

# Direct Regression, Reverse Regression, and Covariance Structure Analysis

CLAES FORNELL\*

*University of Michigan, School of Business Administration, Ann Arbor, MI 48109-1234*

BYONG-DUK RHEE

*Washington University, St. Louis*

YOUJAE YI

*University of Michigan*

**Key words:** Direct Regression, Reverse Regression, Covariance Structure Analysis, Unbiased Estimation

[October 1990]

## *Abstract*

This paper discusses the issues in estimating the effects of marketing variables with linear models. When the variables are not directly observable, it is well known that direct regression yields biased estimates. Several researchers have recently suggested reverse regression as an alternative procedure. However, it is shown that the reverse regression approach also fails to provide unbiased estimates in general, except for some special cases. It is proposed that covariance structure analysis with an appropriate measurement model can ensure the unbiasedness of estimated effects. These issues are examined in the context of assessing market pioneer advantages.

Marketing researchers are often interested in estimating the impact of a certain exogenous variable on the endogenous variable while controlling for the effects of other exogenous variables. For example, one might ask a question: Do market pioneers get higher market shares than equally performing late entrants? A conventional approach to answering this question is to regress a market share variable ( $M$ ) on a pioneer dummy variable ( $D = 1$  for market pioneers,  $D = 0$  for late entrants) and a performance variable ( $P$ ).<sup>1</sup> That is, the following linear model is used:

$$M = \alpha D + \beta P + u \quad (1)$$

where  $\alpha$  indicates the advantage that pioneers have over late entrants after controlling for the firms' performance.

\* The authors thank the editor and the two anonymous reviewers for their helpful comments on the previous version of this paper.

An important task is then to estimate the market pioneer advantage ( $\alpha$ ). When observations on all variables are available, its estimation is straightforward; direct regression of  $M$  on  $D$  and  $P$  will provide unbiased estimates of  $\alpha$ . However, estimating the market pioneer advantage is problematic if another predictor affecting the market share (e.g., performance) is unobservable and is replaced with its correlates that are observable. This problem, identified originally in the economics literature on the theory of distributive justice (e.g., Conway and Roberts 1983; Greene 1984), has recently been introduced to marketing researchers (e.g., Vanhoner and Day 1987).

Researchers (Vanhoner and Day 1987) have identified the case where the usual regression (often-called direct regression) provides biased estimates and in its place suggested an alternative procedure for obtaining an unbiased estimate of  $\alpha$  on the basis of related econometric research (e.g., Goldberger 1984). This estimation procedure is referred to as "reverse regression" since the roles of endogenous and exogenous variables are reversed; that is, exogenous variables (e.g.,  $P$ ) are regressed on endogenous variables (e.g.,  $M$ ). Indeed, reverse regression may yield unbiased estimates under certain circumstances. However, as we show later in this paper, reverse regression fails to provide unbiased estimates in general. Given the widespread use of regression analysis and the difficulty of directly observing variables in marketing, it seems necessary to understand these problems and develop an estimation procedure that can provide unbiased estimates.

The purpose of this paper is therefore (1) to investigate the problems in estimating linear regression models with unobservable variables, (2) to show the limitations of reverse regression, a method recently suggested in place of the often-used direct regression, (3) to propose an alternative method that can yield unbiased estimates, and (4) illustrate these alternative procedures in the context of substantive research (i.e., assessing market pioneer advantages).

## **1. Issues in estimation by direct and reverse regression**

In this section, we examine the issues in estimating linear models via direct and reverse regression in the context of assessing market pioneer advantages. Research in econometrics (e.g., Goldberger 1984) suggests that unbiased estimation of the coefficient (e.g.,  $\alpha$ ) for a predictor (e.g.,  $D$ ) depends upon measurement properties of another predictor (e.g.,  $P$ ). Thus, the issues are examined according to the way  $P$  is measured.

