CHALMERS | UNIVERSITY OF GOTHENBURG

PREPRINT 2014:23

A real life engineering problem in Big Data solved by Pearson's correlation coefficient

SVEN AHLINDER
IVAR GUSTAFSSON

Department of Mathematical Sciences
Division of Mathematics
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg Sweden 2014

Preprint 2014:23

A real life engineering problem in Big Data solved by Pearson's correlation coefficient

Sven Ahlinder and Ivar Gustafsson

Department of Mathematical Sciences
Division of Mathematics
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg, Sweden
Gothenburg, November 2014

Preprint 2014:23
ISSN 1652-9715

Matematiska vetenskaper
Göteborg 2014

A real life engineering problem in Big Data solved by Pearson's correlation coefficient

Sven Ahlinder[1] **and Ivar Gustafsson**[2]

[1]Volvo GTT Advanced Technology and Research, Dept. BF40510 CTP9A, SE-405 08 Gothenburg, Sweden, email: Sven.Ahlinder@volvo.com, correspondent author, phone: +46 31 322 8760

[2]Department of Mathematical Sciences, Division of Mathematics, Chalmers University of Technology, Göteborg, Sweden, email: ivar@chalmers.se

1 Abstract

Over five years, Volvo has sampled 200 variables on trucks in real world use. The sampling frequency has been 10 Hz. This has given a Big Dataset with 500 million rows and 200 columns.

One of the sampled values is the fuel consumption. The wish is to see which of the sampled variables has the largest influence on fuel consumption and therefore be able to adjust those variables for lower fuel consumption.

It is not possible to perform ordinary Multiple Linear Regression (MLR) on the dataset since it is too large for our computers to handle in one piece. Instead we regress each of the 200 variables at a time against fuel consumption. This is achieved by subtraction of the average of each variable followed by application of the Gauss normal equations on each column at a time. This new technique is compared to the classical MLR in the Section 3. In our new method the "explanation ratio" is computed as the Pearson's correlation coefficient. This enables us to approximately answer the following questions;

- How does fuel consumption vary with the different variables?

- How well does the i'th variable explain fuel consumption?

The conclusions are:

1. Normal MLR is not possible on Big Data.

2. It is fairly easy to pin-point the most important variables using the method proposed in the current paper.

Key words: Big Data, Multiple Linear Regression, Multiple Bilinear Regression, Pearson's Correlation Coefficient.

2 Introduction

Datasets, particularly time dependent datasets, rapidly grow large. This is an identified problem in Big Data. Pearson's correlation coefficient is rediscovered in the current paper and applied to a real dataset. The dataset CFF is a sampling at 10 Hz of 200 variables available on the CAN-bus of Volvo trucks. The dataset was collected in the period 2008 to 2013. This takes the form of an SQL database of 200 columns and 500 million rows, some terabytes in size. The aim of the sampling is to conclude which of the sampled variables correlate with fuel consumption of the truck, in order to lower the fuel consumption.

3 Multiple Bilinear Analysis versus Multiple Linear Regression

In this article, a bilinear regression is made on fuel for each input variable, a_j. We call this Multiple Bilinear Analysis, MBA, to differ from Multiple Linear Regression, MLR.

Suppose the system of linear equations $Ax = b$. If all columns a_j of A are orthogonal to each other, MBA and MLR will give the same result vector x, since then x_j depends only on a_j and b.

The difference between the methods occurs when the columns a_j are correlated, If the correlations are strong the matrix A becomes badly conditioned, it has a large condition number. There are many methods for handling badly conditioned matrices, for instance truncated SVD, see Reference [5].

In the common solution by MLR on a badly conditioned matrix, A, the influence on b of the correlated columns will be divided between several different x_j. This leads to smaller x_j for columns a_j that are correlated to other columns, and the magnitude of x_j will be affected by adding extra columns to A.

The aim is to detect high correlations, x_j, between a_j and b. However, the concept of MLR and truncated SVD of badly conditioned matrices can exchange the lead to changes in the ranking of x_j due to effects of correlations realized previously. On the other hand, using MBA, x_j are regarded as independent so this method reveals large x_j independent on illconditioned matrix A.

