Least Square Regression 2

Generalized linear model (GLM), Recursive least square (RLS) and Two stage least square (2SLS)

Part A : Generalized Linear Model

Generalized linear model

- example : logistic regression
- example : poisson regression

Part B : Recursive Least Square

Given N observations $A^{(N)}$ and $B^{(N)}$, and a linear model with parameter X as shown in the following. $\varepsilon^{(N)}$ is uncorrelated Gaussian noise, the covariance matrix $\Sigma^{(N)}$ is thus a diagonal matrix, and its inverse equals to elementwise inverse.

$$\begin{aligned} B^{(N)} &= A^{(N)}X + \varepsilon^{(N)} \\ \Sigma^{(N)} &= E[\varepsilon^{(N)}\varepsilon^{(N)T}] = diag(\sigma_1^2 \quad \sigma_2^2 \quad \sigma_3^2 \quad \dots \quad \sigma_N^2) \\ \sigma_n^2 &= E[\varepsilon_n^{(N)}\varepsilon_n^{(N)T}] \end{aligned}$$

Lets derive the recursive formula for weighted least square, i.e. $X_{WLS}^{(N)}$ based on $X_{WLS}^{(N-1)}$, where bracketed upper index denotes recursion index. Please note that some matrices are generated by **stacking** up all N observations, hence their sizes are dependent on N, while other matrices denote the **snapshot** of the N^{th} recursion, their size are independent of N. In case of stacked matrices, there may be a lower index n, denoting the n^{th} observation among all N observations. Furthermore, lets define the following error terms.

$\eta^{(N)}$	=	$X - X_{WLS}^{(N)}$	= model parameter error	(snapshot of the Nth recursion)
$e^{(N)}$	=	$B^{(N)} - A^{(N)}X_{WLS}^{(N-1)}$	= fitting error with previous model	(stacking all N observations)
$\varepsilon^{(N)}$	=	$B^{(N)} - A^{(N)}X$	= fitting error with ground truth model	(stacking all N observations)

We can express $e^{(N)}$ in terms of $\varepsilon^{(N)}$ by subtracting their definitions above :

$$\begin{aligned} e^{(N)} - \varepsilon^{(N)} &= A^{(N)}X - A^{(N)}X_{WLS}^{(N-1)} \\ \Rightarrow \quad e^{(N)} &= \varepsilon^{(N)} + A^{(N)}(X - X_{WLS}^{(N-1)}) \qquad \text{equation 1} \end{aligned}$$

The size of the matrices are :

$A^{(N)}$	=	$N \times M$ row data matrix	$A_n^{(N)}$	=	$N \times 1$	=	input data of n^{th} observation
$B^{(N)}$	=	$N \times 1$	$b_n^{(N)}$	=	1×1	=	output data of n^{th} observation
$X_{WLS}^{(N)}$	=	$M \times 1$ (snapshot)					
$\eta^{(N)}$	=	$M \times 1$ (snapshot)					
$e^{(N)}$	=	$N \times 1$	$e_n^{(N)}$	=	1×1	=	error of n^{th} observation
$\varepsilon^{(N)}$	=	$N \times 1$	$\varepsilon_n^{(N)}$	=	1×1	=	noise of n^{th} observation

Furthermore :

$$\begin{aligned} A_n^{(N)} &= A_n^{(N-1)} \quad \forall n \in [1, N-1] \\ B_n^{(N)} &\equiv b_n^{(N)} \quad \text{(scalar)} \end{aligned}$$

There are two approaches (1) the minimization of model parameter error and (2) the minimization of fitting error. Both end up with the same result. We will compare the physical meaning of both approaches at the end.

Approach 1 – Minimization of model parameter error

In approach 1, we firstly define our recursive estimator, which recursively update the estimation with the product of a direction (specified by gain vector K) and a magnitude (specified by current fitting error when we apply previous model estimation to the N^{th} observation, i.e. the new data point).

$$X_{WLS}^{(N)} = X_{WLS}^{(N-1)} + K^{(N)}e_N^{(N)} \qquad \text{equation 2}$$

where $e_N^{(N)} = b_N^{(N)} - A_N^{(N)}X_{WLS}^{(N-1)}$

Our objective is to find the optimum gain vector K, that minimize of model paramater error, which is defined as :

$$\begin{aligned} L &= E[\eta^{(N)T}\eta^{(N)}] && \text{inner product of model parameter error} \\ &= E[tr(\eta^{(N)T}\eta^{(N)})] && \text{trick to convert inner product to outer product} \\ &= E[tr(\eta^{(N)}\eta^{(N)T})] && \text{outer product of model parameter error} \\ &= tr(E[\eta^{(N)}\eta^{(N)T}]) && \text{recall that trace is sum of all diagonal elements} \\ &= tr(P^{(N)}) && \text{where P is covariance of model parameter error} \end{aligned}$$

Lets derive the recursive formula for $\eta^{(N)}$ (which will be used to derive the recursive formula for $P^{(N)}$).

$$
\begin{aligned}
\eta^{(N)} &= X - X_{WLS}^{(N)} \\
&= X - X_{WLS}^{(N-1)} - K^{(N)} e_N^{(N)} && \text{i.e. we have expressed } \eta^{(N)} \text{ in terms of } e^{(N)} \\
&= (X - X_{WLS}^{(N-1)}) - K^{(N)}(\varepsilon^{(N)} + A^{(N)}(X - X_{WLS}^{(N-1)})) && \text{using equation 1} \\
&= (I - K^{(N)} A_N^{(N)})(X - X_{WLS}^{(N-1)}) - K^{(N)} \varepsilon_N^{(N)} && \text{i.e. we have expressed } \eta^{(N)} \text{ in terms of } \varepsilon^{(N)} \\
&= (I - K^{(N)} A_N^{(N)})\eta^{(N-1)} - K^{(N)} \varepsilon_N^{(N)} && \text{equation 3}
\end{aligned}
$$

Lets derive the recursive formula for $P^{(N)}$.