### *1.1. Case I: single measures of $P$*

Suppose  $P$  is measured with a single indicator with random error. Such a case can be represented by the following equations.

$$M = \beta P + \alpha D + u$$

$$X = \lambda P + \theta$$

$$P = cD + \varepsilon$$

That is,  $X$  is an observed indicator of  $P$ , and it has the correlation of  $\lambda$  with  $P$ . Also, the correlation between  $D$  and  $P$  is  $c$ .

Several points should be noted with respect to this specification. First, this specification is different from the specification by Goldberger (1984) in that a random error term ( $u$ ) is added to the equation for  $M$ . Goldberger (1984) used a deterministic model without the random error term. However, given a basic and unpredictable element of randomness in market responses, a model with the error term seems to be more justified (see Johnston 1984). Vanhonacker and Day (1987) also used the model containing random error. Furthermore, the model with random error is a general case, because the deterministic model is its special case when random error is zero.

Note also that we have added an explicit relation between  $D$  and  $P$  to the model specification. Specifically, market pioneering is posited to affected firms' performance. This relation is based on previous research in the area. Several researchers have argued that the order of entry gives the pioneer advantages such as broader product lines, higher product quality, lower production cost, lower advertising cost, etc.; for example, market pioneers can develop and position products for the largest and most lucrative segments and leave the smaller and less desirable market niches for late entrants (Robinson 1988; Robinson and Fornell 1985; Schmalensee 1978). Fershtman, Mahajan and Muller (1990) also provide theoretical support for this relationship. However, the estimation issues and results in this paper hold whether exogenous variables are causally related or merely correlated. Given that exogenous variables are often correlated, the model should be relevant to many research settings.

Let us assume that

$$\begin{array}{ll} E(\varepsilon|D) = 0 & V(\varepsilon|D) = \sigma_{\varepsilon}^2 \\ E(\theta|P,D) = 0 & V(\theta|P,D) = \sigma_{\theta}^2 \\ E(u|P,D) = 0 & V(u|P,D) = \sigma_u^2 \end{array}$$

Then, we get

$$\begin{array}{lll} E(X|D) = \lambda cD & V(X|D) = \lambda^2 \sigma_{\varepsilon}^2 + \sigma_{\theta}^2 & \\ E(M|D) = (\alpha + \beta c)D & V(M|D) = \beta^2 \sigma_{\varepsilon}^2 + \sigma_u^2 & \text{Cov}(X, M|D) = \lambda \beta \sigma_{\varepsilon}^2 \end{array}$$

Let us first consider the parameter estimators from direct regression. When the

above models are true, an application of direct regression ( $M = \beta P + \alpha D + u$ ) will give the following estimators.

$$\hat{\beta} = \frac{\lambda \beta \sigma_e^2}{\lambda^2 \sigma_e^2 + \sigma_\theta^2}$$

$$\hat{\alpha} = \alpha + \frac{\beta c \sigma_\theta^2}{\lambda^2 \sigma_e^2 + \sigma_\theta^2} = \alpha + \beta c (1 - \rho_p)$$

where  $\rho_p$  is the reliability of the P measure. Note that the bias in the estimator for the coefficient ( $\alpha$ ) of D is a direct function of (1) the unreliability of the P measure ( $1 - \rho_p$ ), (2) the true coefficient ( $\beta$ ) for P, and (3) the correlation ( $c$ ) between P and D. Since  $\beta$  and  $c$  are expected to be positive in general, the estimator of  $\alpha$  is biased upward in direct regression.

What do these results imply? Estimating the effect of a predictor (the market pioneer advantage) with direct regression is problematic especially when other predictors (e.g., performance) exhibit relatively large measurement errors. For instance, if we measure D with considerable accuracy but have only a rough indicator of P, the effect of D will be exaggerated. This is because the coefficient ( $\beta$ ) of the poorly measured variable will be attenuated, but the coefficient ( $\alpha$ ) of the well-measured variable will be amplified. That is, the estimated impact of one predictor is affected by the measurement properties (i.e., reliabilities) of other predictors incorporated in the regression equation. The estimation is also difficult when other predictors are highly correlated with the focal variable.

