The impact of research designs on R^2 in linear regression models: an exploratory meta-analysis

Heribert Reisinger
Institute of Management, Department of Marketing, University of Vienna, A-1210 Vienna, Austria

In this paper the author tries to identify various influences on the coefficient of determination (R^2) which originate in the research designs of empirical studies, rather than in the research subjects within the framework of an exploratory meta-analysis. The following results are obtained: the larger the sample size and the smaller the number of regressors in a study, the smaller is R^2; time-series studies achieve higher values for R^2 than cross-sectional studies; studies with secondary data achieve higher values for R^2 than studies with primary data; publications in the *Journal of Marketing Research* show on average lower values for R^2 than publications in the *International Journal of Research in Marketing* and in *Marketing Letters*.

KEYWORDS: coefficient of determination, linear regression model, marketing research

INTRODUCTION

The classical linear regression model is the standard procedure for analyzing dependencies between variables that are measured on a metric scale. In the course of model estimation, it is common practice to assess the appropriateness of a single descriptive model for the problem under study with the help of the coefficient of determination, R^2. In empirical studies, the most important benefit of R^2 is that it serves as a fast and easily interpretable measure for the goodness of fit of the estimated model. This advantage, however, comes with a big caveat, i.e. one may over-value the relevance of coefficients of determination, especially in the model selection process, and dedicate only minor interest to the substantive results of the analysis (e.g. the estimated parameters or corresponding t-statistics; see Mayer 1975). R^2 is no absolute indicator of goodness of fit; it is just a relative measure (explained variance relative to total variance in the dependent variable). Some authors largely reject the usage of the coefficient of determination, e.g. Achen (1982):

> "Thus R^2 gives the 'percentage of variance explained' by the regression, an expression that, for most social scientists, is of doubtful meaning but great rhetorical value. (...) But it makes little sense to base decisions on a statistic that, for most social science applications, measures nothing of serious importance. 'Explaining variance' is not what social science is about."

Despite this, practical empirical work shows us that coefficients of determination are regularly published. Anderson-Sprecher's quotation (1994) may give some explanation of Achen's view:

> "The coefficient of multiple determination, R^2, is a measure many statisticians love to hate. This animosity exists primarily because the widespread use of R^2 inevitably leads to at least occasional misuse. (...) The R^2 measure is unlikely to disappear any time soon, however, and any action that improves its usage will be valuable."

Considering the advantages and disadvantages of calculating R^2 in empirical studies, one may ask whether it makes sense to evaluate a model by means of a single descriptive measure at all. For example, from a statistical point of view, the analyzed data set is irrelevant when deciding on the appropriateness of the model under consideration. However, a market researcher clearly distinguishes whether he studies time-series or cross-sectional data. For instance, it is a well-known fact that on average one may expect larger coefficients of determination for time-series data than for cross-sectional data (see e.g. Naert, Leeflang 1978, or Pindyck, Rubinfeld 1991). Starting from this known phenomenon, the question arises whether there are other influences on R^2 which originate in the research designs of empirical studies rather than in the research subjects.

FRAMEWORKOFTHEANALYSIS

In order to find some answers to this question, performing an exploratory meta-analysis seems to be an appropriate starting point for our discussion. In the course of the meta-analysis, various empirical studies have been collected (in this paper we will focus on marketing studies), and a set of potential influencing factors on R^2 has been specified. From a methodological point of view, the data analysis will include different regression analyses with R^2 as the dependent variable, and the potential influencing factors as explanatory variables. The meta-analysis has an exploratory nature, because no *a priori* theory concerning the influences of the various potential impacts on R^2 exists (besides the difference between cross-sectional and time-series data). Recently Peterson (1994) published a comparable (but more extensive) meta-analysis concerning Cronbach's coefficient alpha.