$$
\begin{aligned}
P^{(N)} &= E[\eta^{(N)} \eta^{(N)T}] \\
&= E[((I - K^{(N)} A_N^{(N)})\eta^{(N-1)} - K^{(N)} \varepsilon_N^{(N)})((I - K^{(N)} A_N^{(N)})\eta^{(N-1)} - K^{(N)} \varepsilon_N^{(N)})^T] && \text{using equation 3} \\
&= \begin{aligned}[t] &(I - K^{(N)} A_N^{(N)}) E[\eta^{(N-1)} \eta^{(N-1)T}](I - K^{(N)} A_N^{(N)})^T - (I - K^{(N)} A_N^{(N)}) E[\eta^{(N-1)} \varepsilon_N^{(N)T}] K^{(N)T} \\ &- K^{(N)} E[\varepsilon_N^{(N)} \eta^{(N-1)T}](I - K^{(N)} A_N^{(N)})^T + K^{(N)} E[\varepsilon_N^{(N)} \varepsilon_N^{(N)T}] K^{(N)T} \end{aligned} \\
&= (I - K^{(N)} A_N^{(N)}) P^{(N-1)} (I - K^{(N)} A_N^{(N)})^T + \sigma_N^2 K^{(N)} K^{(N)T} && \text{equation 4}
\end{aligned}
$$

which makes use of :

$$
\begin{aligned}
E[\eta^{(N-1)} \eta^{(N-1)T}] &= P^{(N-1)} \\
E[\eta^{(N-1)} \varepsilon_N^{(N)T}] &= E[\eta^{(N-1)}] E[\varepsilon^{(N)T}] = E[\eta^{(N-1)}] \times 0 = 0 && \text{since } \Sigma \text{ is diagonal} \\
E[\varepsilon_N^{(N)} \eta^{(N-1)T}] &= E[\varepsilon^{(N)}] E[\eta^{(N-1)T}] = 0 \times E[\eta^{(N-1)T}] = 0 && \text{since } \Sigma \text{ is diagonal} \\
E[\varepsilon_N^{(N)} \varepsilon_N^{(N)T}] &= \sigma_N^2
\end{aligned}
$$

Taking derivative of objective function L wrt K, set it to zero.

$$
\begin{aligned}
0 &= \frac{\partial L}{\partial K} \\
&= \frac{\partial}{\partial K} tr(P^{(N)}) \\
&= \frac{\partial}{\partial K} tr((I - K^{(N)} A_N^{(N)}) P^{(N-1)} (I - K^{(N)} A_N^{(N)})^T + \sigma_N^2 K^{(N)} K^{(N)T}) \\
&= \frac{\partial}{\partial K} tr((I - K^{(N)} A_N^{(N)}) P^{(N-1)} (I - K^{(N)} A_N^{(N)})^T) + \frac{\partial}{\partial K} \sigma_N^2 tr(K^{(N)} K^{(N)T}) \\
&= 2(I - K^{(N)} A_N^{(N)}) P^{(N-1)} (-A_N^{(N)T}) + 2\sigma_N^2 K^{(N)} && \text{since } \frac{\partial}{\partial A} tr(ABA^T) = 2AB \\
P^{(N-1)} A_N^{(N)T} &= K^{(N)} A_N^{(N)} P^{(N-1)} A_N^{(N)T} + \sigma_N^2 K^{(N)} \\
&= K^{(N)} (A_N^{(N)} P^{(N-1)} A_N^{(N)T} + \sigma_N^2) \\
K^{(N)} &= \frac{P^{(N-1)} A_N^{(N)T}}{A_N^{(N)} P^{(N-1)} A_N^{(N)T} + \sigma_N^2} = \frac{P^{(N-1)} A_N^{(N)T}}{S^{(N)}} && \text{equation 5}
\end{aligned}
$$

where $S^{(N)} = A_N^{(N)} P^{(N-1)} A_N^{(N)T} + \sigma_N^2$

Recall
- $K^{(N)}$ = M×1 matrix (gain vector)
- $P^{(N)}$ = M×M symmetrical matrix (covariance of model parameter error)
- $S^{(N)}$ = 1×1 scalar (this is not critical, its just for clear presentation)

Lets simplify the recursive formula for $P^{(N)}$, which can be done by putting equation 3 into equation 2.

$$
\begin{aligned}
P^{(N)} &= (I - K^{(N)} A_N^{(N)}) P^{(N-1)} (I - K^{(N)} A_N^{(N)})^T + K^{(N)} R^{(N)} K^{(N)T} \\
&= \begin{aligned}[t] &(I - P^{(N-1)} A_N^{(N)T} (S^{(N)})^{-1} A_N^{(N)}) P^{(N-1)} (I - P^{(N-1)} A_N^{(N)T} (S^{(N)})^{-1} A_N^{(N)})^T + \\ &P^{(N-1)} A_N^{(N)} (S^{(N)})^{-1} R^{(N)} (P^{(N-1)} A_N^{(N)} (S^{(N)})^{-1})^T \end{aligned} \\
&= \begin{aligned}[t] &P^{(N-1)} - P^{(N-1)} A_N^{(N)} (S^{(N)})^{-1} A_N^{(N)} P^{(N-1)} - P^{(N-1)} (P^{(N-1)} A_N^{(N)} (S^{(N)})^{-1} A_N^{(N)})^T + \\ &P^{(N-1)} A_N^{(N)} (S^{(N)})^{-1} A_N^{(N)} P^{(N-1)} (P^{(N-1)} A_N^{(N)} (S^{(N)})^{-1} A_N^{(N)})^T \\ &P^{(N-1)} A_N^{(N)} (S^{(N)})^{-1} R^{(N)} (P^{(N-1)} A_N^{(N)} (S^{(N)})^{-1})^T \end{aligned} && \text{expand the 1st term}
\end{aligned}
$$

$$
\begin{aligned}
& P^{(N-1)} - P^{(N-1)}A_N^{(N)}(S^{(N)})^{-1}A_N^{(N)}P^{(N-1)} - P^{(N-1)}A_N^{(N)T}(S^{(N)})^{-1}A_N^{(N)}P^{(N-1)} + \\
= \quad & P^{(N-1)}A_N^{(N)}(S^{(N)})^{-1}A_N^{(N)}P^{(N-1)}A_N^{(N)T}(S^{(N)})^{-1}A_N^{(N)}P^{(N-1)} \qquad \text{simplify transpose} \\
& P^{(N-1)}A_N^{(N)}(S^{(N)})^{-1}R^{(N)}A_N^{(N)T}(S^{(N)})^{-1}P^{(N-1)}
\end{aligned}
$$

