

Measuring Regressor Information Content In The Presence of Collinearity

Joseph Retzer, Market Probe, Inc., Milwaukee WI

Kurt Pflughoeft, University of Texas, El Paso TX

Abstract

Design matrix collinearity is an all too common occurrence in least squares analyses. A measure of regressor variable information content suggested by Theil and Chung (which represents an extension of a concept originally introduced by Kruskal) is shown to, in part, mitigate this problem.

Kruskal's method measures regressor variable importance by averaging partial correlations over all model orderings. A limiting aspect of Kruskal's weights is that they are non-additive. That is to say the sum of the weights is not intrinsically meaningful. Theil and Chung suggest alternatively examining information theoretic measures which in fact are additive.

The information measures are estimated using the well know Hald Data set (which in turn contains strong collinearities). The authors take advantage of a number of uniquely powerful components of the SAS system.

We begin by describing Kruskal's measurement of importance as it represents the underlying methodology employed. This exposition also serves to aid in the understanding of results obtained at the end of the paper. Next, Theil and Chung's extension to additive measures of information is discussed. Finally, the information measures are estimated using a well-known real world data set.

This paper is targeted toward experienced SAS users familiar with PROC IML and PROC REG in SAS STAT. Additional background knowledge of statistical concepts, particularly with respect to partial correlations and regression analysis is also recommended.

1 Measuring Importance

Numerous methods for measuring importance in multiattribute value models have been presented in the literature (see, e.g., (Soofi and Retzer, 1992)). A review of those which pertain specifically to applications in statistical models may be found in (Soofi and Retzer, 1995). A common method is to rely on the p-values of t-scores attached to the partial regression coefficients. Kruskal notes that this is not unlike the old confusion of substantive with statistical significance. "In fact no necessary connection between importance (a property of the population) and statistical significance (a property of the sample and population) exists." ¹ This represents a confusion between statistical vs. real significance. While methods other than the use of p-values are employed, each in turn has its drawbacks. Boring also provides insight into potential problems when relying on p-values alone. He notes that "Science begins with description but it ends in generalization." (Boring, 1919) This highlights the fact that while statisticians strive for insight regarding the population, they do so while being able to observe only the sample. Insofar as the sample is an accurate representation of the population, our inferences will be justified. We must keep in mind however, that our statistics are descriptions of the sample and no more than that. For that reason we may strive to examine the data in more than one way in our attempt to gain insight into the population.

¹See Kruskal 1978, International Encyclopedia of Statistics.

Kruskal suggests an alternative method of measuring importance in which we average partial correlations over all model orderings. We implement his suggestions by taking advantage of a number of uniquely powerful components of the SAS system.

A limiting aspect of Kruskal's weights is that they are non-additive. That is to say the sum of the weights is not intrinsically meaningful. Theil and Chung suggest examining information theoretic importance measurements which are additive. Computationally, the measurement of information content can be viewed as an extension of Kruskal's analysis. This suggests that once the relative importance measures have been estimated, construction of the information measures is straightforward.

This paper proceeds by examining the computation of Kruskal's weights and the subsequent creation of information measures corresponding to OLS regressor variables. It considers the impact of collinearity in the design matrix by utilizing a well-known real

world data set in which the problem exists. This lends insight into the appropriateness of competing measures of importance under conditions of model instability.

2 Kruskal's Relative Importance Weights, An Illustration

The idea underlying relative importance weighting is to consider the strength of relationship between the regressor and the response variable under varying orderings of inclusion vis à vis the remaining variables. If we were to consider all possible orderings of the variables into the model, and the associated squared partial correlations of a particular variable with that of the response, an average of these partials squared would be considered its relative importance.

Consider the case of a four regressor model in which we are attempting to estimate the relative importance of variable x_1 . The orderings (permutations) and the associated relevant squared partial correlations are given in Table 1 below.

