COLLINEARITY CAN BE RECOGNIZED USING PROGRESSIVELY SEQUENCED REGRESSION ANALYSIS

Prepared in response to a request by participants in the meeting of
ISO/TC131/SC8/WG13/(Math Modeling ad hoc Project), San Antonio, Texas, USA
June 20, 2018
Print Date: June 20, 2018

Jack L Johnson, 2018
Project Leader pro tem, Math Modeling ad hoc Project
Electrohydraulic Engineer
IDAS Electrohydraulics

With Support From:
Dr Jose Garcia,
Assistant Professor, Purdue University
West Lafayette, IN

Paul W Michael
Research Chemist,
Fluid Power Institute
Milwaukee School of Engineering

Pawan Panwar
Master's Candidate
Mechanical Engineering
Milwaukee School of Engineering
Milwaukee, WI

COLLINEARITY CAN BE RECOGNIZED USING PROGRESSIVELY SEQUENCED REGRESSION ANALYSIS

D:\REPORTS\RESEARCH\CollinearityCanBeRecognizedUsingPSRQAnalysis.wpd

Jack L Johnson

ABSTRACT

The original report on the method of Progressively Sequenced Regression Analysis [4] demonstrated how PSR Analysis would track the regression coefficients and chosen figures of merit for a given set of source data and a specific regression function as the number of samples in the source data increased by one observation in an iterative and progressive manner. The basic purpose was to determine when the number of samples had reached a point where further increases in number of samples caused no significant change in the model's regression coefficients and/or figures of merit. In this way the sufficient number of samples could be objectively verified. This study also revealed that the effects of collinearity between regressor terms is displayed in a graphical and clearly visible way. This paper provides an example of collinearity between regressor terms and shows how it reveals itself and can be recognized in PSR Analysis.