$$
\begin{aligned}
= \quad & P^{(N-1)} - 2P^{(N-1)}A_N^{(N)}(S^{(N)})^{-1}A_N^{(N)}P^{(N-1)} + P^{(N-1)}A_N^{(N)}(S^{(N)})^{-1}(A_N^{(N)}P^{(N-1)}A_N^{(N)T} + R^{(N)})(S^{(N)})^{-1}A_N^{(N)}P^{(N-1)} \\
= \quad & P^{(N-1)} - 2P^{(N-1)}A_N^{(N)}(S^{(N)})^{-1}A_N^{(N)}P^{(N-1)} + P^{(N-1)}A_N^{(N)}(S^{(N)})^{-1}S^{(N)}(S^{(N)})^{-1}A_N^{(N)}P^{(N-1)} \\
= \quad & P^{(N-1)} - 2P^{(N-1)}A_N^{(N)}(S^{(N)})^{-1}A_N^{(N)}P^{(N-1)} + P^{(N-1)}A_N^{(N)}(S^{(N)})^{-1}A_N^{(N)}P^{(N-1)} \\
= \quad & P^{(N-1)} - P^{(N-1)}A_N^{(N)}(S^{(N)})^{-1}A_N^{(N)}P^{(N-1)} \\
= \quad & P^{(N-1)} - K^{(N)}A_N^{(N)}P^{(N-1)} \\
= \quad & (I - K^{(N)}A_N^{(N)})P^{(N-1)} \qquad \text{equation 6}
\end{aligned}
$$

Problem solved! The implementation is <u>in turn</u> done by equation 5 (for updating K using current observation), then by equation 2 (for updating X, i.e. the mean of model parameter estimation) and finally equation 6 (for updating P, i.e. the covariance of model parameter estimation).

Approach 2 – Minimization of fitting error

Approach 2 starts with minimization of sum of fitting error square. Unlike approach 1, there is no need to do 1st order derivative as we make use of the result of weighted least square, the model parameter estimation can be expressed as the product of matrix P and matrix Q, thus by finding the recursive formula for P and the recursive formula for Q, we will solve this recursive least square problem. W is the inverse of covariance matrix of data noise.

$$
\begin{aligned}
& L = (A^{(N)}X - B^{(N)})W(A^{(N)}X - B^{(N)})^T \qquad \text{where } W = \Sigma^{-1}\text{, both are diagonal} \\
\text{thus} \quad & X_{WLS}^{(N)} = (A^{(N)T}W^{(N)}A^{(N)})^{-1}(A^{(N)T}W^{(N)}B^{(N)}) = P^{(N)}Q^{(N)} \qquad \text{equation 7} \\
\text{where} \quad & P^{(N)} = (A^{(N)T}W^{(N)}A^{(N)})^{-1} \qquad \text{which is a M×M matrix} \\
& Q^{(N)} = (A^{(N)T}W^{(N)}B^{(N)}) \qquad \text{which is a M×1 matrix}
\end{aligned}
$$

Lets derive the recursive formula for $P^{(N)}$.

$$
\begin{aligned}
P^{(N)} &= (A^{(N)T}W^{(N)}A^{(N)})^{-1} \\
&= (A^{(N-1)T}W^{(N-1)}A^{(N-1)} + w_N A_N^{(N)T}A_N^{(N)})^{-1} && \text{where } w_n = 1/\Sigma_n = 1/\sigma_n^2 \\
&= ((P^{(N-1)})^{-1} + w_N A_N^{(N)T}A_N^{(N)})^{-1} \\
&= P^{(N-1)} - \frac{w_N P^{(N-1)}A_N^{(N)T}A_N^{(N)}P^{(N-1)}}{1 + w_N A_N^{(N)}P^{(N-1)}A_N^{(N)T}} && \text{using Sherman Morrison formula} \\
&= P^{(N-1)} - \frac{P^{(N-1)}A_N^{(N)T}}{A_N^{(N)}P^{(N-1)}A_N^{(N)T} + \sigma_N^2}A_N^{(N)}P^{(N-1)} && (A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u} \\
&= (I - K^{(N)}A_N^{(N)})P^{(N-1)} && \text{hence we have recursive formula for } P^{(N)} \\
\text{where} \quad K^{(N)} &= \frac{P^{(N-1)}A_N^{(N)T}}{A_N^{(N)}P^{(N-1)}A_N^{(N)T} + \sigma_N^2} && \text{defined as the gain vector}
\end{aligned}
$$

Lets derive the recursive formula for $Q^{(N)}$.

$$
\begin{aligned}
Q^{(N)} &= (A^{(N)T}W^{(N)}B^{(N)}) \\
&= (A^{(N-1)T}W^{(N-1)}B^{(N-1)}) + w_N b_N^{(N)}A_N^{(N)T} && \text{note : } b_N^{(N)} \text{ is a scalar} \\
&= Q^{(N-1)} + w_N b_N^{(N)}A_N^{(N)T} \\
&= Q^{(N-1)} + (b_N^{(N)}/\sigma_N^2)A_N^{(N)T} && \text{equation 8}
\end{aligned}
$$

We have defined the gain vector K above, which is :

$$
\begin{aligned}
K^{(N)} &= \frac{P^{(N-1)}A_N^{(N)T}}{A_N^{(N)}P^{(N-1)}A_N^{(N)T} + \sigma_N^2} && \text{we will further simplify it for future use} \\
K^{(N)}\sigma_N^2 &= P^{(N-1)}A_N^{(N)T} - K^{(N)}A_N^{(N)}P^{(N-1)}A_N^{(N)T}
\end{aligned}
$$

$$= (I - K^{(N)} A_N^{(N)}) P^{(N-1)} A_N^{(N)T} \quad \text{since } P^{(N)} = (I - K^{(N)} A_N^{(N)}) P^{(N-1)}$$

$$= P^{(N)} A_N^{(N)T}$$

$$K^{(N)} = P^{(N)} A_N^{(N)T} / \sigma_N^2 \quad \text{equation 9}$$

Lets derive the recursive formula for $X_{WLS}^{(N)}$.