$\left. \begin{array}{l} x_1 \ x_2 \ x_3 \ x_4 \\ x_1 \ x_2 \ x_4 \ x_3 \\ x_1 \ x_3 \ x_2 \ x_4 \\ x_1 \ x_3 \ x_4 \ x_2 \\ x_1 \ x_4 \ x_2 \ x_3 \\ x_1 \ x_4 \ x_3 \ x_2 \end{array} \right\} 6 \cdot r_{y,x_1}^2$	$\left. \begin{array}{l} x_2 \ x_1 \ x_3 \ x_4 \\ x_2 \ x_1 \ x_4 \ x_3 \\ x_3 \ x_1 \ x_2 \ x_4 \\ x_3 \ x_1 \ x_4 \ x_2 \\ x_4 \ x_1 \ x_2 \ x_3 \\ x_4 \ x_1 \ x_3 \ x_2 \end{array} \right\} 2 \cdot (r_{y,x_1 \cdot x_2}^2 + r_{y,x_1 \cdot x_3}^2 + r_{y,x_1 \cdot x_4}^2)$
$\left. \begin{array}{l} x_2 \ x_3 \ x_1 \ x_4 \\ x_2 \ x_4 \ x_1 \ x_3 \\ x_3 \ x_2 \ x_1 \ x_4 \\ x_3 \ x_4 \ x_1 \ x_2 \\ x_4 \ x_2 \ x_1 \ x_3 \\ x_4 \ x_3 \ x_1 \ x_2 \end{array} \right\} 2 \cdot (r_{y,x_1 \cdot (x_2,x_3)}^2 + r_{y,x_1 \cdot (x_3,x_4)}^2 + r_{y,x_1 \cdot (x_2,x_4)}^2)$	$\left. \begin{array}{l} x_4 \ x_2 \ x_3 \ x_1 \\ x_3 \ x_2 \ x_4 \ x_1 \\ x_4 \ x_3 \ x_2 \ x_1 \\ x_2 \ x_3 \ x_4 \ x_1 \\ x_3 \ x_4 \ x_2 \ x_1 \\ x_2 \ x_4 \ x_3 \ x_1 \end{array} \right\} 6 \cdot r_{y,x_1 \cdot (x_2,x_3,x_4)}^2$

Table 1: All permutations in the four variable example

To understand why the relevant squared partial correlations are as given let us consider each block individually. In "Block 1" the relative importance of x_1 is, in all cases, equal to the square of its simple correlation with the response y (r_{y,x_1}^2). This is because x_1 is "entering" the model first, hence no other variables effect need be considered. In "Block 2" however this is no longer the case. For example if the ordering is $x_2 x_1 x_3 x_4$, the relative importance of x_1 must be considered after accounting for the presence of x_2 in the model. This is estimated by the square of the partial correlation of x_1 with the response y given x_2 ($r_{y,x_1 \cdot x_2}^2$). A similar pattern then emerges for all orderings in the remaining blocks. The relative importance of x_1 (RI_{x_1}) is then simply the average of all squared correlations (simple and partial) between x_1 and y . This may be written as follows,

$$RI_{x_1} = (6 \cdot r_{y,x_1}^2 + 2 \cdot (r_{y,x_1 \cdot x_2}^2 + r_{y,x_1 \cdot x_3}^2 + r_{y,x_1 \cdot x_4}^2 + r_{y,x_1 \cdot (x_2,x_3)}^2 + r_{y,x_1 \cdot (x_3,x_4)}^2 + r_{y,x_1 \cdot (x_2,x_4)}^2) + 6 \cdot r_{y,x_1 \cdot (x_2,x_3,x_4)}^2) / 24. \quad (1)$$

3 Partial Correlations Using Proc Reg

To begin, we may note the that simple correlation squared is equivalent to R^2 in simple regression analysis. In addition, the squared partial correlation may be calculated through appropriate use of SSE from relevant simple and / or multiple regressions. For example, the squared partial correlation of x_1 with y given x_2 and x_3 can be calculated as follows,²

$$r_{y,x_1 \cdot (x_2,x_3)}^2 = \frac{SSE_{x_2,x_3} - SSE_{x_1,x_2,x_3}}{SSE_{x_2,x_3}}. \quad (2)$$

Where

- SSE_{x_2,x_3} is the regression sum of squared errors resulting from the regression of y on x_2 and x_3 .
- SSE_{x_1,x_2,x_3} is the regression sum of squared errors resulting from the regression of y on x_1 , x_2 and x_3 .

A relatively simple way of arriving at all possible subset regression SSE's as well as their corresponding R^2 's is by the use of SAS' "proc reg" with the "selection = rsquare" and "sse" model options. It should be noted that if the number of variables in the model exceeds 10, SAS will default to printing only that same number of subset regression results for any fixed number of variables. For example, if there are 12 variables in total, there exist C_2^{12} combinations or 66 unique subset regressions with 2 regressors. SAS however will only report, by default, the first 12 models results. This problem may be overcome by including an additional model option, "best = nnnnnn", where nnnnnn is some number large enough to account for the largest number of subset regression models resulting from a fixed number of regressors. An illustration of the appropriate "proc reg" command could be written as:

```
proc reg data=regdat out=stats noprint;
    model y=x1-x4 / selection=rsquare
                best=99999 sse;
run;
```

The data necessary to calculate relative importance would then be found in the data set "stats."

²See Neter and Wasserman 1974.

Using the information in "stats" we may take advantage of equation (2) to calculate all necessary partial correlations. The indexing of all SSE's (i.e. associating each SSE with a particular regression model) is accomplished using the natural positioning indices associated with matrices via PROC IML. Specifically, the data in "stats" is read into an IML matrix while simultaneously being assigned to rows which indicate the independent variable set contained in a particular model. All that remains in order to calculate the relative importance measurements is to correctly compute the weighted averages of the relevant partial correlations.