When we presented the results of our meta-analysis before, we sometimes had to defend our work against the argument that variations in R^2 are likely to be much more a function of the substantive nature of the study being undertaken than of research design. This argument is of course true, but the main focus of the present study is to find some regularities in R^2 that originate in the research designs of marketing studies, apart from the clear impact of the substantive nature of a study on R^2. The proportion of variance explained is a basic benchmark in evaluating the results of an empirical regression model, and, as we see it, if the results of our meta-analysis show some regularity e.g. between

R^2 and the sample size, this finding is important in assessing the overall fit of the model for the application at hand.

The exploratory nature of this study can be compared to some extent to Ehrenberg's seminal book (1988) on repeat buying. Ehrenberg's main interest lies in finding empirical generalizations in buying behavior. He does not study substantive reasons for buying behavior. In a recent Special Issue of *Marketing Science* entitled *Empirical Generalizations in Marketing*, Bass and Wind (1995) note that during the conference leading to the papers in this Special Issue, some criteria for research remained unresolved: "... At the one extreme are those who feel that empirical generalizations do not have to be based on theory, derived from theory or leading to the development of a theory. Yet, others require that empirical generalizations be theoretically sound." The present paper can be classified as empirically driven research and therefore represents rather the "non-theory-guided" edge of the above mentioned continuum.

Detailed information about the sample collection and the specification of the explanatory variables of the present meta-analysis is given by Reisinger (1996). In the following, only some comments of principle importance on these issues will be made. The data base consists of 105 regression models using OLS estimation, taken from 44 studies published in the *Journal of Marketing Research* (volumes 1992-94), the *International Journal of Research in Marketing* (volumes 1989-94), and *Marketing Letters* (volumes 1989-94). We have chosen these three marketing journals as data sources for our meta-analysis because they emphasize the publication of empirical work. Approximately one third of the data base was taken from each journal. This is the reason why only three volumes of the *Journal of Marketing Research* are used as data sources in the meta-analysis. Various models of the same study have been integrated separately in this analysis, if (a) the regressands are not the same, or (b) the regressands are identical, but all the regressors are different. In all other cases, averages of the variables of interest have been calculated. In the specific situation of nested models, the model with the largest number of regressors has been included in the meta-analysis.

The explanatory variables that are regarded as potential influencing factors on R^2 are summarized in Table 1. The table also includes a list of the different levels of the qualitative variables. Furthermore, the abbreviations that are used throughout the remaining sections of this paper and the encoding of the qualitative variables are given.

TABLE 1. Explanatory variables in the meta-analysis

Variable	Levels	Abbrev.	Encoding
Qualitative variables:			
● Data type	- time-series data		0 0
	- cross-sectional data	D_C	1 0
	- time-series and cross-sectional data (pooled data)	D_P	0 1
● Journal	- Marketing Letters		0 0
	- International Journal of Research in Marketing	IJRM	1 0
	- Journal of Marketing Research	JMR	0 1
● Data collection method		COL	
	- primary		1
	- secondary		0
● Data source		SOU	
	- single source		1
	- multiple source		0
● Examination of the assumption of homoscedasticity		HOM	
	- yes		1
	- no		0
● Examination of the correlation between the regressors		COR	
	- yes		1
	- no		0
Quantitative variables:			
● Number of regressors in the study		K	
● Sample size		N	

FORMULATION OF THE HYPOTHESES

As the analysis has mainly an exploratory character, we will formulate the hypotheses as null hypotheses. The results of the data analysis will then give guidelines for possible future work. Nevertheless, we will discuss possible results in order to reflect our personal expectations before model estimation.

- Null hypothesis concerning the variable "Data type":

  > H_{01}: "Studies with time-series, cross-sectional and pooled data do not differ from each other with regard to the corresponding values of R^2."

  H_{01} is expected to be rejected. As already noted in the introduction, values of R^2 should be higher in time-series analyses than in cross-sectional analyses. The reason for this phenomenon is quite obvious: within a cross-section consisting of a number of different (heterogeneous) objects of investigation, the proportion of variance that cannot be explained is usually higher than with time-series data, where only one object of investigation is studied over a given time period. Furthermore, as pooled data are a combination of time-series and cross-sectional data, we expect the corresponding R^2s to be larger than R^2s in cross-sectional analyses, but smaller than R^2s in time-series analyses.

