ESTIMATED PARAMETERS DO NOT GET THE WRONG SIGN DUE TO COLLINEARITY ACROSS INCLUDED VARIABLES

PHILIP HANS FRANSES AND CHRISTIAAN HEIJ

ERIM REPORT SERIES *RESEARCH IN MANAGEMENT*	
ERIM Report Series reference number	ERS-2002-31-MKT
Publication	March 2002
Number of pages	5
Email address corresponding author	franses@few.eur.nl
URL (electronic version)	http://www.eur.nl/WebDOC/doc/erim/erimrs20020306145935.pdf
Address	Erasmus Research Institute of Management (ERIM) Rotterdam School of Management / Faculteit Bedrijfskunde Erasmus Universiteit Rotterdam P.O. Box 1738 3000 DR Rotterdam, The Netherlands Phone: +31 10 408 1182 Fax: +31 10 408 9640 Email: info@erim.eur.nl Internet: www.erim.eur.nl

Bibliographic data and classifications of all the ERIM reports are also available on the ERIM website: www.erim.eur.nl

ERASMUS **R**ESEARCH **I**NSTITUTE OF **M**ANAGEMENT

REPORT SERIES

RESEARCH IN MANAGEMENT

BIBLIOGRAPHIC DATA AND CLASSIFICATIONS		
Abstract	Estimation results in linear regression models are sometimes in contrast with what was expected on the basis of a certain set of hypotheses or theory, in the sense that one or more parameters have the "wrong sign". One could be inclined to think that this is due to collinearity across explanatory variables, suggesting one should leave out one or more of the collinear variables. In this note we show that this is not a valid approach. Additionally, we show that "wrong signs" can occur because of correlations between included and omitted variables, so that "wrong signs" may occur if the model is not correctly specified. That is, if we find 'wrong signs" we should start questioning our model choice, not the data.	
Library of Congress Classification (LCC)	5001-6182	Business
	5410-5417.5	Marketing
	HA 31.3	Regression Analysis
Journal of Economic Literature (JEL)	M	Business Administration and Business Economics
	M 31 C 44	Marketing Statistical Decision Theory
	M 52	Model evaluation and testing
European Business Schools Library Group (EBSLG)	85 A	Business General
	280 G 255 A	Managing the marketing function Decision theory (general)
	150 K	Applied Econometrics
Gemeenschappelijke Onderwerpsontsluiting (GOO)		
Classification GOO	85.00	Bedrijfskunde, Organisatiekunde: algemeen
	85.40 85.03	Marketing Methoden en technieken, operations research
	85.03	Methoden en technieken, operations research
Keywords GOO	Bedrijfskunde / Bedrijfseconomie	
	Marketing / Besliskunde	
	Regressiemodellen, Collineairiteit,	
Free keywords	Misspecification, collinearity, parameter estimation	

Estimated parameters do not get the "wrong sign" due to collinearity across included variables*

Philip Hans Franses and Christiaan Heij †

March 5, 2002

Abstract

Estimation results in linear regression models are sometimes in contrast with what was expected on the basis of a certain set of hypotheses or theory, in the sense that one or more parameters have the "wrong sign". One could be inclined to think that this is due to collinearity across explanatory variables, suggesting one should leave out one or more of the collinear variables. In this note we show that this is not a valid approach. Additionally, we show that "wrong signs" can occur because of correlations between included and omitted variables, so that "wrong signs" may occur if the model is not correctly specified. That is, if we find "wrong signs" we should start questioning our model choice, not the data.

*This note is motivated by our personal experience that many applied researchers (and referees of academic journals) are inclined to blame the data for estimated effects that contradict a proposed theory (that is the "wrong sign" in the title), by suggesting that this might be due to collinearity across the included explanatory variables. In this note we show that this statement is not correct, at least not in large enough samples. The data used in the illustration originate from the SPSS standard datasets, and can also be obtained from the authors upon request.

†Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR, Rotterdam, The Netherlands, email: franses@few.eur.nl and email: heij@few.eur.nl

1 Introduction

It sometimes happens in empirical analysis that some parameters in a linear regression model get unexpected signs, where the expected sign is based on a theory or a set of hypotheses. A relevant question concerns the cause of this "wrong sign". Of course, the theory or hypotheses could be wrong, but it may also be that something is wrong with the empirical analysis. A popular explanation of wrong signs is to blame the data for this in the sense that the explanatory variables show unhealthy correlations, that is, there is multicollinearity. In this note we prove that in linear regression models such collinearity cannot be the cause of wrong signs. It should be stressed though that our arguments are based on asymptotic theory, that is, our statements hold true for large enough samples. Instead, it may well be that omitted variables cause unexpected signs. That is, if we are surprised by model outcomes then we should not start with blaming the data but with reconsidering our own choice of the regression equation.