$$X_{WLS}^{(N)} = P^{(N)} Q^{(N)} \quad \text{using equation 7}$$

$$= P^{(N)} (Q^{(N-1)} + (b_N^{(N)} / \sigma_N^2) A_N^{(N)T}) \quad \text{using equation 8}$$

$$= P^{(N)} Q^{(N-1)} + (b_N^{(N)} / \sigma_N^2) P^{(N)} A_N^{(N)T}$$

$$= P^{(N)} Q^{(N-1)} + b_N^{(N)} K^{(N)} \quad \text{using equation 9}$$

$$= ((I - K^{(N)} A_N^{(N)}) P^{(N-1)}) Q^{(N-1)} + b_N^{(N)} K^{(N)} \quad \text{using recursive equation for } P^{(N)}$$

$$= (I - K^{(N)} A_N^{(N)}) X_{WLS}^{(N-1)} + b_N^{(N)} K^{(N)} \quad \text{using equation 7}$$

$$= X_{WLS}^{(N-1)} - K^{(N)} A_N^{(N)} X_{WLS}^{(N-1)} + b_N^{(N)} K^{(N)}$$

$$= X_{WLS}^{(N-1)} + K^{(N)} (b_N^{(N)} - A_N^{(N)} X_{WLS}^{(N-1)}) \quad \text{note : } A_N^{(N)} X_{WLS}^{(N)} \text{ and } b_N^{(N)} \text{ are scalar}$$

$$= X_{WLS}^{(N-1)} + K^{(N)} e_N^{(N)} \quad \text{recursive equation for } X^{(N)}$$

where $e_N^{(N)} = b_N^{(N)} - A_N^{(N)} X_{WLS}^{(N-1)}$

Hence we get exactly the same set of updating equations as approach 1. Now lets consider a very common special case, in which we consider a forgetting factor instead of the covariance matrix Σ. Similar RLS is obtained.

Approach 2 – Minimization of fitting error (Specical case : with forgetting factor)

Given time series $\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_N$, each is assigned with a weight λ^{N-n}, where $0<\lambda<1$, that is a lighter weight for older data, λ is called forgetting factor. Hence we have :

$$A^{(N)} = \begin{bmatrix} A_1^{(N)} \\ A_2^{(N)} \\ A_3^{(N)} \\ \ldots \\ A_M^{(N)} \\ A_{M+1}^{(N)} \\ \ldots \\ A_N^{(N)} \end{bmatrix} = \begin{bmatrix} \alpha_1 & 0 & 0 & \ldots & 0 \\ \alpha_2 & \alpha_1 & 0 & \ldots & 0 \\ \alpha_3 & \alpha_2 & \alpha_1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \alpha_M & \alpha_{M-1} & \alpha_{M-2} & \ldots & \alpha_1 \\ \alpha_{M+1} & \alpha_M & \alpha_{M-1} & \ldots & \alpha_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \alpha_N & \alpha_{N-1} & \alpha_{N-2} & \ldots & \alpha_{N-M+1} \end{bmatrix} \qquad B^{(N)} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \ldots \\ b_M \\ b_{M+1} \\ \ldots \\ b_N \end{bmatrix}$$

and $W^{(N)} = diag(\lambda^{N-1} \quad \ldots \quad \lambda^2 \quad \lambda^1 \quad \lambda^0)$

and $L = (A^{(N)} X - B^{(N)}) W (A^{(N)} X - B^{(N)})^T$

$$= \sum_{n=1}^{N} \lambda^{N-n} (A_n^{(N)} X - b_n^{(N)})^2$$

The main difference from previous setting is the weight matrix. Here we have :

$$w_n^{(N)} = w_n^{(N-1)} \quad \forall n \in [1, N-1] \quad \text{in previous setting}$$

$$w_n^{(N)} = w_{n-1}^{(N-1)} \quad \forall n \in [2, N] \quad \text{in current setting}$$

Lets derive the recursive formula for $P^{(N)}$.

$$P^{(N)} = (\sum_{n=1}^{N} \lambda^{N-n} A_n^{(N)T} A_n^{(N)})^{-1}$$

$$= (\sum_{n=1}^{N-1} \lambda^{N-n} A_n^{(N)T} A_n^{(N)} + A_N^{(N)T} A_N^{(N)})^{-1}$$

$$= (\sum_{n=1}^{N-1} \lambda^{N-n} A_n^{(N-1)T} A_n^{(N-1)} + A_N^{(N)T} A_N^{(N)})^{-1} \quad \text{since } A_n^{(N-1)} = A_n^{(N)} \quad \forall n \in [1, N-1]$$

$$= [\lambda (\sum_{n=1}^{N-1} \lambda^{(N-1)-n} A_n^{(N-1)T} A_n^{(N-1)}) + A_N^{(N)T} A_N^{(N)}]^{-1}$$

$$= [\lambda P^{(N)-1} + A_N^{(N)T} A_N^{(N)}]^{-1}$$

$$= \lambda^{-1} P^{(N-1)} - \frac{\lambda^{-1} P^{(N-1)} A_N^{(N)T} A_N^{(N)} \lambda^{-1} P^{(N-1)}}{1 + \lambda^{-1} A_N^{(N)} P^{(N-1)} A_N^{(N)T}} \quad \text{since Sherman Morrison formula}$$

$$= \lambda^{-1}P^{(N-1)} - \frac{P^{(N-1)}A_N^{(N)T}}{A_N^{(N)}P^{(N-1)}A_N^{(N)T} + \lambda} A_N^{(N)}\lambda^{-1}P^{(N-1)}$$

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}$$

$$= \lambda^{-1}(I - K^{(N)}A_N^{(N)})P^{(N-1)}$$ recursive equation for P

where $K^{(N)} = \frac{P^{(N-1)}A_N^{(N)T}}{A_N^{(N)}P^{(N-1)}A_N^{(N)T} + \lambda}$ defined as the gain vector

Lets derive the recursive formula for $Q^{(N)}$.