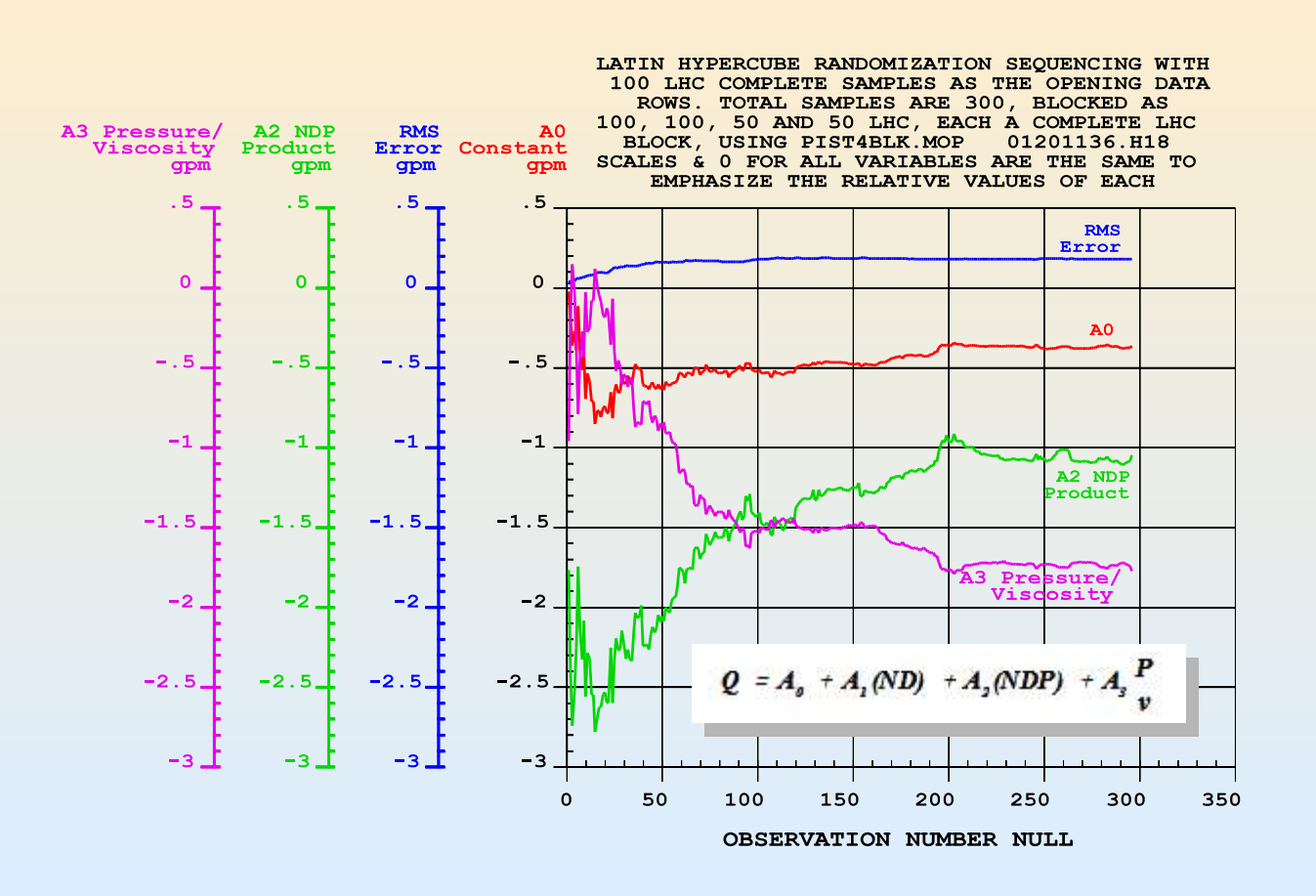

Figure 1 Collinearity between regressor terms in the regression function is signaled in the PSR plots in the mirror images between two coefficients, for example, between A_2 and A_3, above. 9J=

METHOD OF INVESTIGATION

Figure 1 is a graph of the results of Progressively Sequenced Regression Analysis (PSR Analysis) using the data file named "**Pist4Blk.mop**". It is a direct outgrowth from the original source file,"**Pist4Blk.LHC**". The means that were used to assemble this .mop file are discussed in Annex C of the internationally

distributed report regarding PSR Analysis [4]. The .mop extension on the first file simply means that the original .LHC file was subjected to some math operations. In this case, the two Pist4Blk files were modified to include some observation counters that facilitated the graphing of the data.

The original file, "**Pist4Blk.LHC**", is a collection of four different test results from four different tests that are, again, explained in Annex C. Each of the four blocks was started with the creation of a randomly sequenced test set point plan using latin hypercube (LHC) randomization. Both the creation of the test set points and the lab testing and pump data collection processes were conducted by Paul Michael and Pawan Panwar of MSOE's Fluid Power Institute Labs.

The part of the file name, "**4Blk**", is a contraction meaning "4 blocks of source data were used to create the collected '**Pist4Blk.LHC**' result". Again, the explanations of the specific four blocks are contained in Annex C of the earlier report. Because of the four constituent blocks, this file is referred to as having been blocked as a 100, 100, 50, 50 file. That is, it consists of four data blocks that consisted of 100, 100, 50 and 50 samples, respectively, totaling 300 samples (observations) in all, and each constituent block being a complete LHC entity.

After adding the serial observation numbers (it's a convenient variable for plotting the processed data), the file was subjected PSR Analysis using the following equation as the regression function:

$$Q = A_0 + A_1(ND) + A_2(NDP) + A_3\frac{P}{v} \qquad (1)$$

A_1(ND) is the ideal flow, A_2(NDP) is the flow of compressibility and A_3(P/v) is the pressure and viscosity dependent internal leakage.