4 Method

We study overdetermined systems of linear equations

$$Ax = b, \quad A \in R^{m \times n}, m > n, \; b \in R^m, \; x \in R^n. \tag{1}$$

The system has in general no solution so an approximate solution is required. Usually the least squares solution is used i.e. the solution to the minimization problem

$$min_{x \in R^n} \|Ax - b\| \tag{2}$$

This problem (2) can be solved by the Gauss normal system of equations, see Reference [1],

$$A^T Ax = A^T b \tag{3}$$

although there in general exists more numerically stable techniques like the QR-factorization, see Reference [4].

In the current study a new method to compute an approximate solution of (1) is presented. Each separate column $a_j,\ j = 1,\ \ldots\ ,\ n$ of A is considered to form a system of equations with just one column,

$$a_j x_j = b \tag{4}$$

which is solved in the least squares sense by the normal equation technique (3) i.e.

$$a_j^T a_j x_j = a_j^T b \Leftrightarrow x_j = a_j^T b / a_j^T a_j,\ j = 1,\ ...,\ n. \tag{5}$$

When not centering a_j and b, the result is the least squares solution of a proportional line. When subtracting the average, centering a_j and b, we get the affine coefficient x_j, excluding the bias term. The focus of the present study is not the bias term, but which variables (columns) a_j give the most explanation of b, therefore centering of the data is performed. The explanation ratio of each variable is estimated as the ratio between the norm of the model and the norm of the measurments, i.e.

$$p_j = \frac{x_j \|a_j\|}{\|b\|} = x_j \sqrt{\frac{a_j^T a_j}{b^T b}} = \frac{a_j^T b}{\sqrt{a_j^T a_j\ b^T b}} = \frac{a_j^T b}{\|a_j\|\ \|b\|}, \tag{6}$$

where the second (and fourth) equality follows from the definition of the standard Euclidean norm of a vector and the third equality follows from (5). This explanation ratio is the cosine of the angle between a_j and b or the Pearson's correlation coefficient, see Reference [2].

Recall that n systems are solved with just one unknown x_j in least squares sense i.e.

$$min_{x_j \in R} \|a_j x_j - b\|,\quad j = 1,\ \ldots\ ,\ n, \tag{7}$$

where the data is centered in the following manner: $b = \tilde{b} - \bar{b},\quad a_j = \tilde{a_j} - \bar{a_j},\ j = 1,\ \ldots\ ,\ n$ for original data $\tilde{b}$ and columns of the matrix $\tilde{a_j},\ j = 1,\ \ldots\ ,\ n$ and $\bar{v}$ is a notation for the mean value of a vector $\tilde{v}$.

This approach could be described as an affine regression between the result vector b and each column of A. The result of this analysis is that every column in A is judged separately and recieves separate straight line coefficients and rate of explanation of b, the Pearson's correlation coefficient between a_j and b.

The interpretation of this analysis is that columns a_j with high absolute values of the correlation coefficient are strong candidates for explaining the fuel consumption, and those variables should be further investigated. If the correlation coefficient is 95%, then the affine straight line model of this column a_j explains 95% of the fuel consumption, so that from (6)

$$p_j = \frac{x_j \|a_j\|}{\|b\|} = \frac{x_j std(a_j)}{std(b)} = 95\%, \tag{8}$$

where $std(v)$ is a notation for the standard deviation of a vector v, recall that the data is centered.

5 Data processing

The CFF dataset was stored on an SQL server. The following requests were made to the server:

(**o**) Sum of squares of each column, $\tilde{a_j}^T \tilde{a_j},\ \tilde{b}^T \tilde{b}$

(**o**) Average of each column, $\bar{a_j},\ \bar{b}$

(**o**) Feasible number of observation of each column, m_j

(**o**) Scalar product of each column with result column (fuel consumption), $\tilde{a_j}^T \tilde{b}$

This task took approximately 1 week for the 200 variables and 500 million observations. The remaining tasks were performed on a PC in a day. It is important to notice that centering of data may be performed after computing scalar products since for vectors $\tilde{v}$ and $\tilde{u}$ with length m and averages $\bar{v}$ and $\bar{u}$ the following relation holds, where e is the all ones vector:

$$(\tilde{v}-\bar{v}e)^T(\tilde{u}-\bar{u}e) = \tilde{v}^T\tilde{u}-\bar{v}e^T\tilde{u}-\bar{u}\tilde{v}^Te+\bar{v}\bar{u}e^Te = \tilde{v}^T\tilde{u}-m\bar{v}\bar{u}-m\bar{u}\bar{v}+m\bar{v}\bar{u} = \tilde{v}^T\tilde{u}-m\bar{v}\bar{u}. \tag{9}$$

This relation is called Steiner's theorem, see Reference [3]. Using (9) saves floating point operations (flops) since centering of the large columns a_j is not requested. For a matrix $A \in R^{m\times n}$ approximately $2mn$ flops are saved which is significant in the case of large datasets.