The outline of this note is as follows. In Section 2, we discuss three cases that can happen in practice (the model is correct, too large, or too small), and in Section 3, we conclude with mentioning a few other genuine potential causes of wrong signs.

2 Signs and significance of regression coefficients

In this section we consider three possible situations in empirical modelling. The first case concerns a model that perfectly matches with the data generating process [DGP], the other two cases concern a too large or a too small model. In all cases we assume that the DGP is

$$y_i = \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i, \qquad i = 1, \cdots, n,$$

which in words means that the variable y can be explained by two variables x_1 and x_2. Extensions to more than two variables are straightforward, and the conclusions in this note do not change. Furthermore, for simplicity we assume that the random terms ε_i are independent and identically distributed as standard normal, that is with mean 0 and variance 1, that ε_i is uncorrelated with $x_{1,i}$ and with $x_{2,i}$, and finally

that $x_{1,i}$ and $x_{2,i}$ are jointly normally distributed, both with mean 0 and variance 1, and covariance $\text{cov}(x_{1,i}, x_{2,i}) = \rho$.

2.1 Model is correctly specified

In case the empirical model perfectly matches with the DGP, so that the specified model is

$$y_i = \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i,$$

then any textbook will prove that the ordinary least squares estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ are unbiased and consistent. In other words, in large enough samples, the sign of the estimates will be equal to that of the true underlying parameters β_1 and β_2.

In judging the estimated model one usually considers the t-values corresponding to $\hat{\beta}_1$ and $\hat{\beta}_2$, which we denote by t_1 and t_2. The role of multicollinearity is explained by the formula

$$t_1 \approx \hat{\beta}_1 \sqrt{n} \sqrt{1 - \rho^2}.$$

So $\hat{\beta}_1$ will be significant even in case of high collinearity ($|\rho| \approx 1$), provided that the sample size n is large enough.

2.2 Model contains a redundant variable

Suppose now that the DGP is given by

$$y_i = \beta_1 x_{1,i} + \varepsilon_i,$$

so that $x_{2,i}$ actually does not affect y, but one estimates the model

$$y_i = \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i.$$

In this case the least squares estimators are still unbiased and consistent and the expression for the t-value in Section 2.1 remains valid. This means that also in this situation, whatever correlation exists between the explanatory variables, the inclusion of a redundant (unnecessary) variable does not lead to "wrong signs", at least not in large enough samples. The cost of including collinear redundant variables lies in a lower significance, as it reduces the t-value with the factor $\sqrt{1 - \rho^2}$.

2.3 Model lacks a relevant variable

The final case, which is also the most interesting, is the case where the DGP is

$$y_i = \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i,$$

but the practitioner estimates the model

$$y_i = \gamma x_{1,i} + \eta_i,$$

so that the variable x_2 is erroneously omitted. In this case the least squares estimator of the effect of x_1 on y is no longer unbiased, it is not consistent, and in large enough samples it is given by

$$\hat{\gamma} \approx \beta_1 + \beta_2 \rho.$$

So, whereas the true partial effect of x_1 on y (that is, keeping x_2 fixed) is β_1, by estimating the model where the relevant variable x_2 is omitted, one estimates a (total) effect of $\beta_1 + \beta_2 \rho$. Clearly, depending on the values of β_2 and ρ, one may find an effect of x_1 in the model that is opposite to the true effect in the DGP. Because $|\rho|$ is smaller than 1, the wrong sign can only occur if $|\beta_2|$ is larger than $|\beta_1|$. In other words, the variable that is not included should be more important than the one that is included.

Note that omission of a relevant variable from the model is not harmful if the omitted variable is uncorrelated with the included variable (so that $\rho = 0$). However, if one thinks that multicollinearity is a problem (ρ large) and therefore omits a relevant variable ($\beta_2 \neq 0$), then the above result shows that this action may precisely be the cause of wrong signs!

As an illustration, consider the following model. The variable to be explained is salary (in dollars per year, and included in the model after natural logarithmic transformation), and the explanatory variables are age (in years) and education (also in years). There are 473 observations. A regression of salary on an intercept, age, age-squared and education gives the estimates (with standard errors in parentheses) 8.193 (0.255), 0.044 (0.012), -0.0005(0.0001) and 0.088 (0.005), and an R-squared value of 0.502. The correlation between age and age-squared is 0.995 and the correlation between age and education is -0.282. When we consider a regression of salary

on an intercept and age, we obtain 10.675 (0.069) and -0.007 (0.002), which implies that age now has an unexpected (significant) negative effect. Hence, omitting the relevant variables age-squared and education, gives unwanted and inappropriate empirical results.