$$Q^{(N)} = \sum_{n=1}^{N} \lambda^{N-n} b_n^{(N)} A_n^{(N)T}$$

$$= \sum_{n=1}^{N-1} \lambda^{N-n} b_n^{(N)} A_n^{(N)T} + b_N^{(N)} A_N^{(N)T}$$

$$= \sum_{n=1}^{N-1} \lambda^{N-n} b_n^{(N-1)} A_n^{(N-1)T} + b_N^{(N)} A_N^{(N)T}$$ since $b_n^{(N-1)} = b_n^{(N)} \quad \forall n \in [1, N-1]$

$$= \lambda \sum_{n=1}^{N-1} \lambda^{(N-1)-n} b_n^{(N-1)} A_n^{(N-1)T} + b_N^{(N)} A_N^{(N)T}$$

$$= \lambda Q^{(N-1)} + b_N^{(N)} A_N^{(N)T}$$ equation 10

We have defined the gain vector K above, which is :

$$K^{(N)} = \frac{P^{(N-1)}A_N^{(N)T}}{A_N^{(N)}P^{(N-1)}A_N^{(N)T} + \lambda}$$ recursive equation for $K^{(N)}$ (done!!)

$$K^{(N)}\lambda = P^{(N-1)}A_N^{(N)T} - K^{(N)}A_N^{(N)}P^{(N-1)}A_N^{(N)T}$$

$$= (I - K^{(N)}A_N^{(N)})P^{(N-1)}A_N^{(N)T}$$ since $P^{(N)} = \lambda^{-1}(I - K^{(N)}A_N^{(N)})P^{(N-1)}$

$$= \lambda P^{(N)}A_N^{(N)T}$$

$$K^{(N)} = P^{(N)}A_N^{(N)T}$$ equation 11

Lets derive the recursive formula for $X_{WLS}^{(N)}$.

$$X_{WLS}^{(N)} = P^{(N)}Q^{(N)}$$ using equation 7

$$= P^{(N)}(\lambda Q^{(N-1)} + b_N^{(N)}A_N^{(N)T})$$ using equation 10

$$= \lambda P^{(N)}Q^{(N-1)} + b_N^{(N)}P^{(N)}A_N^{(N)T}$$

$$= \lambda P^{(N)}Q^{(N-1)} + b_N^{(N)}K^{(N)}$$ using equation 11

$$= \lambda\lambda^{-1}(I - K^{(N)}A_N^{(N)})P^{(N-1)}Q^{(N-1)} + b_N^{(N)}K^{(N)}$$ using recursive formula for $P^{(N)}$

$$= (I - K^{(N)}A_N^{(N)})X_{WLS}^{(N-1)} + b_N^{(N)}K^{(N)}$$ using equation 7

$$= X_{WLS}^{(N-1)} - K^{(N)}A_N^{(N)}X_{WLS}^{(N-1)} + K^{(N)}b_N^{(N)}$$ note : $A_N^{(N)}X_{WLS}^{(N)}$ and $b_N^{(N)}$ are scalar

$$= X_{WLS}^{(N-1)} + K^{(N)}(b_N^{(N)} - A_N^{(N)}X_{WLS}^{(N-1)})$$

$$= X_{WLS}^{(N-1)} + K^{(N)}e_N^{(N)}$$ recursive equation for $X^{(N)}$

where $e_N^{(N)} = b_N^{(N)} - A_N^{(N)}X_{WLS}^{(N-1)}$

Implementation

Lets summarize the algorithm. We need to initialize $X^{(0)}$ and $P^{(0)}$ only, and start iterating from n=1.

$$X_{WLS}^{(0)} = E(X)$$ i.e. prior knowledge, expected value of X

$$P^{(0)} = E[\eta^{(N)}\eta^{(N)T}] = E[(X - X_{WLS}^{(0)})(X - X_{WLS}^{(0)})^T]$$ i.e. prior knowledge, expected rms error

If we have no prior knowledge, we can simply use any random initial guess for X, and set RMS error infinity large.

$$X_{WLS}^{(0)} = random(M \times 1)$$

$$P^{(0)} = \infty \times I$$

We then update K, X and P in a recursive manner.

$$K^{(N)} = \frac{P^{(N-1)} A_N^{(N)T}}{A_N^{(N)} P^{(N-1)} A_N^{(N)T} + \sigma_N^2}$$ where K is used for current iteration

$$X_{WLS}^{(N)} = X_{WLS}^{(N-1)} + K^{(N)} e_n^{(N)}$$ where $e_n^{(N)} = b_n^{(N)} - A_N^{(N)} X_{WLS}^{(N-1)}$

$$P^{(N)} = (I - K^{(N)} A_N^{(N)}) P^{(N-1)}$$ where P is used for next iteration

Remarks for recursive least square

(1) Can we derive RLS for non diagonal Σ, i.e. when $\varepsilon^{(N)}$ correlates with $\eta^{(N-1)}$?.

(2) Lets show that RLS estimation is unbiased.

$$\begin{aligned} E[\eta^{(N)}] &= E[(I - K^{(N)} A_N^{(N)})\eta^{(N-1)} - K^{(N)}\varepsilon^{(N)}] && \text{using equation 3} \\ &= (I - K^{(N)} A_N^{(N)}) E[\eta^{(N-1)}] - K^{(N)} E[\varepsilon^{(N)}] \\ &= (I - K^{(N)} A_N^{(N)}) E[\eta^{(N-1)}] && \text{if } E[\varepsilon^{(N)}] = 0 \\ &= \prod_{n=2}^{N} (I - K^{(n)} A_n^{(n)}) E[\eta^{(1)}] \\ &= 0 && \text{if } E[\eta^{(1)}] = 0 \end{aligned}$$

$\Rightarrow$ Hence, if the first estimation is unbiased, all subsequent estimations are unbiased.

