Measuring The Information Content Of Regressors In The Linear Model Using Proc Reg* and SAS IML*

"It is important that there be options in explication, and equally important that the candidates have clear population interpretations. ... paintings, diamonds, and data should be examined in several ways." (Kruskal, 1987)

Joseph Retzer, Market Probe, Inc., Milwaukee WI

Kurt Pflughoeft, University of Texas, El Paso TX

Abstract

This paper begins by describing an implementation of Kruskal's relative importance analysis using SAS STAT and SAS IML. While Kruskal's weights lend insight into the relative importance of each regressor, they are non-additive in nature and therefore limit potential interpretation. In order to overcome the non-additivity drawback, an information theoretic measure (as suggested by Theil and Chung in "Information-theoretic measures of fit for univariate and multivariate linear regressions", *The American Statistician* 1988) is implemented. In addition, the impact of regressor variable collinearity is examined using simulated data.

This paper is targeted toward experienced SAS users familiar with 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

[1] See Kruskal 1978, International Encyclopedia of Statistics.

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.

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 in fact 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 linear model regressor variables. It will also consider the impact of collinearity in the design matrix by utilizing simulated data. 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 x1. The orderings (permutations) and the associated relevant squared partial correlations are given in Table 1 below.

$$\left.\begin{array}{cccc} x1 & x2 & x3 & x4 \\ x1 & x2 & x4 & x3 \\ x1 & x3 & x2 & x4 \\ x1 & x3 & x4 & x2 \\ x1 & x4 & x2 & x3 \\ x1 & x4 & x3 & x2 \end{array}\right\} 6 \cdot r^2_{y,x1}$$

$$\left.\begin{array}{cccc} x2 & x1 & x3 & x4 \\ x2 & x1 & x4 & x3 \\ x3 & x1 & x2 & x4 \\ x3 & x1 & x4 & x2 \\ x4 & x1 & x2 & x3 \\ x4 & x1 & x3 & x2 \end{array}\right\} 2 \cdot (r^2_{y,x1\cdot x2} + r^2_{y,x1\cdot x3} + r^2_{y,x1\cdot x4})$$

$$\left.\begin{array}{cccc} x2 & x3 & x1 & x4 \\ x2 & x4 & x1 & x3 \\ x3 & x2 & x1 & x4 \\ x3 & x4 & x1 & x2 \\ x4 & x2 & x1 & x3 \\ x4 & x3 & x1 & x2 \end{array}\right\} 2 \cdot (r^2_{y,x1\cdot(x2,x3)} + r^2_{y,x1\cdot(x3,x4)} + r^2_{y,x1\cdot(x2,x4)})$$

$$\left.\begin{array}{cccc} x4 & x2 & x3 & x1 \\ x3 & x2 & x4 & x1 \\ x4 & x3 & x2 & x1 \\ x2 & x3 & x4 & x1 \\ x3 & x4 & x2 & x1 \\ x2 & x4 & x3 & x1 \end{array}\right\} 6 \cdot r^2_{y,x1\cdot(x2,x3,x4)}$$

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 x1 is, in all cases, equal to the square of its simple correlation with the response y ($r^2_{y,x1}$). This is because x1 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 x2 x1 x3 x4, the relative importance of x1 must be considered after accounting for the presence of x2 in the model. This is estimated by the square of the partial correlation of x1 with the response y given x2 ($r^2_{y,x1\cdot x2}$). A similar pattern then emerges for all orderings in the remaining blocks. The relative importance of x1 (RI_{x1}) is then simply the average of all squared correlations (simple and partial) between x1 and y. This may be written as follows,

$$\begin{aligned} RI_{x1} &= (6 \cdot r^2_{y,x1} + 2 \cdot (r^2_{y,x1\cdot x2} + r^2_{y,x1\cdot x3} \\ &\quad + r^2_{y,x1\cdot x4} + r^2_{y,x1\cdot (x2,x3)} \\ &\quad + r^2_{y,x1\cdot (x3,x4)} + r^2_{y,x1\cdot (x2,x4)}) \\ &\quad + 6 \cdot r^2_{y,x1\cdot (x2,x3,x4)})/24. \end{aligned} \tag{1}$$

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 x1 with y given x2 and x3 can be calculated as follows, [2]

$$r^2_{y,x1\cdot (x2,x3)} = \frac{SSE_{x2,x3} - SSE_{x1,x2,x3}}{SSE_{x2,x3}}. \tag{2}$$

[2] See Neter and Wasserman 1974.