3 Conclusion

This note shows that "wrong signs" can be caused by omitted variables that are more important in explaining the dependent variable y than the included variables. Further, collinearity across included variables cannot cause "wrong sings", at least not in large enough samples.

For the sake of simplicity the results were presented for the case of two independent variables, but they hold equally well true for the case of $k > 2$ independent variables $x_1, \cdots, x_k$. The expression for the t-value of $\hat{\beta}_1$ becomes

$$t_1 \approx \beta_1 \sqrt{n} \sqrt{1 - R^2}$$

where R^2 is the R-squared of the regression of x_2 on the other independent variables $x_2, \cdots, x_k$. As concerns omitted variables, if the variables $x_2, \cdots, x_k$ are erroneously omitted from the model and all these variables have mean 0 and variance 1, then the regression of y on x_1 alone provides an estimated effect of

$$\hat{\gamma} \approx \beta_1 + \beta_2 \rho_2 + \cdots + \beta_k \rho_k$$

where ρ_j is the correlation between x_1 and x_j for $j = 2, \cdots, k$. The wrong sign occurs (in large enough samples) if $\beta_2 \rho_2 + \cdots + \beta_k \rho_k$ is larger than and has the opposite sign of β_1.

There are of course many other possible causes of unexpected and unwanted signs. Some of the variables may be endogenous, the data may contain measurement errors, the sample may not be representative, and so on. Concerning the statistical significance of the estimates, multicollinearity is but one of the possible problems. Other causes of low significance may be that the sample is too small or that the error terms are not distributed as assumed (non-normal, heteroskedastic, and so on). In large samples, significance is not really an issue, but wrong model choice (neglecting

relevant variables) and endogeneity are the main concerns for proper interpretation of the estimated coefficients.

Publications in the Report Series Research* in Management

ERIM Research Program: "Marketing"

2002

Suboptimality of Sales Promotions and Improvement through Channel Coordination
Berend Wierenga & Han Soethoudt
ERS-2002-10-MKT

The Role of Schema Salience in Ad Processing and Evaluation
Joost Loef, Gerrit Antonides & W. Fred van Raaij
ERS-2002-15-MKT

The Shape of Utility Functions and Organizational Behavior
Joost M.E. Pennings & Ale Smidts
ERS-2002-18-MKT

Competitive Reactions and the Cross-Sales Effects of Advertising and Promotion
Jan-Benedict E.M. Steenkamp, Vincent R. Nijs, Dominique M. Hanssens & Marnik G. Dekimpe
ERS-2002-20-MKT

Do promotions benefit manufacturers, retailers or both?
Shuba Srinivasan, Koen Pauwels, Dominique M. Hanssens & Marnik G. Dekimpe
ERS-2002-21-MKT

How cannibalistic is the internet channel?
Barbara Deleersnyder, Inge Geyskens, Katrijn Gielens & Marnik G. Dekimpe
ERS-2002-22-MKT

Evaluating Direct Marketing Campaigns; Recent Findings and Future Research Topics
Jedid-Jah Jonker, Philip Hans Franses & Nanda Piersma
ERS-2002-26-MKT

The Joint Effect of Relationship Perceptions, Loyalty Program and Direct Mailings on Customer Share Development
Peter C. Verhoef
ERS-2002-27-MKT

Estimated parameters do not get the "wrong sign" due to collinearity across included variables
Philip Hans Franses & Christiaan Hey
ERS-2002-31-MKT

2001
Predicting Customer Potential Value. An application in the insurance industry
Peter C. Verhoef & Bas Donkers
ERS-2001-01-MKT

---

* A complete overview of the ERIM Report Series Research in Management:
http://www.ers.erim.eur.nl

ERIM Research Programs:
LIS Business Processes, Logistics and Information Systems
ORG Organizing for Performance
MKT Marketing
F&A Finance and Accounting
STR Strategy and Entrepreneurship

Modeling Potenitally Time-Varying Effects of Promotions on Sales
Philip Hans Franses, Richard Paap & Philip A. Sijthoff
ERS-2001-05-MKT

Modeling Consideration Sets and Brand Choice Using Artificial Neural Networks
Björn Vroomen, Philip Hans Franses & Erjen van Nierop
ERS-2001-10-MKT

Firm Size and Export Intensity: A Transaction Costs and Resource-Based Perspective
Ernst Verwaal & Bas Donkers
ERS-2001-12-MKT