Let us now consider the parameter estimators from reverse regression. If reverse regression ( $X = gM + dD + v$ ) is used, the following estimators are obtained.

$$\hat{g} = \frac{\lambda \beta \sigma_e^2}{\beta^2 \sigma_e^2 + \sigma_u^2}$$

$$\hat{d} = -\frac{\alpha \beta \lambda \sigma_e^2}{\beta^2 \sigma_e^2 + \sigma_u^2} + \lambda c \left( 1 - \frac{\beta^2 \sigma_e^2}{\beta^2 \sigma_e^2 + \sigma_u^2} \right)$$

$$\hat{\alpha}^* = -\frac{\hat{d}}{\hat{g}} = \alpha - \frac{c \sigma_u^2}{\beta \sigma_e^2}$$

The estimator of  $\alpha$  from reverse regression is biased downward. When the reverse regression approach is taken, the estimator of  $\alpha$  is still biased, but the direction of the bias is downward. The magnitude of the bias is a function of the ratio between the error variance for M and the error variance for P, as well as the correlation between D and P. In sum, when P is measured with a single indicator with random error, neither direct regression nor reverse regression provides an unbiased estimate for the effect of D.

Now let us examine some special cases of the model and implications for unbiased estimation. Let us first look at the case when the random error term is zero ( $u=0$  and  $\sigma_u^2=0$ ), which is the model used by Goldberger (1984). We can note that the estimator of  $\alpha$  from reverse regression is unbiased in such a case. A second special case might occur when D and P are totally uncorrelated (i.e.,  $c=0$ ). It can be noted that the estimator of  $\alpha$  from direct or reverse regression is unbiased in this case. That is, under certain special circumstances direct and reverse regression can provide unbiased estimators. However, these situations are unlikely to occur in practice, and direct or reverse regression is of limited applicability for unbiased estimation.

### 1.2. Case II: multiple measures of P

Marketing variables (e.g., firms' performance P) are often measured with a set of indicators ( $X'=(X_1, \dots, X_n)$ ), rather than with a single indicator. Since such a case is most likely, it will be the focus of this paper. Based on Goldberger's (1984) results, Vanhoneracker and Day (1987) consider two measurement models for P with multiple measures: Formative and Reflective Measurement Models. The measurement vector X will be imperfectly correlated with P, a latent variable interpreted as performance. These two models differ in epistemic relationships between X and P.

Formative Measurement Model specifies that the elements of X ( $X_i$ 's) are multiple causes of P. That is,  $X_i$ 's are *formative indicators* of P, and P is a linear combination of  $X_i$ 's. The equation for this measurement model is as follows:

$$P = \Gamma X + v \quad (2)$$

Reflective Measurement Model hypothesizes that the elements of X ( $X_i$ 's) are multiple indicators of P. In other words, P is an unobservable factor underlying a set of measures ( $X_i$ 's), and  $X_i$ 's are *reflective indicators* of P. The following is the equation for the Reflective Measurement Model:

$$X = \Lambda P + \Theta \quad (3)$$

Let us consider the latter case where P is measured by multiple reflective indicators with measurement errors. This case can be represented by the following equations:

$$\begin{aligned} M &= \beta P + \alpha D + u \\ X &= \Lambda P + \Theta \\ P &= cD + \varepsilon \end{aligned}$$

Let us assume

$$\begin{array}{ll} E(e|D) = 0 & V(\varepsilon|D) = \sigma_e^2 \\ E(\Theta|P, D) = 0 & V(\Theta|P, D) = \Omega \\ E(u|P, D) = 0 & V(u|P, D) = \sigma_u^2 \end{array}$$