Where

- $SSE_{x2,x3}$ is the regression sum of squared errors resulting from the regression of y on x2 and x3.
- $SSE_{x1,x2,x3}$ is the regression sum of squared errors resulting from the regression of y on x1, x2 and x3.

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^{12}_2 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`."

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,

$$I(R^2) = \log_2(1 - R^2) \tag{3}$$

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

$$\begin{aligned} 1 - R^2 &= (1 - r^2_{y,x1})(1 - r^2_{y,x2\cdot x1}) \cdots \\ &\quad (1 - r^2_{y,xP\cdot(x1,x2,x3,..x(p-1))}) \end{aligned} \tag{4}$$

The information measure associated with the decomposition illustrated in (4) is arrived at by taking the base 2 log of both sides to give,

$$\begin{aligned} I(R^2) &= I(r^2_{y,x1}) + I(r^2_{y,x2\cdot x1}) + \cdots \\ &\quad + I(r^2_{y,xP\cdot(x1,x2,x3,..x(p-1))}) \end{aligned} \tag{5}$$

Theil notes that if a natural ordering were suggested to be say $x1, x2, \ldots xP$ 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 unwarranted interpretations from the regression output.

To examine differences in interpretation between Theil's information measures and regression analysis, 1000 observations of 5 variables were randomly generated. The independent variables, $x1 - x3$ were drawn from normal distributions with $\mu = 10$ and $\sigma = 2$. Multicollinearity was induced into the sample by setting the correlation, ρ, between $x3$ and $x4$ to 0.70. This correlation was achieved using the following transformation,
$x4 = (\rho \cdot x3) + (z \cdot \sqrt{1 - \rho^2})$, where $z \sim N(0, 1)$. The correlation matrix for the dependent and independent variables is given in Table 2.

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

CORR	x1	x2	x3	x4	y
x1	1.00	0.03	0.02	0.03	0.67
x2	0.03	1.00	0.02	0.07	0.04
x3	0.02	0.02	1.00	0.72	0.68
x4	0.03	0.07	0.72	1.00	0.51
y	0.67	0.04	0.68	0.51	1.00

Table 2: Pairwise Correlations

The parameter estimates as well as p-values resulting from a regression employing all independent variables is given in Table 3.

Variable	Parameter Estimate	Prob > | T |
INTERCEP	-0.307257	0.2491
x1	0.999880	0.0001
x2	0.020599	0.1789
x3	1.008582	0.0001
x4	0.002871	0.8979

Table 3: Regression I, Entire Independent Variable Set

The Condition Index for this model was approximately 21.9. A Condition Index in the neighborhood of 15-30 seems to represent a borderline situation in which collinearity may become a problem (Belsley et al., 1980). A ranking of the variable's importance based upon p values, would be $x1$, $x3$, $x2$, and $x4$ respectively. Note that variables $x2$ and $x4$ are not statistically significant. It appears that the multicollinearity between $x3$ and $x4$ is effecting the contribution credited to the latter variable. Indeed, it appears $x4$ is less valuable than $x2$ in explaining y. Similar results for a reduced model in which $x2$ and $x3$ are eliminated are given in Table 4. [4]

Variable	Parameter Estimate	Prob > |T|
INTERCEP	2.740916	0.0001
x1	0.993133	0.0001
x4	0.741698	0.0001

Table 4: Regression II, Reduced Independent Variable Set

In this model the true contribution of $x4$ is more readily apparent.

The information levels contained in the regressor set, as measured by Theil's technique, are given in Table 5.

x1	x2	x3	x4
1.67258	.0021914	1.33968	0.27348

Table 5: Theil's Information Measures. (Information in Regressors on Dependent Variable y)

Using Theil's information measures the ranking of the variables' information content is $x1$, $x3$, $x4$ and $x2$ respectively. The percent of total information for these variables are $x1$ (50.83%), $x3$ (40.72%), $x4$ (8.31%) and $x2$ (0%). Note that Theil's measures correctly assign higher information content to $x4$ as opposed to $x2$.

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 various 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.

[4] The Condition Index for this model was approximately 14.8

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.

Kruskal, W. (1987). Relative importance by averaging over orderings. *The American Statistician*, 41(1).

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