Customs-Related Transaction Costs, Firm Size and International Trade Intensity
Ernst Verwaal & Bas Donkers
ERS-2001-13-MKT

The Effectiveness of Different Mechanisms for Integrating Marketing and R & D
Mark A.A.M. Leenders & Berend Wierenga
ERS-2001-20-MKT

Intra-Firm Adoption Decisions: Departmental Adoption of the Common European Currency
Yvonne M. van Everdingen & Berend Wierenga
ERS-2001-21-MKT

Econometric Analysis of the Market Share Attraction Model
Dennis Fok, Philip Hans Franses & Richard Paap
ERS-2001-25-MKT

Buying High Tech Products: An Embeddedness Perspective
Stefan Wuyts, Stefan Stremersch & Philip Hans Franses
ERS-2001-27-MKT

Changing Perceptions and Changing Behavior in Customer Relationships
Peter C. Verhoef, Philip Hans Franses & Bas Donkers
ERS-2001-31-MKT

How and Why Decision Models Influence Marketing Resource Allocations
Gary L. Lilien, Arvind Rangaswamy, Katrin Starke & Gerrit H. van Bruggen
ERS-2001-33-MKT

An Equilibrium-Correction Model for Dynamic Network Data
David Dekker, Philip Hans Franses & David Krackhardt
ERS-2001-39-MKT

Aggegration Methods in International Comparisons: What Have We Learned?
Bert M. Balk
ERS-2001-41-MKT

The Impact of Channel Function Performance on Buyer-Seller Relationships in Marketing Channels
Gerrit H. van Bruggen, Manish Kacker & Chantal Nieuwlaat
ERS-2001-44-MKT

Incorporating Responsiveness to Marketing Efforts when Modeling Brand Choice
Dennis Fok, Philip Hans Franses & Richard Paap
ERS-2001-47-MKT

Competitiveness of Family Businesses: Distinghuising Family Orientation and Business Orientation
Mark A.A.M. Leenders & Eric Waarts
ERS-2001-50-MKT

The Effectiveness of Advertising Matching Purchase Motivation: An Experimental Test
ERS-2001-65-MKT
Joost Loef, Gerrit Antonides & W. Fred van Raaij

Using Selective Sampling for Binary Choice Models to Reduce Survey Costs
ERS-2001-67-MKT
Bas Donkers, Philip Hans Franses & Peter Verhoef

Deriving Target Selction Rules from Edogenously Selected Samples
ERS-2001-68-MKT
Bas Donkers, Jedid-Jah Jonker, Philip Hans Franses & Richard Paap

2000
Impact of the Employee Communication and Perceived External Prestige on Organizational Identification
Ale Smidts, Cees B.M. van Riel & Ad Th.H. Pruyn
ERS-2000-01-MKT

Forecasting Market Shares from Models for Sales
Dennis Fok & Philip Hans Franses
ERS-2000-03-MKT

The Effect of Relational Constructs on Relationship Performance: Does Duration Matter?
Peter C. Verhoef, Philip Hans Franses & Janny C. Hoekstra
ERS-2000-08-MKT

Informants in Organizational Marketing Research: How Many, Who, and How to Aggregate Response?
Gerrit H. van Bruggen, Gary L. Lilien & Manish Kacker
ERS-2000-32-MKT

The Powerful Triangle of Marketing Data, Managerial Judgment, and Marketing Management Support Systems
Gerrit H. van Bruggen, Ale Smidts & Berend Wierenga
ERS-2000-33-MKT

Consumer Perception and Evaluation of Waiting Time: A Field Experiment
Gerrit Antonides, Peter C. Verhoef & Marcel van Aalst
ERS-2000-35-MKT

Broker Positions in Task-Specific Knowledge Networks: Effects on Perceived Performance and Role Stressors in an Account Management System
David Dekker, Frans Stokman & Philip Hans Franses
ERS-2000-37-MKT

Modeling Unobserved Consideration Sets for Household Panel Data
Erjen van Nierop, Richard Paap, Bart Bronnenberg, Philip Hans Franses & Michel Wedel
ERS-2000-42-MKT

A Managerial Perspective on the Logic of Increasing Returns
Erik den Hartigh, Fred Langerak & Harry Commandeur
ERS-2000-48-MKT

The Mediating Effect of NPD-Activities and NPD-Performance on the Relationship between Market Orientation and Organizational Performance
Fred Langerak, Erik Jan Hultink & Henry S.J. Robben
ERS-2000-50-MKT

Sensemaking from actions: Deriving organization members' means and ends from their day-to-day behavior
Johan van Rekom, Cees B.M. van Riel & Berend Wierenga
ERS-2000-52-MKT