Next all values are centered using (9):

(**o**) Centered $\tilde{a_j}^T \tilde{a_j}$: $a_j^T a_j = \tilde{a_j}^T \tilde{a_j} - m_j \bar{a_j}^2$

(**o**) Centered $\tilde{a_j}^T \tilde{b}$: $a_j^T b = \tilde{a_j}^T \tilde{b} - m_j \bar{a_j} \bar{b}$

(**o**) Centered $\tilde{b}^T \tilde{b}$: $b^T b = \tilde{b}^T \tilde{b} - m_j \bar{b}^2$

Finally regression is performed, using formulas (5) and (6) from Section 4:

(**o**) Slope: $x_j = a_j^T b / a_j^T a_j$

(**o**) Bias: $bias_j = \bar{b} - x_j \bar{a_j}$

(**o**) Pearson's coefficient (absolute value): $p_j = \sqrt{\frac{x_j^2 a_j^T a_j}{b^T b}} = \frac{|a_j^T b|}{\|a_j\|\ \|b\|}$

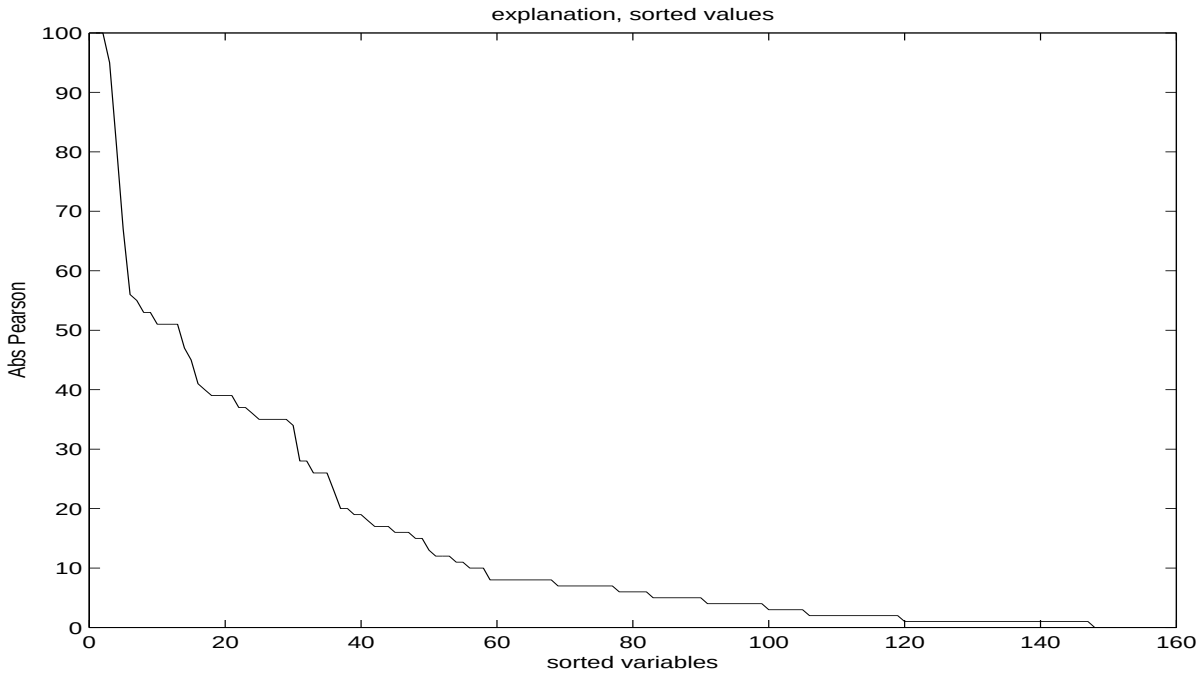

Figure 1: Explanation, all variables

6 Result

The work presented in the current paper is a regression analysis of large systems variables, amounting to several terabytes of data. As highlighted in Section 5, calculation of the required values i.e. scalar products and mean values took one week on the database server. Calculation of the linear coefficients and correlation are cheaply performed on a desk top computer in a single day. After removal of infeasible variables (i.e. those variables with few observations), 151 variables remained. The original variable numbering was retained.