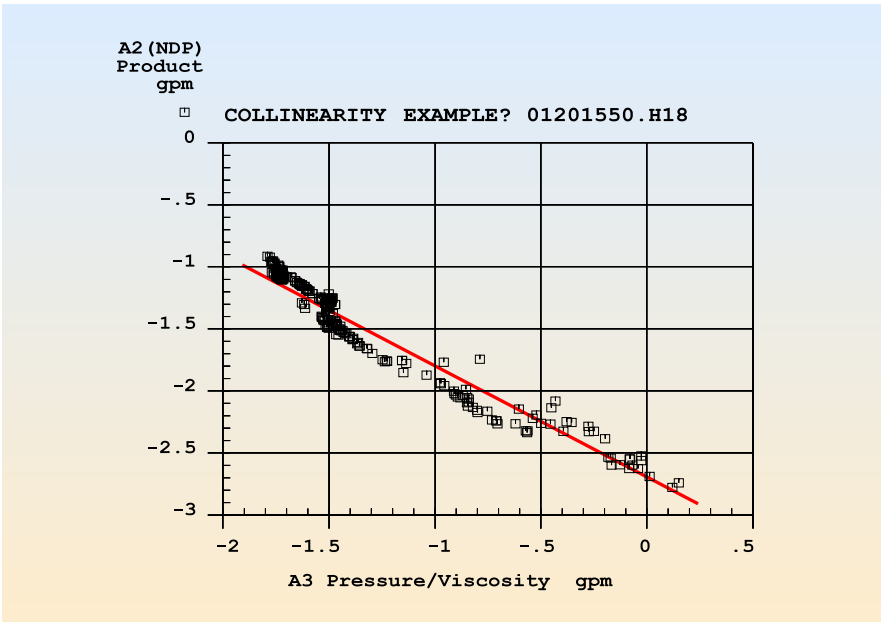

Figure 2

A scan of the coefficient plots in Figure 1 shows that A_2 and A_3 appear to be top-to-bottom mirror images of one another. When the plotting axes are changed, A_2 and A_3 can be plotted against one another. Figure 3 shows that result.

The relationship is not a perfect straight line, however, it is nearly so. It begs a question as to this being an example of collinearity in which one regressor data column can be calculated from another. If so, the model could lead to erroneous results if used to calculate pump output flow. Is this an example of collinearity?

Post-PSR Analysis Processing of Results: The aim of PSR Analysis, as stated in the original reports [4], is to track the evolution of the regression coefficients and selected FoMs as PSR moves through the source file, one sample at a time. The result is a set of coefficients that form the solutions to each regression problem. It is difficult to compare the significance of each of the coefficients because alone they have no meaning by themselves. This is caused by the fact that each regressor, being a function of non-linear combinations of pump variables, have values in some that range very high and others are very low. That

is, there may be, say, a regression coefficient that has a value of 0.000079, that on the surface may look insignificant. However if the regressor range has values that are, for example, as high as 397658.2, the product of the coefficient and the regressor can be significant indeed. A simple but effective strategy was applied that gives a better picture of the relative significance of each coefficient. In the pump of this report, the ideal flow is about 26.6 gpm.

The procedure involves converting it into a measure of its contribution to the regression dependent variable by "converting" each to engineering units of the dependent variable, in this case, to gpm. Here is how it was done: Each regressor data column (a vector) was scanned for its absolute maximum value and then multiplied by its respective regression coefficient. Recall that the coefficientschabge throughout PSR Analysis, but the peaks are in the unchanging source data file. This bit of arithmetic was done with the coefficients before making the graph of Figure 1. Additionally, all four plotted variables were forced to share a common axis scaling, further facilitating the ability for the human eye and brain to interpret and compare the relative importance of each variable. In this manner and with the graph of Figure 2 it is easily seen that the $\boldsymbol{A_2(NDP)}$ vector is linearly calculable by knowing only $\boldsymbol{A_3(P/v)}$.

ELIMINATING THE OFFENDING REGRESSOR TERM

When collinearity exists, there is essentially a redundancy in the regression terms. That is, one terms is unnecessary, but there is a question as to which one should be eliminated. On the surface, either one can be labeled as the "offending term" and threfore be eliminated, however, some thought can help to make the selection on more pragmatic basis. This author is an advocate of selecting the regression terms on the basis of known and accepted and applicable theories of the underlying physics, that is, on the basis of applying first principles. Review of the literature through the decades will lead to the formation of a list of "candidate" regression terms. The modeler can select a candidate, or candidates, and test them as to applicability and suitability, as in the case of collinearity.

Unfortunately, the inevitable and unavoidable measurement errors prevent extraction of all but the more obvious effects. This was demonstrated by Johnson [1] in Figure 61 on page 96. This graph shows how the display of internal leakage as a function of the ideal flow has an "amplifying" effect on the values. The basic model for a pump is that its output flow is the ideal flow minus internal leakage. Internal leakage is not directly measurable, but must be calculated from an estimation of the ideal flow, based on direct measurement of the actual output flow, minus the actual flow. The result is that the errors in estimating the ideal flow (displacement times speed, actually) and the errors in the measurement of the output flow, propagate directly and undiminished to the internal leakage calculated value. When the leakage is compared to the ideal flow, the error in and the scatter of the leakage are expanded in inverse proportion to the volumetric efficiency of the pump.

In the case at hand, it's increasingly difficult to absolutely separate the compressibility flow from the pressure-dependent flow due to internal clearances, especially as pumps become more efficient. Now, this is absolutely true: Certainly, both compressibility effects (***NDP***) and pressure-dependent effects (***P/v***) are both at work in the pump simultaneously. It's impossible to identify with any degree of certainty, or by any direct measurement, how much is attributable to either physical process, so one is simply eliminated.