- Null hypothesis concerning the variable "Journal":

  > H_{02}: "Studies in the *International Journal of Research in Marketing*, in the *Journal of Marketing Research* and in *Marketing Letters* do not differ from each other with regard to the corresponding values of R^2."

  There seems to be no reason why differences should exist between the three journals. We therefore expect H_{02} not to be rejected.

- Null hypothesis concerning the variable "Data collection method":

  > H_{03}: "Studies with primary and secondary data do not differ from each other with regard to the corresponding values of R^2."

  Since primary data are collected especially for the problem under consideration, we expect studies with primary data to achieve higher values for R^2 than studies with secondary data.

- Null hypothesis concerning the variable "Data source":

  > H_{04}: "Studies with single source and multiple source data do not differ from each other with regard to the corresponding values of R^2."

  The merging of different data sources in the case of multiple source data could be responsible for a higher inexplicable variance than in studies with single source data, resulting in lower R^2-values.

- Null hypotheses concerning the variables "Examination of the assumption of homoscedasticity" and "Examination of the correlation between the regressors":

> H_{05}: "Studies that incorporate an explicit examination of the assumption of homoscedasticity, and studies that do not incorporate such an examination, do not differ from each other with regard to the corresponding values of R^2."
>
> H_{06}: "Studies that incorporate an examination of the correlation between the regressors, and studies that do not incorporate such an examination, do not differ from each other with regard to the corresponding values of R^2."

There seems to be no obvious reason why such examinations should have an influence on R^2. We therefore expect H_{05} and H_{06} not to be rejected.

- Null hypothesis concerning the variable "Number of regressors in the study":

> H_{07}: "The number of regressors in a study has no influence on the value of R^2."

H_{07} is inconsistent with a fundamental feature of R^2 which says that as more regressors are integrated into a model with the same regressand, the value of R^2 will increase or at least stay constant. Analogous to this feature, we also expect R^2 to achieve higher values when more regressors are considered in the case of different studies, so that H_{07} will be rejected.

- Null hypothesis concerning the variable "Sample size":

> H_{08}: "The sample size has no influence on the value of R^2."

As far as we know, no theories have been published concerning the relationship between R^2 and the sample size. From a statistical point of view, a larger sample size will lead to more precise estimation results. However, a greater precision of parameter estimates, resulting from a larger sample size, does not imply that the variance which cannot be explained will diminish. We therefore expect H_{08} not to be rejected. Another comment regarding the sample size has to be made. As the sample sizes of the empirical studies in our data set range from 17 to 21 600, the resulting parameter estimates for the corresponding variable would be very small. Therefore, the logarithms of the variable N are calculated before the regression analyses are made.

DATA ANALYSIS AND RESULTS

The eight null hypotheses H_{01} to H_{08} will be tested performing various simple and one multiple regression analyses. Because we are estimating several models with the same data set, it is possible to judge the stability of the results achieved. The advantage of the multiple model is that the effects of various explanatory variables on R^2 can be analyzed simultaneously. A possible disadvantage is that in the multiple case the influences of certain variables may not be identified exactly as they overlap due to multicollinearity relations.

The results of the eight simple regression analyses (estimated parameters, *F*-values, corresponding *p*-values and achieved R^2s) are summarized in Table 2. Furthermore, the conclusions regarding the corresponding null hypotheses are included if a Type I error of 5 % is postulated. In the following, we will use the symbol R^2 denoting the coefficients of determination achieved in the course of model estimation in order to avoid any mixing up with R^2 as the dependent variable.

The average R^2s in the data set within the various levels of the qualitative explanatory variables can be derived directly from the corresponding parameter estimates in Table 2. For instance, studies with time-series, cross-sectional and pooled data achieve average R^2s of, respectively, 0.60, 0.31, and 0.52.