(3) Meaning of matrix P in both approaches
From approach 1 and 2, we have two different meanings for matrix P :

- In approach 1 $P = E[\eta\eta^T]$ i.e. covariance of model parameter error
- In approach 2 $P = (A^T\Sigma^{-1}A)^{-1}$ i.e. inverse of covariance of data matrix

Are they equivalent eventually? The answer is yes!

$$\begin{aligned} E[\eta\eta^T] &= E[(X_{WLS} - X)(X_{WLS} - X)^T] \\ &= E[(X_{WLS} - E[X_{WLS}])(X_{WLS} - E[X_{WLS}])^T] && \text{since recursive least square is unbiased} \\ &= \mathrm{cov}[X_{WLS}] \\ &= (A^T\Sigma^{-1}A)^{-1} && \text{please read bias and variance analysis of } X_{WLS} \end{aligned}$$

(4) RLS is about growing set of data. How about growing set of regressor (variable)? Suppose we have N data in $\Re^{M-1}$:

$$\begin{aligned} AX &= B && \text{A is a N×(M–1) matrix} \\ QRX &= B && \text{by QR decomposition} \\ X &= R^{-1}Q^T B && \text{by backward substitution} \end{aligned}$$

where A_n = n[th] row vector (data)

A_m = m[th] column vector (regressor)

Now a new regressor A_M (a column vector) is added to the system, i.e. a new variable is introduced to the system, the dimension of regression space is enhanced from $\Re^{M-1}$ to $\Re^M$, then the system can still be solved recursively using least square by adding new column vectors Q_M and R_M to matrix Q and R respectively :

$$Q_M = \frac{A_M - \sum_{m=1}^{M-1} proj_{Q_m}(A_M)}{\left\| A_M - \sum_{m=1}^{M-1} proj_{Q_m}(A_M) \right\|}$$

$$= \frac{A_M - \sum_{m=1}^{M-1} Q_m({Q_m}^T A_M)}{\left\| A_M - \sum_{m=1}^{M-1} Q_m({Q_m}^T A_M) \right\|}$$ if ${Q_m}^T Q_m = 1$, i.e. orthonormal basis

$$= \frac{A_M - \sum_{m=1}^{M-1} Q_m({Q_m}^T Q_m)^{-1}({Q_m}^T A_M)}{\left\| A_M - \sum_{m=1}^{M-1} Q_m({Q_m}^T Q_m)^{-1}({Q_m}^T A_M) \right\|}$$ if ${Q_m}^T Q_m \neq 1$, i.e. non orthonormal basis

$$R_M = \begin{bmatrix} {A_M}^T Q_1 \\ {A_M}^T Q_2 \\ \dots \\ {A_M}^T Q_M \\ 0 \\ \dots \end{bmatrix}$$

Part C : Two Stage Least Square

In all the previous least square models, we make the assumption that each regressor (or variable) in A (i.e. columns in A) is non stochastic and of course, uncorrelated with noise ε, this is called exogenous regressor assumption.

$$E[A^T\varepsilon] = E[A^T]E[\varepsilon] = E[A^T]\times 0 = 0 \quad \text{uncorrelated assumption}$$
$$E[\varepsilon \mid A] = E[\varepsilon] = 0 \quad \text{independent assumption (stronger)}$$

Now we assume regressor to be stochastic. Suppose A is a N×M data matrix, i.e. there are N noisy observations and M regressors, among which, **some of them** are correlated with noise, they are called endogenous regressor. Please note that we are talking about **correlation between noise and regressor**, do not confuse with **correlation in noise across two observations**. In the presence of endogenous regressors, least square estimate is no longer unbiased. Lets recall :

(1) independent RV implies uncorrelated RV — i.e. $E[\varepsilon \mid A]=0 \Rightarrow E[A^T\varepsilon]=0$

(2) correlated RV implies dependent RV — i.e. $E[A^T\varepsilon]\neq 0 \Rightarrow E[\varepsilon \mid A]\neq 0$

Lets perform bias analysis and consistence analysis for the least square with existence of endogenous regressors.

$$\begin{aligned} X_{LS} &= (A^TA)^{-1}(A^TB) \\ &= (A^TA)^{-1}(A^T(AX+\varepsilon)) \\ &= (A^TA)^{-1}(A^TA)X + (A^TA)^{-1}(A^T\varepsilon) \\ &= X + (A^TA)^{-1}(A^T\varepsilon) \end{aligned}$$

Bias analysis :

$$\begin{aligned} E[X_{LS}] &= X + E[E[(A^TA)^{-1}(A^T\varepsilon)\mid A]] && \text{A is stochastic in current setting.} \\ &= X + E[(A^TA)^{-1}A^TE[\varepsilon\mid A]] \\ &\neq X && \text{using } E[\varepsilon\mid A]\neq 0 \end{aligned}$$

Consistence analysis :

$$\begin{aligned} p\lim(A^TA)/N &= p\lim(\textstyle\sum_{n=1}^{N} A_n^TA_n)/N && \text{where } A_n \text{ is the } n^{th} \text{ row of A} \\ &= E[A_n^TA_n] \quad (\text{M×M constant matrix}) && \text{using weak law of large number} \\ p\lim(A^T\varepsilon)/N &= p\lim(\textstyle\sum_{n=1}^{N} A_n^T\varepsilon_n)/N \\ &= E[A_n^T\varepsilon_n] \quad (\text{M×1 constant matrix}) && \text{using weak law of large number} \\ p\lim X_{LS} &= X + p\lim(A^TA)^{-1}(A^T\varepsilon) && \text{using Slutsky theorem} \\ &= X + p\lim(A^TA/N)^{-1}(A^T\varepsilon/N) \\ &= X + (E[A_n^TA_n])^{-1}(E[A_n^T\varepsilon_n]) && \text{using Slutsky theorem} \\ &\neq X && \text{using } E[A_n^T\varepsilon_n]\neq 0 \end{aligned}$$

Hence X_{LS} is both **biased and not consistent**!! Now we are going to introduce the solution with instrumental variable, however least square estimate with instrumental variable is still biased, yet it is consistent.

Instrumental variable

When regressor is endogenous, we can then make use of instrumental variables, these are observable variables, from which we can build new exogenous regressors that are correlated with existing endogenous regressors, and at the same time, uncorrelated with noise. Suppose the first K regressors (among M regressors) are exogenous regressors, while the other M-K regressors are endogenous regressors, and suppose that we can successfully find M-K observable instrumental variables, then we can build a new matrix Z by replacing the last M-K columns in A with instrumental variables.

$$A = [A_1 \;\; A_2 \;\; \ldots \;\; A_K \;\; A_{K+1} \;\; A_{K+2} \;\; \ldots \;\; A_M] \quad \text{where } A_m \text{ is the } m^{th} \text{ column regressor}$$
$$Z = [A_1 \;\; A_2 \;\; \ldots \;\; A_K \;\; \Lambda_1 \;\; \Lambda_2 \;\; \ldots \;\; \Lambda_{M-K}] \quad \text{where } \Lambda_m \text{ is the } m^{th} \text{ instrumental variable}$$
$$\text{then } X_{IV} = (Z^TA)^{-1}(Z^TB) \quad \text{least square with instrumental variable}$$

where
$$E[A^TZ] \neq E[A]^TE[Z] \quad \text{A and Z are correlated}$$
$$E[A^T\varepsilon] \neq E[A]^TE[\varepsilon] \quad \text{A and } \varepsilon \text{ are correlated}$$
$$E[Z^T\varepsilon] = E[Z]^TE[\varepsilon] = E[Z]^T0 = 0 \quad \text{Z and } \varepsilon \text{ are uncorrelated}$$

