Research Report
Department of Statistics
Göteborg University
Sweden

# A stepwise regression procedure applied to an international production study - a multiple inference solution

**Tommy Johnsson**
**Inge Ivarsson**

**Research Report 1997:4**
**ISSN 0349-8034**

| Mailing address: | Fax | Phone | Home Page: |
|---|---|---|---|
| Department of Statistics | Nat: 031-773 12 74 | Nat: 031-773 10 00 | http://www.stat.gu.se |
| Göteborg University | Int: +46 31 773 12 74 | Int: +46 31 773 10 00 | |
| Box 660 | | | |
| SE 405 30 Göteborg | | | |
| Sweden | | | |

# A stepwise regression procedure applied to an international production study - a multiple inference solution

**Tommy Johnsson**
**Inge Ivarsson**[*)]

[*)] **Department of Geography, University of Göteborg**

# Contents

# 1 Introduction

The aim of this paper is to apply a certain procedure for stepwise regression analysis to a problem included in a study of integrated international production. The question, which factors are important for the transnational company's trade behaviour in certain respects, have been analysed by means of traditional methods. Hence the more interesting findings on the subject are likely to be reported already and the results here are only to be regarded as supplementary ones. Comparisons between old and new outcomes, and to some extent the interpretations of them, is however one point here.

A brief description of the underlying, real world, problem is given in section 2 and the basic ideas behind the regression procedure is presented in section 3. The major results are given in section 4 while some concluding remarks are made in section 5.

# 2 The problem

The problems analysed here is part of a major study on foreign transnational corporations in Sweden, Ivarsson (1996), including an empirical survey of the company's behaviour in certain respects. After noting that some well-defined measurements tend to vary between companies, factors likely to explain this variation are looked for. That is, models were built in order to verify the connections in terms of independent variables explaining the dependent ones. Among the latter were percentage of manufactured output being exported i) outside Sweden, ii) outside the Nordic countries and iii) inside the firm. Also analysed were the proportion of intra-firm import . In the original study the degree of intra-corporate technology co-operation were estimated on an ordinal scale and also transformed to dichotomous variable but that one has, due to the questionable distribution properties, been excluded from the dependent variables here. The former, explanatory, factors included categorical ones such as whether the company was a raw material based industry and the extent of technological co-operation with external firms in Sweden as well as real valued variables such as size and age of the affiliate. For a complete list of factors see appendix 1.

Descriptive statistics are given in tables 1 and 2. The former gives means etc for observations used in the analyses of $Y_1$, $Y_3$ and $Y_4$ while the latter, since one values is missing, the remaining data for the analysis of $Y_2$ . It should be noted that a factor with a large variance, when competing with other factors and everything else equal, are more likely to be regarded significant than one with small variance.

| Table 1 | Var | Mean | Std | Min | Max | n=288 |
|---|---|---|---|---|---|---|
| | $Y_1$ | 42.7 | 33.3 | 0 | 100 | |
| | $Y_3$ | 34.3 | 36.5 | 0 | 100 | |
| | $Y_4$ | 23.2 | 32.9 | 0 | 100 | |
| | $X_1$ | 0.13 | 0.33 | 0 | 1 | |
| | $X_2$ | 48.0 | 32.2 | 0 | 100 | |
| | $X_3$ | 0.51 | 0.50 | 0 | 1 | |
| | $X_4$ | 0.56 | 0.50 | 0 | 1 | |
| | $X_5$ | 45.3 | 40.6 | 1 | 100 | |
| | $X_6$ | 0.87 | 0.34 | 0 | 1 | |
| | $X_7$ | 11.1 | 14.0 | 0 | 89 | |
| | $X_8$ | 100 | 121 | 3 | 917 | |
| | $X_9$ | 0.12 | 0.33 | 0 | 1 | |

| Table 2 | Var | Mean | Std | Min | Max | n=287 |
|---|---|---|---|---|---|---|
| | $Y_2$ | 29.9 | 32.7 | 0 | 100 | |
| | $X_1$ | 0.13 | 0.33 | 0 | 1 | |
| | $X_2$ | 48.2 | 32.2 | 0 | 100 | |
| | $X_3$ | 0.52 | 0.50 | 0 | 1 | |
| | $X_4$ | 0.56 | 0.50 | 0 | 1 | |
| | $X_5$ | 45.5 | 40.6 | 1 | 100 | |
| | $X_6$ | 0.87 | 0.33 | 0 | 1 | |
| | $X_7$ | 10.9 | 13.7 | 0 | 89 | |
| | $X_8$ | 100 | 122 | 3 | 917 | |
| | $X_9$ | 0.12 | 0.33 | 0 | 1 | |

# 3 The procedure

When non-experimental data is used for showing linear correlations or other associations between investigated variables there is always a possibility of confounding factors. Since several factors, observed as well as not observed ones, tend to vary at the same time it is difficult to tell which of them do have the original influence on the response. Variables outside the study in question are of course impossible to take into account; concerning them we just have to make a reservation in the interpretations of the results. However, for factors measured and included in the study there are means of avoiding more serious confounding. The first and general rule is of course to not exclude possible confounders from the models, even if they sometimes are of minor or no interest. But even so, there are still problems when interpreting for instance a regression model, with multicollinearity between the regressors.