Then

$$\begin{array}{lll} E(X|D) = \Lambda c D & V(X|D) = \Lambda \Lambda' \sigma_e^2 + \Omega & \\ E(M|D) = (\alpha + \beta c) D & V(M|D) = \beta^2 \sigma_e^2 + \sigma_u^2 & \text{Cov}(X, M|D) = \Lambda B \sigma_e^2 \end{array}$$

When the above models are true, direct regression ( $M = bX + \alpha D + v$ ) would give the following estimators:

$$\hat{b} = (\Lambda' \Omega^{-1} \Lambda \sigma_e^2 + 1)^{-1} \Omega^{-1} \Lambda \beta \sigma_e^2$$

$$\hat{\alpha} = \alpha + \frac{\beta c}{\Lambda' \Omega^{-1} \Lambda \sigma_e^2 + 1}$$

If reverse regression ( $\hat{b}'X = \hat{b}'gM + \hat{b}'dD + \omega$ ) were used, the estimators would be

$$\hat{b}'\hat{g}' = \frac{\hat{b}'\Lambda\beta\sigma_e^2}{\beta^2\sigma_e^2 + \sigma_u^2}$$

$$\hat{b}'\hat{d} = \hat{b}'\Lambda \left[ \left(1 - \frac{\beta^2\sigma_e^2}{\beta^2\sigma_e^2 + \sigma_u^2}\right)c - \frac{\alpha\beta\sigma_e^2}{\beta^2\sigma_e^2 + \sigma_u^2} \right]$$

$$\hat{\alpha}^* = -\frac{\hat{b}'\hat{d}}{\hat{b}'\hat{g}'} = \alpha - \frac{c\sigma_u^2}{\beta\sigma_e^2}$$

It can be noted that the estimators from direct regression are biased upward, while the estimators from reverse regression are biased downward. The bias is a function of the ratio between the error variance for  $M$  ( $\sigma_u^2$ ) and the error variance for  $P$  ( $\sigma_e^2$ ), the effect of  $P$  on  $M$  ( $\beta$ ), and the correlation ( $c$ ) between  $D$  and  $P$ .

We can again examine the special cases of the model and implications for unbiased estimation. We can note that direct and reverse regression may provide unbiased estimators under certain circumstances. Specifically, when the random error term is zero ( $u = 0$  and  $\sigma_u^2 = 0$ ), as modeled by Goldberger (1984), the estimator of  $\alpha$  from reverse regression is unbiased. When  $D$  and  $P$  are uncorrelated (i.e.,  $c = 0$ ), the estimator of  $\alpha$  from direct or reverse regression is unbiased. These findings are the same as those in the single-indicator case.

## 2. Estimation by covariance structure analysis

We have seen that direct and reverse regression may fail to provide unbiased estimates of  $\alpha$  when variables are unobservable. Then, a question arises: How can we obtain an unbiased estimate? It is suggested in this paper that unbiased estimation can be achieved by a covariance structure approach. The measurement models in Equations 1 and 2 can be combined with the structural model in Equation 1 into overall causal models. Let us first consider Formative Measurement Model. If we substitute Equation 2 into Equation 1, we get

$$M = \alpha D + \beta(\Gamma X + v) + u = \alpha D + \beta \Gamma X + (\beta v + u) = \alpha D + \beta^* X + u^*,$$

where

$$\beta^* = \beta \Gamma \text{ and } u^* = \beta v + u.$$

This equation can be represented as an overall model provided in Figure 1A. Under Reflective Measurement Model, the overall model can be represented as a set of measurement and structural models, as illustrated in Figure 1B.