Lets perform bias analysis and consistence analysis for the least square with instrumental variable.

$$\begin{aligned} X_{IV} &= (Z^T A)^{-1}(Z^T B) \\ &= (Z^T A)^{-1}(Z^T (AX+\varepsilon)) \\ &= (Z^T A)^{-1}(Z^T A)X + (Z^T A)^{-1}(Z^T \varepsilon) \\ &= X + (Z^T A)^{-1}(Z^T \varepsilon) \end{aligned}$$

Bias analysis :

$$\begin{aligned} E[X_{IV}] &= X + E[E[(Z^T A)^{-1}(Z^T \varepsilon) \mid A, Z]] && \text{A is stochastic in current setting.} \\ &= X + E[(Z^T A)^{-1} Z^T E[\varepsilon \mid A, Z]] \\ &\neq X && \text{using } E[\varepsilon \mid A] \neq 0 \text{, still biased} \end{aligned}$$

Consistence analysis :

$$\begin{aligned} p\lim(Z^T A)/N &= p\lim(\textstyle\sum_{n=1}^{N} Z_n{}^T A_n)/N && && \text{where } Z_n \text{ is the } n^{th} \text{ row of Z} \\ &= E[Z_n{}^T A_n] && \text{(M×M cons matrix)} && \text{using weak law of large number} \\ p\lim(Z^T \varepsilon)/N &= p\lim(\textstyle\sum_{n=1}^{N} Z_n{}^T \varepsilon_n)/N \\ &= E[Z_n{}^T \varepsilon_n] && \text{(M×1 const matrix)} && \text{using weak law of large number} \\ p\lim X_{IV} &= X + p\lim(Z^T A)^{-1}(Z^T \varepsilon) && && \text{using Slutsky theorem} \\ &= X + p\lim(Z^T A/N)^{-1}(Z^T \varepsilon/N) \\ &= X + (E[Z_n{}^T A_n])^{-1}(E[Z_n{}^T \varepsilon_n]) && && \text{using Slutsky theorem} \\ &= X && && \text{using } E[Z_n{}^T \varepsilon_n] = 0 \text{, it is consistent!!} \end{aligned}$$

Hence, least square estimate with instrumental variable is still biased, yet it is consistent. Least square estimate with instrumental variable can be interpreted as the minimization of the following objective function.

$$X_{IV} = \arg\min_X (AX-B)^T (ZX-B)$$

Two stage least square

Now suppose we have more intrumental variables than regressors, i.e. K>M, and if we do not want to discard any of available instruments, then we have to make use of two stage least square. Suppose Z is a N×K matrix composed by stacking all K instrumental variables, then we have :

$$\begin{aligned} A &= [A_1 \quad A_2 \quad A_3 \quad \dots \quad A_M] && \text{where } A_m \text{ is the } m^{th} \text{ column regressor} \\ Z &= [\Lambda_1 \quad \Lambda_2 \quad \Lambda_3 \quad \dots \quad \Lambda_K] && \text{where } \Lambda_k \text{ is the } k^{th} \text{ instrumental variable} \end{aligned}$$

then

$$\begin{aligned} X_{2SLS} &= (A^T Z(Z^T Z)^{-1} Z^T A)^{-1}(A^T Z(Z^T Z)^{-1} Z^T B) \\ &= (A^T (Z(Z^T Z)^{-1} Z^T) A)^{-1}(A^T (Z(Z^T Z)^{-1} Z^T) B) \\ &= (A^T P A)^{-1}(A^T P B) \\ P &= Z(Z^T Z)^{-1} Z^T && \Re^N \to \Re^N \text{ projection matrix for regressors of A} \end{aligned}$$

where

$$\begin{aligned} E[A^T Z] &\neq E[A]^T E[Z] && \text{A and Z are correlated} \\ E[A^T \varepsilon] &\neq E[A]^T E[\varepsilon] && \text{A and } \varepsilon \text{ are correlated} \\ E[Z^T \varepsilon] &= E[Z]^T E[\varepsilon] && \text{Z and } \varepsilon \text{ are uncorrelated} \\ &= E[Z]^T 0 = 0 \end{aligned}$$

Lets perform bias analysis and consistence analysis for the two stage least square.

$$\begin{aligned} X_{2SLS} &= (A^T P A)^{-1}(A^T P B) \\ &= (A^T P A)^{-1}(A^T P (AX+\varepsilon)) \\ &= (A^T P A)^{-1}(A^T P A)X + (A^T P A)^{-1}(A^T P \varepsilon) \\ &= X + (A^T P A)^{-1}(A^T P \varepsilon) \end{aligned}$$

Bias analysis :

$$E[X_{2SLS}] = X + E[E[(A^T PA)^{-1}(A^T P\varepsilon) \mid A, Z]]$$ A is stochastic in current setting.

