Multivariate Regression, leading up to Principal Component Regression/Analysis
-- an introductory tutorial to some of the most important ideas in multivariate regression.
Instructor: Nam Sun Wang

Multivariate Regression.

Let us expand the number of independent variables and dependent variables. Here, we are given a set of data consisted of a series of m+1 independent variables $x^{<0>}$, $x^{<1>}$, ..., $x^{<m>}$, and l+1 dependent variables $y^{<0>}$, $y^{<1>}$, ..., $y^{<l>}$. An example is how the quality, thickness, and strength of a paper product (Y) depend on water content, source of fiber, digestion temperature, pH, etc. (X). Another example is how the yield and composition in a chemical reactor (Y) depend on stirrer speed, feed flow rate, reactant concentrations, ... (X). The chemical composition (Y) measured with a chemical sensor may be related to the response of an array sensor (X). The mechanical or chemical property of a material (Y) may depend on its color spectrum (X). An economic example may be how the stock price and trading volume (Y) depend on the prevailing interest rate, the company's earning, the quarter in the calendar, ... (X). The gross national product (Y) may depend on a country's population, literacy rate, average age, level of rainfall, ... (X). The probability of death, thus, the premium of a life insurance policy, may depend on the many attributes of the insured. The salary and popularity of a football player (Y) may depend on his height, weight, running speed, strength, running yards gained, passing yards gained, number of touchdowns, number of fumbles, hours of practice per day, ... (X). The standardized test scores or the grade point average of a student (Y) may depend on the number of hours spent in school, amount of daily TV time, the household income, gender, the time of the day the test is taken, and maybe even the number of whip lashes received since one's birth or the average number of glasses of milk one consumes daily (X). Furthermore, a student's standardized test scores and grade point average may be closely correlated. The examples are endless.

What we include as an independent variable need not actually affect the dependent variables in any way. It is not necessarily a reflection of what we believe to affect the process. If we so desire, we can throw in everything that may remotely affect the dependent variables. One thing regression tells us is whether there is indeed any correlation between the two. A word of caution: existence of a correlation does not imply the existence of an actual connection or the existence of a direct cause-effect relationship. It is often true that "look and thou shall find." To judge whether a particular degree of correlation is significant, we need to resort to tools from probability, hypothesis testing, metrics, reliability, controlled experimentation, etc. In addition, we need to worry about a lot of other things: how to include representative samples, adequate sample size, define the domain of validity to avoid extrapolation, and detection and rectification of outliers and gross errors -- none of which will be addressed in this worksheet.

Let us start with some independent data x and some dependent data y.

Raw data: $\left(x^{<0>}\right)_i \left(x^{<1>}\right)_i \ldots \left(x^{<m>}\right)_i$ and $\left(y^{<0>}\right)_i \left(y^{<1>}\right)_i \ldots \left(y^{<l>}\right)_i$ $i := 0 .. n$

Combine these data into an independent matrix X and a dependent matrix Y.

Raw data: X and Y

As before, our task is to find a set of regression coefficients a such that the given model function f(X,a) passes through the given data (X,Y) as closely as possible. That is to say that we try to minimize some sort of error between Y and f(X,a).

$$\underset{a}{\text{Minimize}} \quad \text{error} := Y - f(X, a)$$

There are many measures of this error. One of them is the absolute error, which is mathematically cumbersome to work with because the absolute function is not differentiable at zero.

$$\underset{a}{\text{Minimize}} \quad |\text{error}| = |Y - f(X, a)|$$

As before, we try to minimize the sum of squared errors, which is mathematically more tractable.

$$\underset{a}{\text{Minimize}} \quad \text{sse} := \sum_{i=0}^{n} \left(\text{error}_i\right)^2 = \sum_{i=0}^{n} \sum_{k=0}^{l} \left(y_{i,k} - f\left(X_i, a\right)^{<k>}\right)^2$$

$$:= E^T \cdot E := (Y - f(X, a))^T \cdot (Y - f(X, a))$$

Multivariate Linear Regression (MLR).