A second issue when multiple analyses are used is that of protection against false statements. When several regressors are included, and hence several test are performed, the probability of committing a type-I error will increase with the number of tests and eventually reach unacceptable levels. That is, if no further steps are taken to keep the multiple level of significance at certain pre-determined values

The procedure for stepwise regression analysis used here can be regarded as a technique for dealing with the problems mentioned above. Multicollinearity is taken into account for by testing whether a factor showing some influence on the response, really maintain that impact after other correlations with other factors has been adjusted for. After analysing all regressors the final result is the forming of a number of groups containing either i) only one regressor proven to have it's own impact on the response, ii) one regressor with it's own impact as well as one or several factors correlated with the response but overruled by and associated with the first member of the group or iii) the remaining factors not proven to have any correlation what so ever with the response. Each test is also performed under the protection of a predetermined multiple significance level. In this case this means that the probability of forming a group such as i) or ii) containing only nuisance variables is kept at a low level. The procedure is described in Johnsson (1992) and also included, along with some evaluations, in Johnsson (1989).

# 4 Results

The regressors found to be significant in the original analysis is given in Table 3. With various p-values two or three factors, different among the responses, were shown to have any influence. Using the stepwise procedure described above gave only a slightly different outcome. The latter results, in terms of the groups formed, are summarised in Table 4. For these analyses the multiple level of significance was set to 0.10 which, if there are nine possible regressors tested simultaneously, corresponds to a level at about 0.01 for the single tests.

Since the traditional analysis only picks out the regressors one by one it is natural to compare the significant ones there with the first ranked variables in each group formed by the stepwise procedure. Doing so, the results are identical for the responses $Y_1$ ,$Y_2$ and $Y_4$while there is a minor difference for $Y_3$; the original analysis gave three significant factors, taken one by one, where the stepwise procedure regarded two of those to be highly correlated and the third one to be insignificant.

Looking at the second or third ranked regressors within the groups provide information not given by the traditional models. For the responses $Y_1$ ,$Y_2$ and $Y_4$ one or two variables are pointed out, not as significant in there own rights, but as possible confounders and hence interesting when the structure of dependencies are being described.

When models from the multiple inference stepwise procedure are to be used for prediction purpose it is reasonable to form regression equations based on the first ranked factors only. This also holds when regressioncoefficents are compared with the ones from the original analysis. From a theoretical point of view if could however be interesting to see the values, and especially, the signs of the second and third ranked factors if they were included in the models. Table 5 gives the former while table 6 gives the latter. Compared to the original analyses no major differences could be noted.

**Table 3 Significant regressors, original analysis**

| Response variable | Regressors | | |
|---|---|---|---|
| $Y_1$ | $X_5^{***}$ | $X_8^{**}$ | $X_9^{***}$ |
| $Y_2$ | $X_5^{***}$ | $X_8^{***}$ | $X_9^{***}$ |
| $Y_3$ | $X_1^{*}$ | $X_3^{*}$ | $X_4^{***}$ |
| $Y_4$ | | $X_2^{***}$ | $X_5^{***}$ |

***p-value<0.001 **p-value<0.01 *p-value<0.05

## Table 4 Groups formed by stepwise procedure

| Response variable | Groups of regressors | | |
|---|---|---|---|
| $Y_1$ | $(X_5\ X_6\ X_7)$ | $(X_8)$ | $(X_9)$ |
| $Y_2$ | $(X_5)$ | $(X_8\ X_6)$ | $(X_9)$ |
| $Y_3$ | | | $(X_4\ X_1)$ |
| $Y_4$ | | $(X_2\ X_6)$ | $(X_5)$ |

multiple level of significance<0.10

## Table 5 Regressioncoefficients, only first ranked ones

| Factor/Dep var | Y1 | Y2 | Y3 | Y4 |
|---|---|---|---|---|
| $X_1$ | | | | |
| $X_2$ | | | | -0.24 |
| $X_3$ | | | | |
| $X_4$ | | | -16.2 | |
| $X_5$ | 0.25 | 0.29 | | -0.17 |
| $X_6$ | | | | |
| $X_7$ | | | | |
| $X_8$ | 0.04 | 0.05 | | |
| $X_9$ | 14.1 | 13.8 | | |