4 Measuring Regressor Information Content

Theil suggests in his 1987 paper that an information theoretic measure which quantifies the amount of information contained in a particular set of regressor variables may be decomposed to allow for assignment of importance to each independent variable. Specifically, Theil suggests that $I(R^2)$ ³ be used to quantify the information in the regressors regarding the response variable. The $I(\cdot)$ function is a well know information based tool for quantifying information. The implied measure would be given as,

$$I(R^2) = \log_2(1 - R^2) \quad (3)$$

In addition we may note that $(1 - R^2)$ may be decomposed as,

$$1 - R^2 = (1 - r_{y,x1}^2)(1 - r_{y,x2 \cdot x1}^2) \cdots (1 - r_{y,xP \cdot (x1,x2,x3,\dots,x(p-1))}^2) \quad (4)$$

The information measure associated with the decomposition illustrated in (4) is arrived at

by taking the base 2 log of both sides to give,

$$I(R^2) = I(r_{y,x1}^2) + I(r_{y,x2 \cdot x1}^2) + \cdots + I(r_{y,xP \cdot (x1,x2,x3,\dots,x(p-1))}^2) \quad (5)$$

Theil notes that if a natural ordering were suggested to be say $x1, x2, \dots, xP$ then (5) would give a unique additive decomposition of the importance assigned to each variable in the model. If no natural ordering is agreed upon then we may follow Kruskal's method of averaging over all orderings.

Computationally, once Kruskal's weights are estimated, the extension to Theil's weights is straightforward. Note that the components on the RHS of (5) are the estimated partial correlations squared which then are transformed to yield the information measure. This transformation is then all that is required in addition to the previous routine.

5 The Impact of Collinearity

The absence of multicollinearity would lead to a situation in which the importance/information content of independent variables is more easily determined. However, the presence of multicollinearity is more of the rule than the exception. Such conditions may exasperate the decision maker's efforts to construct an appropriate model. Although the decision maker may try to circumvent the problem by running a series of full and reduced regression models, choosing amongst those models could prove to be problematic. Likewise, if all possible regression models from a set of independent variables are not considered, the decision maker may make

³ R^2 being the multiple correlation coefficient between the response and regressor variables.

unwarranted interpretations from the regression output.

To examine differences in interpretation between Theil's information measures and regression analysis, the Hald data set was analyzed.

5.1 Hald data set

The data set used is taken from an experiment illustrated in Hald's "Statistical Theory With Engineering Applications" (Hald, 1952). The variables used in the analysis are given as follows,

- y (Response) : Heat evolved in calories per gram of cement.
- x_1 : Amount of tricalcium aluminate, $3CaO \cdot Al_2O_3$.
- x_2 : Amount of tricalcium silicate, $3CaO \cdot SiO_2$.
- x_3 : Amount of tetracalcium aluminoferrite, $4CaO \cdot Al_2O_3 \cdot Fe_2O_3$.
- x_4 : Amount of dicalcium silicate, $2CaO \cdot SiO_2$.

The experiment was aimed at measuring the effect of composition on heat evolved during cement hardening.

Initial regression estimates reveal substantial collinearity in the data. This is illustrated in both the correlation matrix (Table 2) as well as regression diagnostic output (Table 3) generated by SAS' Proc Reg.

CORR	x1	x2	x3	x4	y
x1	1.00	0.23	-0.82	-0.25	0.73
x2	0.23	1.00	-0.14	-0.97	0.82
x3	-0.82	-0.14	1.00	0.03	-0.54
x4	-0.25	-0.97	0.03	1.00	-0.82
y	0.73	0.82	-0.54	-0.82	1.00

Table 2: Pairwise Correlations, Hald Data.

The R^2 for the initial OLS was .9824 and adjusted R^2 was .9736. In addition the condition index was 249.58. This would indicate strong collinearities in the data.

Below are the parameter estimates from that model.

Parameter		
Variable	Estimate	Prob > T
INTERCEP	62.405369	0.3991
x1	1.551103	0.0708
x2	0.510168	0.5009
x3	0.101909	0.8959
x4	-0.144061	0.8441

Table 3: Entire Independent Variable Set, Hald Data.

Variance inflation factors (Table 4 below) support the observed strong relationship between variables x_2 and x_4 that can be noted in the correlation matrix above.

Variable	Variance Inflation
INT	0.00000000
x1	38.49621149
x2	254.42316585
x3	46.86838633
x4	282.51286479

Table 4: Variance Inflation Factors, Hald Data.

A summary of a stepwise analysis on the model is given in Table 5.

Step	Var		Num In	Partial R^2	Model R^2
	Ent	Rmvd			
1	x4		1	0.675	0.675
2	x1		2	0.298	0.973
3	x2		3	0.010	0.982
4		x4	2	0.004	0.979

Table 5: Summary of Stepwise Procedure for Dependent Variable y , Hald Data.