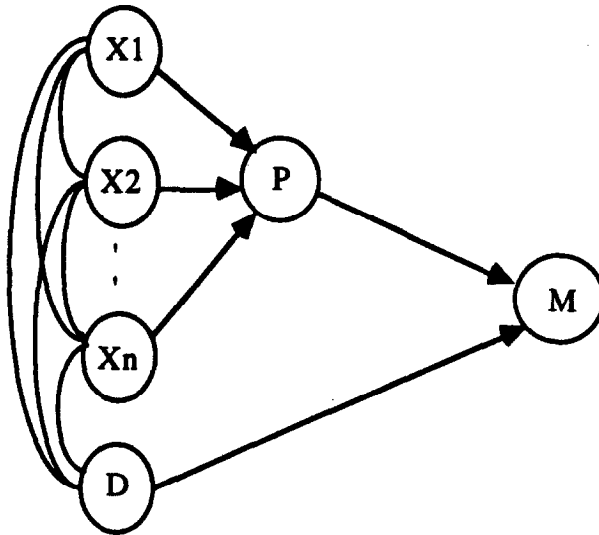
If one knows which of the two overall models is correct, the estimation of that model via covariance structure analysis should provide asymptotically unbiased estimates of  $\alpha$  (Jöreskog and Sörbom 1984). An important task for unbiased estimation is therefore to choose between Models A and B in Figure 1. Next, we will examine this issue in detail.

### 2.1. Formative vs. reflective specification

As mentioned earlier, an unbiased estimation of  $\alpha$  will depend on the choice of the correct model in Figure 1, which in turn depends upon the choice of the measurement model. That is, the choice between reflective and formative indicator models is critical to deciding which model should be used to estimate  $\alpha$ .

The choice of indicator mode should be made primarily on the basis of the substantive theory behind the model: the way in which variables are conceptualized (e.g., Fornell and Bookstein 1982, pp. 441–442). Constructs such as “attitude” or “personality” are typically viewed as underlying factors that give rise to something that is observed. In such a case, the reflective indicator model would be used. In contrast, constructs such as “socioeconomic status (SES)” might be conceived as composites rather than as factors. That is, instead of SES generating variables such as education, income, and occupational prestige, these variables are more appropriately seen as causing changes in SES. In such a case, constructs can be seen as explanatory combinations of indicators, and their indicators should be represented as formative.

(a) Overall model with formative indicators



(b) Overall model with reflective indicators

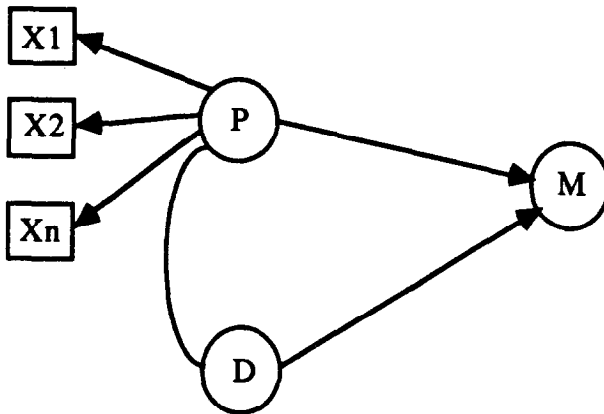


Figure 1. Causal model representations.

To gain further insights into the difference between formative vs. reflective specification, let us compare the two formally. If  $O$  is the observed measure,  $T$ , the true score and  $e$  an error component, it is well-known that the reflective specification is:

$$O = T + e$$



with the assumptions that

$$\begin{aligned} E(e) &= 0, \text{Cov}(T, e) = 0, \text{Cov}(e_i, e_j) = 0, \text{ which imply} \\ \text{Var}(O) &= \text{Var}(T) + \text{Var}(e) \text{ and} \\ \text{Var}(T) &< \text{Var}(O). \end{aligned}$$

That is, the variance in the true scores is *smaller* than the variance in the measured variables. However, in the formative specification, the opposite is true. Now we have:

$$T = O + e \text{ or [since } E(e) = 0]$$

we can write this as

$$T = O - e$$

which brings us back to the reflective equation

$$O = T + e.$$

But since  $e$  represents all remaining causes of  $T$  other than  $O$ , we also have

$$\begin{aligned} \text{Cov}(T, e) &\neq 0 \text{ and instead} \\ \text{Cov}(O, e) &= 0, \text{ which imply} \\ \text{Var}(T) &> \text{Var}(O). \end{aligned}$$

In addition to conceptual issues, the choice of formative vs. reflective measurement models has implications for the predictive power (within the data) as well as assessing the individual contribution of one's measures. Clearly, the reflective formulation can never account for more variance in the dependent variable than the formative specification. Typically, the latter will do better on this score. On the other hand, in case of high multicollinearity among the measured variables, formative specifications make it difficult to assess the individual contribution of these variables.