In Figure 1 we present the explanation rates (Pearson's coefficients) for the 151 variables, sorted in decreasing order. We can see that 10% of the 151 variables have an explanation rate over 50%. These variables are worth further analysis, since they probably have a strong influence on fuel consumption. 60% of the variables have an explanation rate lower than 10% which make them poor candidates for further analysis.

Figure 2 shows the 12 variables with an explanation rate higher than 50% plus the 100% explanation rate of fuel consumption itself. We can see that variable number 221 also has 100% explanation rate which is expected as the variable is an alternative measure of fuel consumption. The unexpected candidate, the real "dark horse", is variable 48 with an explanation rate of 95%. This variable has never previously been identified as explaining fuel consumption.

From Figure 3, it can be seen that the regression coefficients x_j give no indication of which variables provide the best explanation of fuel consumption. Scaling the regression

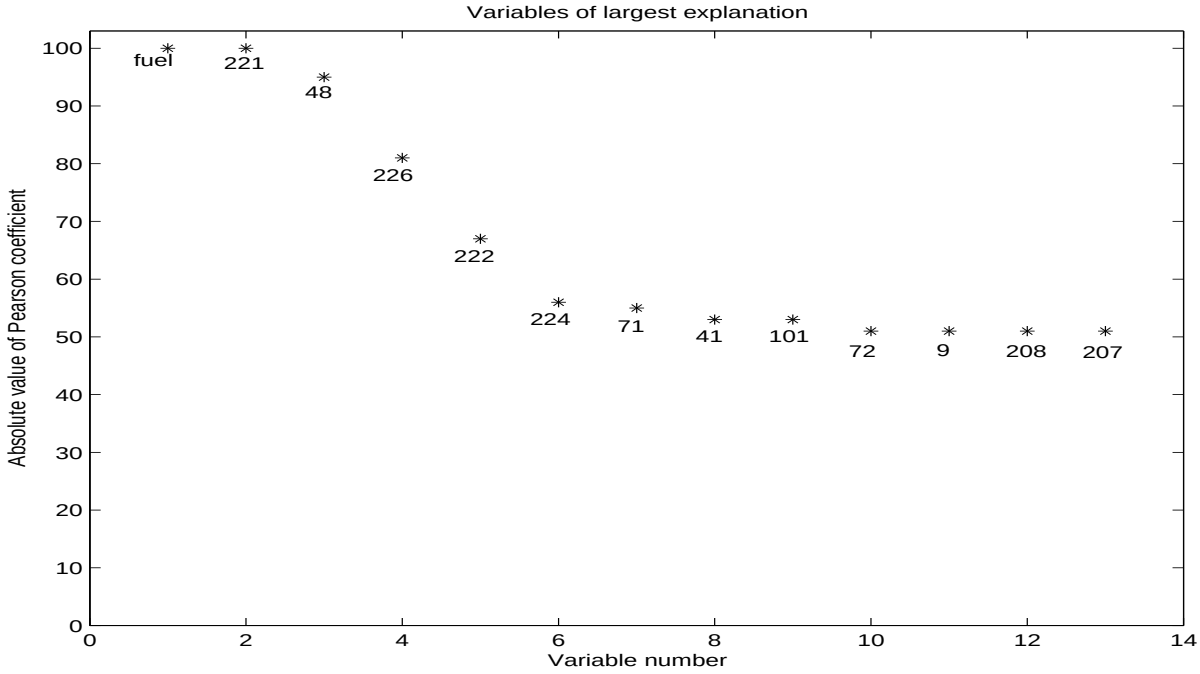

Figure 2: Variables of largest explanation - zoom in from Figure 1

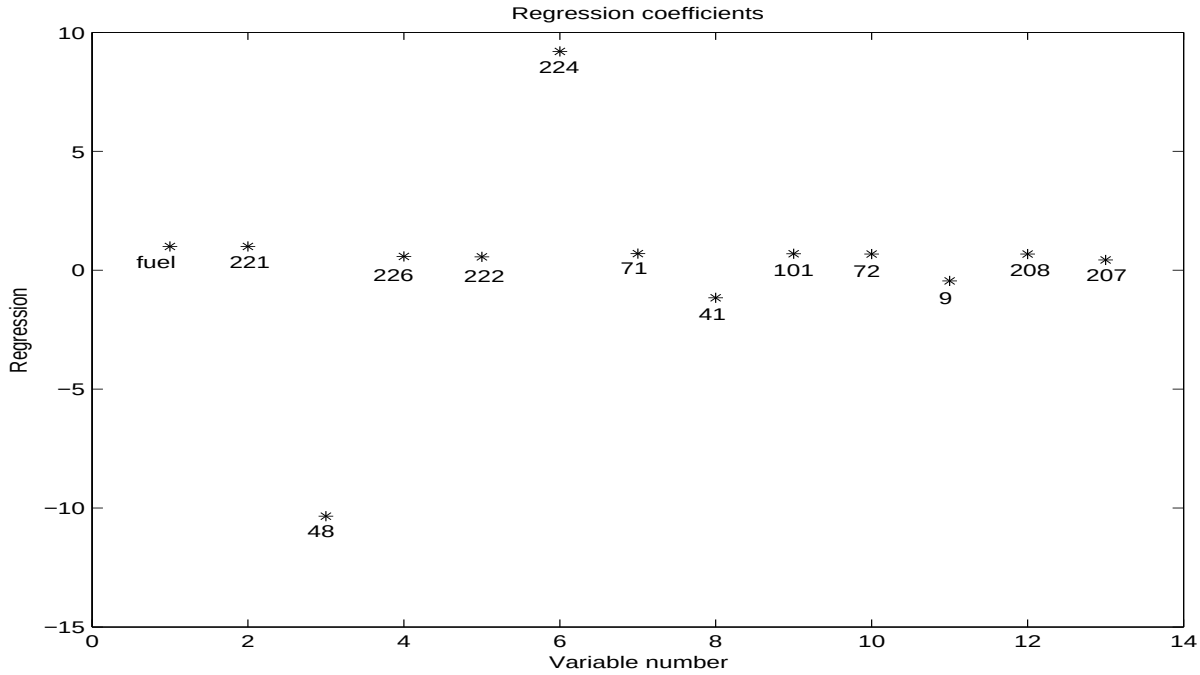

Figure 3: Regression coefficients

coefficient by the standard deviation of each variable, would solve this problem.

The regression coefficients are very valuable when understanding the physics in later analysis. Variable 48, the dark horse, has an $x_j = -10$. This means that for an increase of this variable of 0.01 units, the fuel consumption will drop 0.1 units, if everything else was constant.

7 Discussion