The simplest model is a linear one where the X matrix is simply the plain given set of dependent variables $x^{<0>}$, $x^{<1>}$, ..., $x^{<l>}$. We can also have functions and combinations of $x^{<\bullet>}$ in X (e.g., auto- or cross-terms of two independent variables such as $x^{<0>} \cdot x^{<0>}$ and $x^{<0>} \cdot x^{<1>;}$ or functions of one or several independent variables such as $x^{<0>}/x^{<1>}$, $\sin(x^{<0>})$, $x^{<0>} \cdot \exp(x^{<1>})$, etc.). At any rate, the linear combination of these terms is expressed as:

$$Y := X \cdot a + E$$

Example -- the Simplest Model. Let us generate some artificial data to demonstrate multivariate regression.

Number of points: $N := 50$ $i := 0 .. N$

Dimension: $m := 2$ $j := 0 .. m$

The independent variables vary randomly in three directions, but with more variation in the 0th direction, a bit less in the 1st direction, and even less in the 2nd direction. To generate a different set of artificial data, mark any part of the following equation with the cursor and press the F9 key on a PC.

$$X^{\langle i \rangle} := (\mathrm{rnd}(1) - 0.5)\cdot\begin{pmatrix}10\\0\\0\end{pmatrix} + (\mathrm{rnd}(1) - 0.5)\cdot\begin{pmatrix}0\\1\\0\end{pmatrix} + (\mathrm{rnd}(1) - 0.5)\cdot\begin{pmatrix}0\\0\\0.1\end{pmatrix} \qquad X := X^{T}$$

Generate the dependent variable which varies linearly with the first two independent variables plus a small amount of noise.

$$Y_i := X_{i,0} + 3\cdot X_{i,1} + (\mathrm{rnd}(0.1) - 0.05) \qquad y(x) := x\cdot\begin{pmatrix}1\\3\\0\end{pmatrix}$$

$Y := y(X) + (\mathrm{rnd}(0.1) - 0.05)$

↑ This formula is toggled off because the same noise is added uniformly to all data points.

The least squares solution is again given by the same normal equation as before.

$$a := \left(X^{T}\cdot X\right)^{-1}\cdot X^{T}\cdot Y$$

$$a = \begin{pmatrix}1.003\\3\\-0.109\end{pmatrix}$$

Thus, we are able to recover the underlying structural relationship between X and Y, namely that Y is equally dependent on $x^{\langle 0 \rangle}$ and $x^{\langle 1 \rangle}$ but was not at all affected by $x^{\langle 2 \rangle}$.

The regression equation, which is valid for both a single point and multiple number of points, is:

$$y_{regress}(x) := x\cdot a$$

Examples: $y_{regress}((5 \quad 0.5 \quad 0.05)) = 6.507$ $y((5 \quad 0.5 \quad 0.05)) = 6.5$

The following arguments are out of the calibration range. **Extrapolation is dangerous**.

$y_{regress}((1 \quad 1 \quad 1)) = 3.894$ $y((1 \quad 1 \quad 1)) = 4$

$$y_{regress}\left(\begin{pmatrix}1 & 1 & 1\\5 & 0.5 & 0.05\end{pmatrix}\right) = \begin{pmatrix}3.894\\6.507\end{pmatrix} \qquad y\left(\begin{pmatrix}1 & 1 & 1\\5 & 0.5 & 0.05\end{pmatrix}\right) = \begin{pmatrix}4\\6.5\end{pmatrix}$$

Goodness of fit:

$sse_{old} := Y\cdot Y$ $sse_{old} = 414.48$

$sse := \left(Y - y_{regress}(X)\right)\cdot\left(Y - y_{regress}(X)\right)$ $sse = 0.043$

$$r2 := \frac{sse_{old} - sse}{sse_{old}} \qquad r2 = 99.99\ \%$$

$r := \sqrt{r2}$ $r = 99.995\ \%$

Example -- Multivariate Y. If there are more than one dependent variable, we perform regression analysis no differently.

$Y_{i,1} := X_{i,0} + X_{i,1} + (\text{rnd}(0.1) - 0.05)$

$y(x) := x \cdot \begin{pmatrix} 1 & 1 \\ 3 & 1 \\ 0 & 0 \end{pmatrix}$

$Y := y(X) + (\text{rnd}(0.1) - 0.05)$

↑ Toggled off because the noise is not added correctly with this formula.