**Table 6 Regressioncoefficients, all, also confounders**

| Factor/Dep var | Y1 | Y2 | Y3 | Y4 |
|---|---|---|---|---|
| $X_1$ | | | -13.6 | |
| $X_2$ | | | | -0.23 |
| $X_3$ | | | | |
| $X_4$ | | | -13.4 | |
| $X_5$ | 0.24 | 0.28 | | -0.16 |
| $X_6$ | 4.20 | 11.2 | | -8.16 |
| $X_7$ | -0.24 | | | |
| $X_8$ | 0.04 | 0.04 | | |
| $X_9$ | 13.4 | 13.5 | | |

# 5 Conclusions

The fact that there are no major differences between the results given by the two methods could be regarded as a confirmation of the results already presented in Ivarsson (1996) while the supplementary information regarding the multicollinearity among the regressors just adds some details to the picture. However, since the study is based on a rather large number of observations making it possible to detect even week correlations, this is hardly surprising. On the contrary, different results would have been quite confusing. In this context one must also remember that, regardless of the chosen technique, significant factors does not necessarily mean important factors and that all conclusions based on the statistical inference performed are valid only within the models. This holds for the traditional analysis as well as for the stepwise procedure, the amount of information contained in the sample could not be artificially increased; it could only be utilised in a more efficient way. The latter could perhaps be achieved by using the described procedure.

## References

Ivarsson, I. (1996): Integrated International Production- A Study of Foreign Transnational Corporations in Sweden, Dept of Geography, University of Göteborg

Johnsson, T. (1989): On stepwise procedures for some multiple inference problems, Almqvist & Wiksell International, Stockholm, Sweden

Johnsson, T. (1992): A procedure for stepwise regression analysis, Statistical Papers 33, Springer-Verlag

# Appendix

## Dependent variables

$Y_1$ = **Affiliates' export propensity** - percentage of manufactured output sold outside Sweden

$Y_2$ = **Affiliates' extra-Nordic export propensity** - percentage of manufactured output sold outside the Nordic countries

$Y_3$ = **Intra-firm export** - percentage of total export

$Y_4$ = **Intra-firm import** - percentage of total import

## Independent variables

$X_1$ = **Raw material based industry** - dummy variable with the value 1 for wood products, furniture, paper and pulp, iron and steel and non-ferrous metals; 0 for others

$X_2$ = **Domestic purchase of inputs** - percentage of total purchases of material input

$X_3$ = **Degree of domestic inter-firm technology co-operation** - an ordinal scale 0-4 transformed into a dummy variable with 0 versus 1-4

$X_4$ = **Affiliates operating in competitive Swedish industry-clusters** - dummy variable with the value 1 for affiliates related to such cluster; 0 for others

$X_5$ = **Specialised affiliates** - sales-value of affiliate's major product as share of parent corporation's total sale of the same product

$X_6$ = **Mode of entry** - dummy variable with 0 for green-field investments and 1 for acquisitions

$X_7$ = **Age of affiliate** - number of years since incorporated or established by parent corporation

$X_8$ = **Size of affiliate** - in percentage of average size within type of corporation

$X_9$ = **Home country of affiliate** - dummy variable with 1 for European and 0 for others

Research Report

| | | |
|---|---|---|
| 1995:1 | Arnkelsdottir, H | Surveillance of rare events. On evaluations of the sets method. |
| 1995:2 | Sveréus, A | Detection of gradual changes. Statistical methods in post marketing surveillance. |
| 1995:3 | Ekman, C | On second order surfaces estimation and rotatability. |
| 1996:1 | Ekman, A | Sequential analysis of simple hypotheses when using play-the-winner allocation. |
| 1996:2 | Wessman, P | Some principles for surveillance adopted for multivariate processes with a common change point. |
| 1996:3 | Frisén, M. &<br>Wessman, P | Evaluations of likelihood ratio methods for surveillance. |
| 1996:4 | Wessman, P. | Evaluation of univariate surveillance procedures for some multivariate problems. |
| 1996:5 | Särkkä, A. | Outlying observations and their influence on maximum pseudo-likelihood estimates of Gibbs point processes. |
| 1997:1 | Ekman, A. | Sequential probability ratio tests when using randomized play-the-winner allocation. |
| 1997:2 | Pettersson, M. | Monitoring a Freshwater Fishpopulation: Statistical Surveillance of Biodiversity. |
| 1977:3 | Jonsson, R. | Screening-related prevalence and incidence for non-recurrent diseases. |