Finally, the information measures in both actual as well as percentage terms is given in Table 6.

x1	x2	x3	x4
1.87083	1.52270	0.77905	1.65371
32.1102	26.1350	13.3712	28.3835

Table 6: Theil's Information Measures/Percentages, Hald Data.

In step 1, the x_4 variable was entered into the model; this variable is eventually removed in step 4. In step two, the x_1 variable was entered to improve the model's overall R^2 . In step 3, the x_2 variable is entered but it is obvious that x_2 and x_4 can not both be in the regression model due to their high collinearity. Now for forecasting purposes it appears that the model identified by the stepwise regression procedure is a good one. However, the decision maker may be tempted to state that x_3 and x_1 are not important in determining how much heat is generated during hardening. This may or may not be true and the only way to verify this is to have a deep understanding of the problem at hand. However, without such knowledge reliance on only the regression

output may lead the decision maker down the wrong path. It is clear from the correlation matrix that x_4 's importance as measured by r is about equivalent to x_2 's when explaining the variation in y . It may also be possible for the decision maker to rank the importance of each variable by examining the simple R^2 's between the dependent variable and a single independent variable. But these measures do not inform the decision maker as to what the relative or absolute distances between the measures may be, i.e. how much more important is x_2 than x_4 ? and what percent of y 's information is explained by x_2 in the context of the remaining independent variables?

Fortunately, the decision maker can gain a better grasp to the answers of these questions by examining Theil's measures. Table 6 lists Theil's relative importance weights in the first row and the second row transforms these measures into percentages. From the table we see that x_1 actually contains more information about y than any other variable. This insight is not evident when examining the correlation and/or regression output. The percent associated with x_1 indicates that 32% of the information in y is explained by x_1 . Next, we see that x_2 and x_4 measures are similar being 26.14 and 28.38 respectively. Finally, x_3 appears to be the least important but still contains some explanatory power with respect to y .

6 Conclusion

Looking at simple p -values for assessing relative importance may be misleading and/or incomplete. The p -values indicate only the significance of the regression coefficient with respect to the sample and do

not supply information concerning interrelationships among variables. A reasonable method for examining "relative" importance is offered by Kruskal. Calculation of Kruskal's measure can be tedious and computationally intense. This article presents an approach which employs certain aspects of the SAS programming language to derive an effective algorithm which handles the calculation of these weights in models with varying sized sets of regressors. In addition, the non-additivity limitation of Kruskal's weights can be overcome by a straightforward transformation suggested by Theil and Chung. This new measure, which has a solid foundation in information theory, provides additional insight into the data and in part mitigates the effect of design matrix collinearity.

References

- Belsley, D., Kuh, E., and Welsch, R. (1980). *Regression Diagnostics, Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons Inc.
- Boring, E. G. (1919). Mathematical vs. scientific significance. *Psychological Bulletin*, 16(10):335-338.
- Hald, A. (1952). *Statistical Theory with Engineering Applications*. John Wiley & Sons Inc.
- Kruskal, W. and Majors, R. (1989). Concepts of relative importance in scientific literature. *The American Statistician*, 43(1).
- Kruskal, W. H., editor (1978). *International Encyclopedia of Statistics*, chapter Tests of Significance, pages 944-958. New York Free Press.

Neter, J. and Wasserman, W. (1974). *Applied Linear Statistical Models*. Richard D. Irwin Inc.

SAS Inc. (a) (1988). *SAS IML User's Guide Release 6.03*. SAS Institute Inc.

SAS Inc. (b) (1989). *SAS Language and Procedures Version 6*. SAS Institute Inc., first edition.

SAS Inc. (c) (1989). *SAS Stat User's Guide Version 6, Volumes 1 and 2*. SAS Institute Inc., fourth edition.

Soofi, E. and Retzer, J. (1992). Adjustment of importance weights in multiattribute value models by minimum discrimination information. *European Journal of Operational Research*, 60.

Soofi, E. and Retzer, J. (1995). A review of relative importance measures in statistics. *The Proceedings of the American Statistical Association Bayesian Statistical Sciences Section*.

Theil, H. (1987). How many bits of information does an independent variable yield in a multiple regression? *Statistics and Probability Letters*, 6(2).

Authors

Joseph Retzer, Ph.D.
Market Probe, Inc.
Milwaukee WI, 53226
414-778-6000 (voice)
jjr@execpc.com

Kurt Pflughoeft, Ph.D.
University of Texas
El Paso TX, 79968-0544
915-747-7733 (voice)
pflug@mail.utep.edu

*SAS, SAS STAT, Proc Reg and SAS IML are registered trademarks of SAS Institute Inc., Cary NC, USA.