$a := \left(X^T \cdot X\right)^{-1} \cdot X^T \cdot Y$

$$a = \begin{pmatrix} 1.003 & 1.001 \\ 3 & 0.957 \\ -0.109 & 0.037 \end{pmatrix}$$

← Linear regression has captured the underlying structure. Compare it to the matrix in y(x).

$y_{regress}(x) := x \cdot a$

Examples: $y_{regress}((5 \quad 0.5 \quad 0.05)) = (6.507 \quad 5.486)$ $y((5 \quad 0.5 \quad 0.05)) = (6.5 \quad 5.5)$

The following arguments are out of the calibration range. Extrapolation is dangerous.

$y_{regress}((1 \quad 1 \quad 1)) = (3.894 \quad 1.996)$ $y((1 \quad 1 \quad 1)) = (4 \quad 2)$

$$y_{regress}\begin{pmatrix} 1 & 1 & 1 \\ 5 & 0.5 & 0.05 \end{pmatrix} = \begin{pmatrix} 3.894 & 1.996 \\ 6.507 & 5.486 \end{pmatrix} \quad y\begin{pmatrix} 1 & 1 & 1 \\ 5 & 0.5 & 0.05 \end{pmatrix} = \begin{pmatrix} 4 & 2 \\ 6.5 & 5.5 \end{pmatrix}$$

Example -- Linear Combinations of Nonlinear Basis Functions.

Generate the dependent variable which varies nonlinearly with the first two independent variables plus a small amount of noise.

$Y := 0$ ... reset the variable Y for a new assignment.

$$Y_i := X_{i,0} + 3 \cdot X_{i,1} + X_{i,0} \cdot X_{i,1} + 10000 \cdot \left(X_{i,2}\right)^2 + (\text{rnd}(0.1) - 0.05)$$

$$y(x) := x \cdot \begin{pmatrix} 1 \\ 3 \\ 0 \end{pmatrix} + x \cdot \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 10000 \end{pmatrix} \cdot x^T$$

The least squares solution that ignores the auto- and cross-terms is again given by the same normal equation.

$$a := \left(X^T \cdot X\right)^{-1} \cdot X^T \cdot Y$$

$$a = \begin{pmatrix} 0.644 \\ 7.461 \\ 56.487 \end{pmatrix}$$

As expected, with just the plain X terms, we fail to recover the underlying structural relationship between X and Y.

The regression equation is:

$$y_{regress}(x) := x \cdot a$$

Examples: $y_{regress}((5 \quad 0.5 \quad 0.05)) = 9.774$ $\qquad$ $y((5 \quad 0.5 \quad 0.05)) = 34$

Goodness of fit:

$sse_{old} := Y \cdot Y$ $\qquad$ $sse_{old} = 5.781 \cdot 10^3$

$sse := \left(Y - y_{regress}(X)\right) \cdot \left(Y - y_{regress}(X)\right)$ $\qquad$ $sse = 5.318 \cdot 10^3$

$r2 := \dfrac{sse_{old} - sse}{sse_{old}}$ $\qquad$ $r2 = 8.019 \cdot \%$

$r := \sqrt{r2}$ $\qquad$ $r = 28.318 \cdot \%$ ← Not too much variation in Y has been captured.

Repeat regression by including the cross terms. Below we expand the independent variable X to include the cross terms. The resulting expanded matrix is Xx.