The compressibility term will be eliminated. It is the feeling of this researcher that the leakage's direct proportionality to pressure, and its inverse proportionality to viscosity are well-known and effectively postulated in the Hagen-Poisieulle equations. A crude estimate of the maximum compressibility flow is ½ of 1% per 1000 psi based on a bulk modulus of 200,000 psi. At its maximum value, the flow loss due to compressibility in a 3500 psi pump test is less than 1/10 of 1% of maximum ideal flow. Data error simply prevents such a small flow to be extracted from the measured flow. It is insignificant.

Compressibility effects could, reasonably and rationally, be added to the total internal leakage by including a term, for example, that is based on the ½ of 1%, per the above argument.

RESULTS OF ELIMINATING THE COMPRESSIBILITY FROM THE REGRESSION FUNCTION

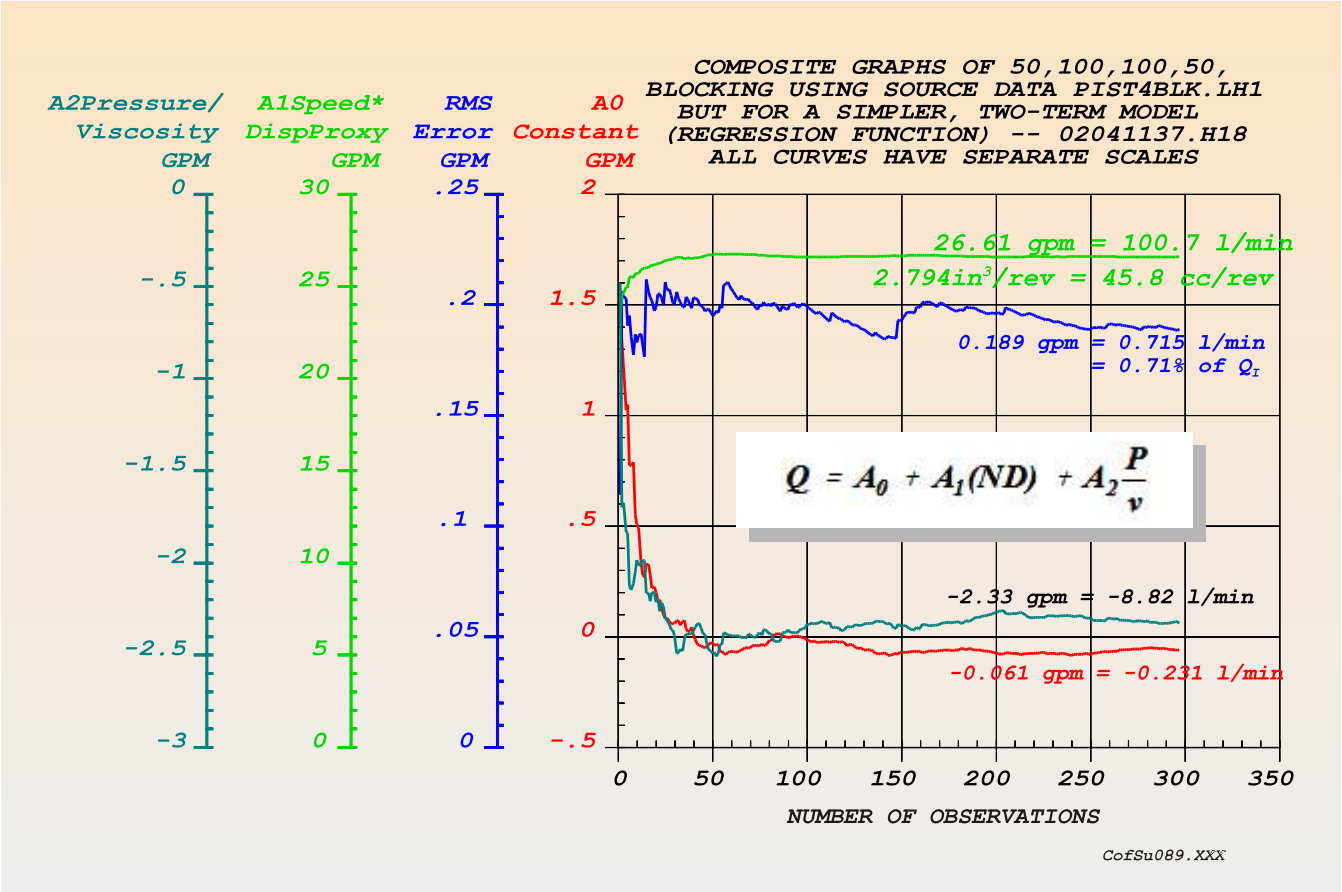

Figure 3 When the offending regressor term is deleted from the regression function, volatility of the regression coefficients is significantly reduced. 9KC

Changes to the Input Data File: One change was made to the input data file, but it is judged to have little to no affect on the results or conclusions drawn therefrom. Specifically, blocking was used as described in [5], and the blocking was changed from a file with 100,100,50,50 blocking to 50,100,100,50 which shortens the learning zone by 50%. If anything, the shortened learning zone would probably make the trends more erratic, not less.

