CO902

Supporting Lecture Notes:

Least Squares Derivation in Matrix Mode & the Pseudoinverse

Lecture 5 — 4 Feb 2013

# 1 Fitting Variables to Data with Least Squares

We have data $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ that we wish to fit with a linear combination of the columns of $\mathbf{X}$, where

$$\mathbf{X} = [\, \mathbf{1}\, \mathbf{X}_1\, \cdots\, \mathbf{X}_D\, ]$$

and $\mathbf{X}_d = \{X_{di}\}_{i=1}^n$ are the data on the $d$th variable, $d = 1, \ldots, D$. To find the "least squares solution" we need to find length vector $\mathbf{w}$ of length $(D+1)$ that minimizes the sum of squared errors $(\mathbf{Xw} - \mathbf{Y})'(\mathbf{Xw} - \mathbf{Y})$. It is useful to see how to do this in "matrix mode".

The two following results from matrix calculus are useful. For column vectors $\mathbf{x}$ and $\mathbf{a}$ of the same length

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{a}'\mathbf{x} = \frac{\partial}{\partial \mathbf{x}} \mathbf{x}'\mathbf{a} = \mathbf{a}.$$

For a column vector $\mathbf{x}$ and matrix $\mathbf{A}$

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{x}'\mathbf{A}\mathbf{x} = (\mathbf{A} + \mathbf{A}')\mathbf{x},$$

and, if $\mathbf{A}$ is symmetric, $\frac{\partial}{\partial \mathbf{x}} \mathbf{x}'\mathbf{A}\mathbf{x} = 2\mathbf{A}\mathbf{x}$ (think $(d/dx)x^2 = 2x$).

We take the derivative of the sum of squared errors w.r.t. $\mathbf{w}$, set to zero and solve for $\hat{\mathbf{w}}$:

$$\begin{aligned}
\mathbf{0} &= \frac{\partial}{\partial \mathbf{w}} (\mathbf{Xw} - \mathbf{Y})'(\mathbf{Xw} - \mathbf{Y}) \\
&= \frac{\partial}{\partial \mathbf{w}} \left[\mathbf{w}'\mathbf{X}'\mathbf{X}\mathbf{w} - \mathbf{w}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\mathbf{w} - \mathbf{Y}'\mathbf{Y}\right] \\
&= \frac{\partial}{\partial \mathbf{w}} \mathbf{w}'\mathbf{X}'\mathbf{X}\mathbf{w} - 2\frac{\partial}{\partial \mathbf{w}} \mathbf{w}'\mathbf{X}'\mathbf{Y} - 0 \\
&= 2\mathbf{X}'\mathbf{X}\mathbf{w} - 2\mathbf{X}'\mathbf{Y}.
\end{aligned}$$

This leads to the definition of the Least Squares "Normal Equations"

$$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\mathbf{w}.$$

This system of $D + 1$ equations defines the Least Squares solutions... any vector $\mathbf{w}$ that satisfies it is a "Least Squares" estimate. Of course, if $\mathbf{X}$ is full rank, then we have the standard answer

$$\hat{\mathbf{w}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

*However* you should avoid at all costs ever using this formulate to estimate $\hat{\mathbf{w}}$. It is numerically unstable if $\mathbf{X}$ is poorly conditioned. The "best practice" estimates for least squares is to make a Q-R decomponsition of the Normal Equations (this is what happens inside any regression program). But, in Matlab, the easiest thing to use is the Moore Penrose pseudoinverse.

# 2 Moore-Penrose pseudoinverse

For an arbitrary matrix $\mathbf{A}$, not necessarly square nor even full rank, the Moore-Penrose pseudoinverse is written $\mathbf{A}^-$ and satisfies the following conditions

1. $\mathbf{A}\mathbf{A}^-$ is symmetric,
2. $\mathbf{A}^-\mathbf{A}$ is symmetric,
3. $\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$,
4. $\mathbf{A}^-\mathbf{A}\mathbf{A}^- = \mathbf{A}^-$.

From these you can show that $(\mathbf{A}^-)^- = \mathbf{A}$ and $(\mathbf{A}^-)' = (\mathbf{A}')^-$, and also this useful result: $\mathbf{A}' = \mathbf{A}'\mathbf{A}\mathbf{A}^-$. The gory details to this last one are

$$\begin{array}{rrcll} & (\mathbf{A}\mathbf{A}^-)' & = & \mathbf{A}\mathbf{A}^- & (\textit{by condition 1}) \\ \Leftrightarrow & (\mathbf{A}^-)'\mathbf{A}' & = & \mathbf{A}\mathbf{A}^- & \\ \Leftrightarrow & \mathbf{A}'(\mathbf{A}^-)'\mathbf{A}' & = & \mathbf{A}'\mathbf{A}\mathbf{A}^- & (\textit{premultiply by } \mathbf{A}') \\ \Leftrightarrow & \mathbf{A}'(\mathbf{A}')^-\mathbf{A}' & = & \mathbf{A}'\mathbf{A}\mathbf{A}^- & (\textit{by } (\mathbf{A}^-)' = (\mathbf{A}')^-) \\ \Leftrightarrow & \mathbf{A}' & = & \mathbf{A}'\mathbf{A}\mathbf{A}^- & (\textit{by condition 3}). \end{array}$$

But this means that the pseudo inverse of $\mathbf{X}$ times $\mathbf{Y}$ is a solution to Normal Equations! Consider $\tilde{\mathbf{w}} = \mathbf{X}^-\mathbf{Y}$, the Normal Equations are then

$$\begin{aligned} \mathbf{X}'\mathbf{Y} &= \mathbf{X}'\mathbf{X}\,\tilde{\mathbf{w}} \\ &= \mathbf{X}'\mathbf{X}\,\mathbf{X}^-\mathbf{Y} \\ &= \mathbf{X}'\mathbf{Y}, \end{aligned}$$

using the previous result.

The reason to use $\hat{\mathbf{w}} = \mathbf{X}^-\mathbf{Y}$ instead of $\hat{\mathbf{w}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is: (1) the pseudo inverse is numerically stable, and (2) it will work even if $\mathbf{X}$ is rank deficient. Of course, if $\mathbf{X}$ is rank deficient then there are an infinite number of solutions to the Normal Equations, but we can be happy in the knowledge that the Moore-Penrose pseudo inverse will always give a single unique solution.

In matlab, use `pinv(X)` to get the pseudo inverse.