$Xx := X$

$Xx_{i,3} := X_{i,0} \cdot X_{i,0}$ $\quad$ $Xx_{i,4} := X_{i,1} \cdot X_{i,1}$ $\quad$ $Xx_{i,5} := X_{i,2} \cdot X_{i,2}$

$Xx_{i,6} := X_{i,0} \cdot X_{i,1}$ $\quad$ $Xx_{i,7} := X_{i,0} \cdot X_{i,2}$ $\quad$ $Xx_{i,8} := X_{i,1} \cdot X_{i,2}$

The least squares solution that ignores the auto- and cross-terms is again given by the same normal equation.

$$a := \left(Xx^T \cdot Xx\right)^{-1} \cdot Xx^T \cdot Y$$

$$a^T = (\ 0.998 \quad 2.967 \quad -0.036 \quad 3.329 \cdot 10^{-4} \quad -0.098 \quad 1 \cdot 10^4 \quad 0.99 \quad 0.081 \quad 0.528\ )$$

The regression equation is:

$$y_{regress}(x) := x \cdot \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} + x \cdot \begin{bmatrix} a_3 & a_6 & a_7 \\ 0 & a_4 & a_8 \\ 0 & 0 & a_5 \end{bmatrix} \cdot x^T$$

Examples: $y_{regress}((5 \quad 0.5 \quad 0.05)) = 33.961$ $y((5 \quad 0.5 \quad 0.05)) = 34$

Goodness of fit:

$sse_{old} := Y \cdot Y$ $sse_{old} = 5.781 \cdot 10^3$

$$sse := \sum_{i=0}^{last(Y)} \left(Y_i - y_{regress}\left(\left(X^T \right)^{<i>^T} \right) \right)^2$$ $sse = 0.031$

$r2 := \frac{sse_{old} - sse}{sse_{old}}$ $r2 = 99.999\,\%$

$r := \sqrt{r2}$ $r = 100\,\%$ ← Now, with the quadratic terms, practically all variation in Y is captured.

Example -- Linearly correlated X. The independent variables have the same range as before; however, the first two independent variables $x^{<0>}$ and $x^{<1>}$ are mostly dependent, with $x^{<0>}$ being 10 times of $x^{<1>}$.

$\text{conde}\left(X^T \cdot X\right) = 1.205 \cdot 10^4$ ... condition number of the last example

$X := 0$ ... reset X to prepare for a new assignment.

$$X^{<i>} := (\text{rnd}(1) - 0.5) \cdot \begin{pmatrix} 10 \\ 1 \\ 0 \end{pmatrix} + (\text{rnd}(1) - 0.5) \cdot \begin{pmatrix} 0 \\ 0.001 \\ 0 \end{pmatrix} + (\text{rnd}(1) - 0.5) \cdot \begin{pmatrix} 0 \\ 0 \\ 0.1 \end{pmatrix} \qquad X := X^T$$

↑ Without this small noise term, $x^{<0>}$ and $x^{<1>}$ are completely dependent and X^TX is singular.

$Y := 0 \qquad Y_i := X_{i,0} + 3 \cdot X_{i,1} + (\text{rnd}(0.1) - 0.05)$

$$y(x) := x \cdot \begin{pmatrix} 1 \\ 3 \\ 0 \end{pmatrix}$$

$a := \left(X^T \cdot X\right)^{-1} \cdot X^T \cdot Y$

$$a = \begin{pmatrix} 2.001 \\ -7.008 \\ 0.102 \end{pmatrix}$$

← The recovered structure from linear regression is <u>not</u> what we had put in. (Compare to the vector in y(x))

$\text{conde}\left(X^T \cdot X\right) = 8.735 \cdot 10^7$ ← The condition number is very large, which means $X^T \cdot X$ is almost singular. This provides a warning that linear regression is breaking down.

Regression model (which is not to be trusted):

$y_{\text{regress}}(x) := x \cdot a$

Examples. The first example works O.K. because $x^{<0>}$ and $x^{<1>}$ are correlated. On the other hand, the output numbers in the second example do not agree at all with the original model because the given $x^{<0>}$ and $x^{<1>}$ are not correlated the same way as the calibration data are. Note that although the input numbers are each within the range of the calibration data, the uncorrelated pattern in the second example is not included in the calibration data. Technically, this, too, is a case of extrapolation.

$y_{\text{regress}}((5 \quad 0.5 \quad 0.05)) = 6.507 \qquad y((5 \quad 0.5 \quad 0.05)) = 6.5$ ← O.K.

$y_{\text{regress}}((5 \quad -0.5 \quad 0.05)) = 13.516 \qquad y((5 \quad -0.5 \quad 0.05)) = 3.5$ ← totally off -- extrapolation.