The results of eliminating the compressibility term are shown in Figure 3. It is immediately obvious that the trajectories are better behaved, ie, less volatile than with the compressibility term in place. Surprisingly, the RMS Error (blue curve) has no significant variations, even in the learning zone (first 50 observations). This is felt to be more coincidence in that the input source data had reinforcing information content during the earliest samples such that the RMS Error was at or near the plateau right at the start of the analysis. But the coefficients do not conform as well. They require the entire learning zone for convergence. This is, perhaps, to be expected and predictable because the first data block, the learning block, has 50 samples in it.

The results in Figure 3 show no tendency of the volatility that accompanies collinearity. Furthermore, the simplicity of the regression function, the low value of A_0 (it's only0.23% of the maximum ideal flow) and

the absence of collinearity all support the notion that this is a valid and useful mathematical model of this hydraulic pump. The RMS Error is only 0.71% of the maximum ideal flow which is more less expected given that the accuracy requirements for flow measurement in ISO 4409 [3] require uncertainty to be not more than 0.5% of maximum measured flow. The calculated value is just outside that window, however, the model is influenced by ALL of the pump's variables, speed, pressure, displacement and viscosity, and not just the measured flow. In all, the data and the model are deemed to be remarkably useful and reliable.

Extrapolation of the model into operational areas that are outside the tested boundaries will have to be taken up by any ISO committee that is going to standardize mathematical models of hydraulic machinery. All purist modelers will caution against extrapolation of empirically-based models because any untested operating regions are, strictly speaking, unknown regions. Few would make any guarantees. Yet, almost all users of the models will likely press their luck from time-to-time and do analyses and simulations that are outside the tested bounds. Thus, it is incumbent upon the practical modeler to consider extrapolabilityof the model. The model that is characterized by the equation in Figure 3 is eminently extrapolable.

CONCLUSIONS

The existence of regressor collinearity reveals itself in the PSR output in two distinct ways: First, and arguably most obvious, there will be two terms that are mirror images of one another, and second, there is more volatility in the two correlated terms, Figure 1. An analysis of the data and the regression function with Minitab [2] confirmed the collinarity in the Variation Inflation Factor (VIF).

Additionally, the model is remarkably simple, a consequence of steps, not detailed in this paper, that systematically eliminated collinearity. When all the collinearity was eliminated from an orginal candidate list that included as many as 7 regressors, the final model consisted of only two regressors, and three regression coefficients, including the constant.

REFERENCES

[1] Jack L Johnson; **Two Related Monographs on Indexes and Modeling**, (Formerly referred to as **Two Part Monograph**); unpublished working document distributed to participants in The Math Modeling Project of ISO\TC131\SC8\WG13\ISO\TC-131\USTAG, 2015, IDAS Electrohydraulics (Electronic copies available on request to interested participants) 9IX

[2] http://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/multiple-regression/methods-and-formulas/methods-and-formulas/ (**Methods and Formulas for Multiple Regression**) 9J8

[3] ISO 4409, **Hydraulic fluid power - Positive displacement pumps, motors and integral transmissions - Methods of testing and presenting basic steady state performance data**, 2007, ISO copyright office, Case postale 56 • CH-1211 Geneva 20, 2007 9AM

[4] Jack L Johnson, **Progressively Sequenced Regression Helps to Establish Minimum Sample Size at Test Time**, Unpublished report distributed to active members ISO\TC131\SC8\WG13 Mathematical Modeling Ad Hoc Project, January, 2018 9KA

[5] Jack L Johnson, **A Sequel - Progressively Sequenced Regression Helps to Establish Minimum Sample Size at Test Time**, Unpublished report distributed to active members ISO\TC131\SC8\WG13 Mathematical Modeling Ad Hoc Project, February, 2018 9KB

D:\REPORTS\RESEARCH\CollinearityCanBeRecognizedUsingPSRQAnalysis.wpd