In sum, then, the choice of reflective vs. formative measurement specifications rests primarily on conceptual considerations and the purpose of the modelling effort. Empirical matters also play a role. Ideally, the choice should be made a priori on the basis of theoretical reasoning. If one's theory does not provide an unequivocal decision on this score, one may have to depend on empirical evidence. However, it is more difficult to determine measurement specification on the basis of covariance structures because the formative model is always just identified with zero degrees of freedom. It is possible however to evaluate the covariance fit of a reflective model as long as it has more than three indicators.

## 2.2. An illustration

It might be useful to illustrate alternative estimation procedures and examine the potential biases with an example. Recall that direct regression provides unbiased estimators in the case of formative indicators. Given that the case with reflective indicators causes problems for analysis, we have focused on the model with reflective indicators. We have constructed a “true” model with plausible parameter values and generated data accordingly, rather than using actual data for which the true parameter values are unknown, which will cause difficulty in assessing the magnitude of bias.

Suppose the model in Figure 2A represents a true model under investigation. A latent variable  $P$  has three reflective indicators  $X_1 - X_3$ . The correlation between  $D$  and  $P$  is 0.40, while the coefficients relating  $P$  to  $D$  and  $M$  are 0.70 and 0.25, respectively. That is, a true value for  $\alpha$  is known to be 0.25. From this model, a data set of 100 cases is generated via the normal random number generator. The resulting correlation matrix for the data is given in Figure 2B. For the sake of facilitating the argument and without loss of generality, we will treat this as the population. The three alternative estimation procedures are then applied to this data set, and the estimates for  $\alpha$  are compared.

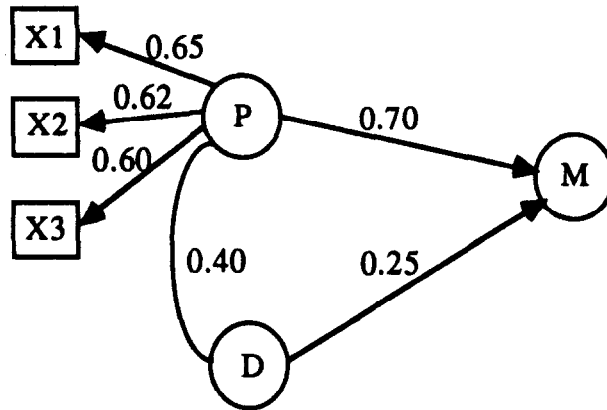
When the direct regression approach is used, the estimate of  $\alpha$  is 0.357, which is higher than the true value of 0.25. As expected, direct regression yielded an overestimate of  $\alpha$  (i.e., with an upward bias of 0.107). When the reverse regression approach is employed, the estimate of  $\alpha$  is 0.041. Thus, there is a downward bias (0.209) from reverse regression, which is again consistent with the analytic results presented earlier. In contrast, the covariance structure approach yields an unbiased estimate of 0.25. Also, the fit is perfect ( $X^2(4) = 0.00, p = 1.00$ ), since we are working with the population correlation matrix for the correct model specification.

This example shows that estimates from direct and reverse regression are biased when a true model has an unobservable variable with reflective indicators. We have used the model and parameter values that are likely to occur in typical research settings. However, the magnitude of bias is substantial enough to yield misleading inferences about the effect of predictors in the model. These results suggest that researchers should pay special attention to estimation when unobservable variables are included in the model.

