 2y(j) \sum_{i=0}^{M} c_i x_i(j) + \left( \sum_{i=0}^{M} c_i x_i(j) \right)^2 \right] \]

- Expand the last term containing the square of a sum using dummy indices, distribute the \( j \) sum over to all terms giving
  \[ \text{Var}(\hat{\varepsilon}) = \frac{1}{N} \sum_{j=1}^{N} y(j)^2 - \frac{2}{N} \sum_{j=1}^{N} y(j) \sum_{i=0}^{M} c_i x_i(j) + \frac{1}{N} \sum_{j=1}^{N} \sum_{i=0}^{M} c_i x_i(j) \sum_{k=0}^{M} c_k x_k(j) \]

- Rearrange sums and take the inner sum over \( j \).
  \[ \text{Var}(\hat{\varepsilon}) = \frac{1}{N} \sum_{j=1}^{N} y(j)^2 - 2 \sum_{i=0}^{M} c_i \left( \frac{1}{N} \sum_{j=1}^{N} x_i(j) y(j) \right) + \sum_{k=0}^{M} c_k \sum_{i=0}^{M} c_i \left( \frac{1}{N} \sum_{j=1}^{N} x_i(j) x_k(j) \right) \]
Minimum Variance and Prediction Efficiency 2

\[ \text{Var}(\tilde{c}) = \frac{1}{N} \sum_{j=1}^{N} y(j)^2 - 2 \sum_{i=0}^{M} c_i \left( \frac{1}{N} \sum_{j=1}^{N} x_i(j) y(j) \right) + \sum_{k=0}^{M} \sum_{i=0}^{M} c_i \left( \frac{1}{N} \sum_{j=1}^{N} x_i(j) x_k(j) \right) \]

- The condition for minimum variance is

\[ \sum_{i=0}^{M} \left( \frac{1}{N} \sum_{j=1}^{N} x_i(j) x_k(j) \right) c_i = \frac{1}{N} \sum_{j=1}^{N} x_k(j) y(j) \]

- Compare these two equations and note that we can substitute the LHS of the minimum variance condition into the last term of the top (variance) equation. This guarantees that the variance will be a minimum. The substitution introduces a positive term that cancels half of the preceding term leaving us with

\[ \text{Min}(\text{Var}(\tilde{c})) = \frac{1}{N} \sum_{j=1}^{N} y(j)^2 - \sum_{i=0}^{M} c_i \left( \frac{1}{N} \sum_{j=1}^{N} x_i(j) y(j) \right) \]

- Reorder the terms

\[ \text{Min}(\text{Var}(\tilde{c})) = \frac{1}{N} \sum_{j=1}^{N} y(j)^2 - \frac{1}{N} \sum_{j=1}^{N} \left( \sum_{i=0}^{M} c_i x_i(j) \right) y(j) \]
Minimum Variance and Prediction Efficiency 3

\[ \text{Min}(\text{Var}(\tilde{e})) = \frac{1}{N} \sum_{j=1}^{N} y(j)^2 - \frac{1}{N} \sum_{j=1}^{N} \left( \sum_{i=0}^{M} c_i x_i(j) \right) y(j) \]

- The inner sum in the last term is the model prediction \( \hat{y}(j) \). The second term on the right hand side is the scalar product between the prediction and the data.

- If the model is a perfect fit to the data the two terms on the RHS will cancel giving zero as the minimum residual variance.

- If there is no relation between y and x then all possible values of y can occur for any x. In this case the model value for y will just be the mean value of y or \( \hat{y}(j) = \bar{y}(j) \).

- Substituting this into the RHS gives
  \[ \text{Min}(\text{Var}(\tilde{e})) = \frac{1}{N} \sum_{j} y(j)^2 - \frac{1}{N} \bar{y} \sum_{j} y(j) \]

- Thus the maximum of the minimum residual variance is just the variance of the original data.