After gathering of the dataset CFF, amounting to over 1 terabyte data, analysis initially seemed to be unfeasable. However, by use of Pearson's correlation coefficient and regression of each varable in isolation, understanding of the relationships within the dataset, was achieved.

The method is particularly effective in the case where several datasets are incoming with the same variables. For each analysis only the four values presented in Section 5 for each variable i.e. sum of squares of each column, average of each column, feasible number of observation of each column, and scalar product of each column with result column (fuel consumption) are needed. This compresses a dataset of size nm to a dataset of size $4n$, which still contains all necessary values for future analyses. Recall that $m = 500$ millions and $n = 200$ in the present study.

8 Conclusion

The final conclusion is that the terabyte sized dataset, CFF, is analysed in a reasonable and very economic way, taking less than 1 week. A multiple linear regression of the same dataset would require many times longer, of overall computational time, estimated to over 100 weeks.

9 Acknowledgement

The authors thank Volvo Corporation for provision of the dataset CFF, and for sponsoring the development work contained in the current paper.

10 References

[1] Box, Hunter, Hunter. Statistics for Experimenters ISBN 0-417-09315-7 1978 John Wiley & Sons, Inc.
[2] Larsen, Marx. An Introduction to Mathematical Statistics and Applications ISBN 0-13-487174-X 1986 Prentice-Hall.

[3] Råde, Westergren. Matematics Handbook ISBN 91-44-25053-3 1995 Studentlitteratur, Lund Third edition.
[4] Ahlinder, Gustafsson. On some numerical methods for solving highly overdermined systems of linear equations, in progress.
[5] Demmel. Applied Numerical Linerar Algebra, Sociaty for Industrial and Applied Mathematics, 1997.