## 3. Discussion and conclusion

The suggested method based on covariance structure analysis has several advantages over other practices (e.g., direct or reverse regression) in estimating linear models. First, reverse regression fails to provide unbiased estimates when the dependent variable exhibits random error. Reverse regression gives unbiased estimates only when no random errors exist for dependent variables (Goldberger

(a) Specification of the true model.



(b) Correlation matrix for the data generated from the model.

---

<b>M</b>	1.00				
<b>X<sub>1</sub></b>	.520	1.00			
<b>X<sub>2</sub></b>	.496	.403	1.00		
<b>X<sub>3</sub></b>	.480	.390	.372	1.00	
<b>D</b>	.530	.260	.248	.240	1.00

---

Figure 2. Model and data used for illustration.

1984). Such a situation is unlikely for marketing data, which usually contain stochastic errors due to the imperfectness in measurement or the randomness in phenomena. Reverse regression is therefore likely to yield biased estimates. On the other hand, the suggested approach provides asymptotically unbiased estimates even when dependent variables contain random errors.

Second, the suggested method is consistent in logic. The previous practice consists of two steps: 1) obtain estimates of  $\beta$  from direct regression and 2) use these estimates in reverse regression. But if the reverse regression model is correct, why should one use estimates that are obtained from a different, incorrect model (i.e., direct regression model)? In contrast, the suggested procedure uses the correctly specified model in estimation. There is no other specification.

We have examined the issues in estimating the effects of marketing variables with an example of market pioneer advantages. When there are errors in variables, as is typical in marketing data, the often-used direct regression and the recently suggested reverse regression yield biased estimates for and misleading conclusions about the impact of a predictor. It is proposed that covariance structure analysis in conjunction with an appropriate measurement model be used instead.

## Notes

1. Although we focus on the impact of a discrete variable (i.e., a pioneer dummy variable) in this paper, this question is generalizable to a continuous exogenous variable. For example, what are the effects of advertising expenditures on revenue? Also, the results in this paper are valid whether the exogenous variable is discrete or continuous.

## References

- Conway, Delores A., and Harry V. Roberts. (1983). "Reverse Regression, Fairness, and Employment Discrimination," *Journal of Business and Economic Statistics* 1 (January), 75-85.
- Fershtman, Chaim, Vijay Mahajan, and Eitan Muller. (1990). "Market Share Pioneering Advantage: A Theoretical Approach," *Management Science* 36 (August), 900-918.
- Fornell, Claes, and Fred L. Bookstein. (1982). "Two Structural Equation Models: LISREL and PLS Applied to Consumer Exit-Voice Theory," *Journal of Marketing Research* 19 (November), 440-452.
- Goldberger, Arthur S. (1984). "Redirecting Reverse Regression," *Journal of Business and Economic Statistics* 2 (April), 114-116.
- Greene, William H. (1984). "Reverse Regression: The Algebra of Discrimination," *Journal of Business and Economic Statistics* 2 (April), 117-120.
- Johnston, J. (1984). *Econometric Methods*, 3rd ed., NY: McGraw-Hill.
- Jöreskog, Karl G., and Dag Sörbom. (1984). *LISREL: Analysis of Linear Structure Relationships by the Method of Maximum Likelihood*. Mooresville, IN: Scientific Software.
- Robinson, William T. (1988). "Sources of Market Pioneer Advantages: The Case of Industrial Goods Industries," *Journal of Marketing Research* 25 (February), 87-94.
- Robinson, William T., and Claes Fornell. (1985). "Sources of Market Pioneer Advantages in Consumer Goods Industries," *Journal of Marketing Research* 22 (August), 305-317.
- Schmalensee, Richard. (1978). "Entry Deterrence in the Ready-to-eat Breakfast Cereal Industry," *The Bell Journal of Economics* 9, 305-327.
- Vanhonacker, Wilfred R., and Diana Day. (1987). "Cross-Sectional Estimation in Marketing: Direct Versus Reverse Regression," *Marketing Science* 6 (Summer), 254-267.