$$= X + E[(A^T PA)^{-1} A^T PE[\varepsilon \mid A, Z]]$$

$$\neq X$$ using $E[\varepsilon \mid A] \neq 0$, still biased

Consistence analysis :

$$p\lim(Z^T A)/N = p\lim(\textstyle\sum_{n=1}^{N} Z_n{}^T A_n)/N$$

$$= E[Z_n{}^T A_n]$$ (K×M cons matrix) using weak law of large number

$$p\lim(Z^T \varepsilon)/N = p\lim(\textstyle\sum_{n=1}^{N} Z_n{}^T \varepsilon_n)/N$$

$$= E[Z_n{}^T \varepsilon_n]$$ (K×1 const matrix) using weak law of large number

$$p\lim(Z^T Z)/N = p\lim(\textstyle\sum_{n=1}^{N} Z_n{}^T Z_n)/N$$

$$= E[Z_n{}^T Z_n]$$ (K×K const matrix) using weak law of large number

$$p\lim\left(\frac{A^T Z}{N}\left(\frac{Z^T Z}{N}\right)^{-1}\frac{Z^T A}{N}\right) = E[A_n{}^T Z_n]E[Z_n{}^T Z_n]E[Z_n{}^T A_n]$$

$$p\lim\left(\frac{A^T Z}{N}\left(\frac{Z^T Z}{N}\right)^{-1}\frac{Z^T \varepsilon}{N}\right) = E[A_n{}^T Z_n]E[Z_n{}^T Z_n]E[Z_n{}^T \varepsilon_n]$$

$$p\lim X_{2SLS} = X + p\lim(A^T Z(Z^T Z)^{-1} Z^T A)^{-1}(A^T Z(Z^T Z)^{-1} Z^T \varepsilon)$$

$$= X + p\lim\left(\frac{A^T Z}{N}\left(\frac{Z^T Z}{N}\right)^{-1}\frac{Z^T A}{N}\right)^{-1}\left(\frac{A^T Z}{N}\left(\frac{Z^T Z}{N}\right)^{-1}\frac{Z^T \varepsilon}{N}\right)$$

$$= X + (E[A_n{}^T Z_n]E[Z_n{}^T Z_n]E[Z_n{}^T A_n])^{-1}(E[A_n{}^T Z_n]E[Z_n{}^T Z_n]E[Z_n{}^T \varepsilon_n])$$

$$= X + (E[A_n{}^T Z_n]E[Z_n{}^T Z_n]E[Z_n{}^T A_n])^{-1}(E[A_n{}^T Z_n]E[Z_n{}^T Z_n] \times 0)$$

$$= X$$

Interpretation 1 : An exogenizing step followed by a least square in exogenized domain

Two stage least square estimate can be regarded as the result of two stages. In stage one, we perform M regressions in K dimensional space together, by solving for Y in the following linear system :

$$A = ZY + \eta$$

which can be regarded as performing M regressions together, i.e. for all m∈[1,M], we have :

$$A_m = ZY_m + \eta_m$$

where A = N×M matrix = $[A_1 \;\; A_2 \;\; A_3 \;\; \ldots \;\; A_M]$ = data in stage one

Z = N×K matrix = data in stage one

Y = K×M matrix = $[Y_1 \;\; Y_2 \;\; Y_3 \;\; \ldots \;\; Y_M]$ = model parameter in stage one

η = N×M matrix = $[\eta_1 \;\; \eta_2 \;\; \eta_3 \;\; \ldots \;\; \eta_M]$ = uncorrelated noise

Therefore Y_m is the weight vector, so that the linear combination of instrumental variables ZY_m gives the endogenous regressor A_m. Hence the least square estimate for the stage one is :

$$Y_{2SLS} = (Z^T Z)^{-1} Z^T A$$

then $$ZY_{2SLS} = Z(Z^T Z)^{-1} Z^T A$$

$$= PA$$ (i.e. projection of regressors on space spanned by instrumental variables)

In stage two, we perform regression in M dimensional space, by solving for X in the following linear system :

$$B = (PA)X + \varepsilon$$

$$X_{2SLS} = ((PA)^T PA)^{-1}((PA)^T PB)$$

$$= (A^T PPA)^{-1}(A^T PPB)$$ since P is symmetric

$$= (A^T PA)^{-1}(A^T PB)$$ since PP = P

$$= (A^T Z(Z^T Z)^{-1} Z^T A)^{-1}(A^T Z(Z^T Z)^{-1} Z^T B)$$

In other words, two stage least square can be regarded as performing least square with exogenized regressors (i.e. projection of endogenous regressors on space spanned by instrumental variables).

Interpretation 2 : Generalized least square

Two stage least square estimate can also be regarded as generalized least square on a transformed problem. Lets apply a transformation Z^T on regressors, we have :

$$B = AX + \varepsilon \quad \Rightarrow \text{the original problem}$$
$$Z^T B = Z^T AX + Z^T \varepsilon \quad \Rightarrow \text{the transformed problem, with } \Re^N \rightarrow \Re^K \text{ transformation } Z^T$$

The uncorrelated noise becomes correlated, with expected value and covariance matrix :

$$\begin{aligned} E[Z^T\varepsilon] &= Z^T E[\varepsilon] \\ &= 0 \\ \text{var}[Z^T\varepsilon] &= E[(Z^T\varepsilon)(Z^T\varepsilon)^T] - E[Z^T\varepsilon]E[Z^T\varepsilon]^T \\ &= E[(Z^T\varepsilon)(Z^T\varepsilon)^T] \\ &= E[Z^T\varepsilon\varepsilon^T Z] \\ &= Z^T E[\varepsilon\varepsilon^T]Z \\ &= \Sigma \end{aligned}$$

Generalized least square gives the following solution :

$$\begin{aligned} X_{2SLS} &= ((Z^TA)^T\Sigma^{-1}(Z^TA))^{-1}((Z^TA)^T\Sigma^{-1}(Z^TB)) \\ &= ((Z^TA)^T(Z^TE[\varepsilon\varepsilon^T]Z)^{-1}(Z^TA))^{-1}((Z^TA)^T(Z^TE[\varepsilon\varepsilon^T]Z)^{-1}(Z^TB)) && \text{(equation 10)} \\ &= (\sigma^2(Z^TA)^T(Z^TZ)^{-1}(Z^TA))^{-1}(\sigma^2(Z^TA)^T(Z^TZ)^{-1}(Z^TB)) && \text{since } E[\varepsilon\varepsilon^T]=\sigma^2 I \\ &= (A^TZ(Z^TZ)^{-1}Z^TA)^{-1}(A^TZ(Z^TZ)^{-1}Z^TB) \end{aligned}$$

Remark : Can you derive the two stage least square with correlated noise? Equation 10 gives us some hints.

Reference

(1) Recursive Least Square Estimation, Com S477/577, Lowa State University. (for RLS approach 1)
(2) A Tutorial on Recursive methods in Linear Least Squares Problems, Arvind Yedla. (for RLS approach 2)
(3) Instrumental Variable and Two Stage Least Square Estimators, Hisayuki Yoshimoto.
(4) Endogenous Regressors and Instrumental Variables, James L Powell, University of California, Berkeley.