TABLE 2. Results of the simple regression analyses

Dependent variable: R^2

Explanatory variables	$\hat{\beta}$	*p*-value	*F*-value	*p*-value	R^2	Conclusion
D_C	-.29	.00	12.81	.00	.20	H_{01} can be rejected.
D_P	-.08	.41				
(Constant)	.60	.00				
IJRM	.10	.14	9.01	.00	.15	H_{02} can be rejected.
JMR	-.14	.03				
(Constant)	.41	.00				
COL	-.25	.00	26.38	.00	.20	H_{03} can be rejected.
(Constant)	.54	.00				
SOU	-.16	.07	3.47	.07	.03	H_{04} cannot be rejected.
(Constant)	.53	.00				
HOM	.15	.13	2.35	.13	.02	H_{05} cannot be rejected.
(Constant)	.38	.00				
COR	.09	.08	3.19	.08	.03	H_{06} cannot be rejected.
(Constant)	.34	.00				
K	.01	.04	4.29	.04	.04	H_{07} can be rejected.
(Constant)	.34	.00				
ln N	-.05	.00	11.43	.00	.10	H_{08} can be rejected.
(Constant)	.66	.00				

- The null hypothesis H_{01} can be rejected as expected. Cross-sectional analyses show significantly lower R^2-values than time-series analyses. Analyses with pooled data do not differ from time-series analyses regarding R^2. The difference between time-series and cross-sectional data conforms with the theoretical considerations in the previous section. However, it is possible to make another interpretation of this finding. Time-series data are often measured at an aggregate level; cross-sectional data can be available at the same level of aggregation but probably often involve data at a lower level (e.g. households). An important consequence of aggregation is that some variation which cannot be explained is usually averaged out, resulting in high R^2-values. Therefore, another contributing factor to the difference in R^2 between time-series and cross-sectional data could be the different aggregation level of the two types of data.

- Hypothesis H_{02} can be rejected, contrary to our expectations. Publications in the JMR show lower average R^2-values than publications in the IJRM and *Marketing Letters*. It is possible to explain the difference between the JMR and the IJRM with the same aggregation argument as before. Nearly all of the studies in the data base with time-series data and pooling data were published in the IJRM. Therefore, average R^2-values in the IJRM tend to be rather high. However, we cannot explain the difference between the JMR and *Marketing Letters* with the aggregation issue, because most of the studies in these two journals work with cross-sectional data. Further interpretations of the differences between the three journals would be highly speculative and have therefore been avoided.

- Hypothesis H_{03} can be rejected, but the sign of the parameter estimate does not correspond to our expectations. Studies with primary data show lower R^2-values than studies with secondary data. This result is highly significant ($p_{\text{COL}} \leq 0.001$). We assume this occurs primarily because of the high correlation between the use of primary data and the performance of a cross-sectional study ($r_{\text{COL,D\_C}} = 0.68$). Again, aggregation could be a contributing factor to the difference between primary and secondary data, as secondary data may be measured at a higher aggregation level.

- Hypotheses H_{04}, H_{05} and H_{06} cannot be rejected.

- The rejection of null hypothesis H_{07} corresponds to our expectations. The fundamental feature "the larger the number of regressors in a study, the higher is R^2" can also be observed comparing various studies.

- The final null hypothesis H_{08} can be rejected, contrary to our expectations. The larger the sample size, the lower is R^2! The implications of this result are remarkable. Because of statistical properties, models based on a larger sample size usually give better (in the statistical sense of more exact) parameter estimates than models with a smaller sample size. Referring to the determination of the sample size in an empirical

> study, the targets of achieving (a) the most precise estimation result, and (b) the highest possible proportion of explained variance, are competing with each other. Given the negative sign of the parameter estimate in the empirical results in Table 2, we can interpret the findings as follows. As the sample size becomes smaller, the (unadjusted) R^2 tends to increase. Because we have used the logarithm of the variable N, we can see that the magnitude of this effect becomes greater as the sample size becomes smaller. A possible explanation of this finding may be based on the difference between the adjusted and unadjusted coefficient of determination. The adjusted R^2 (which accounts for degrees of freedom) is an approximately unbiased estimator of the R^2 in a population, whereas the unadjusted R^2 is biased upward (it overstates true explanatory power). The size of the upward bias depends primarily on the sample size, and secondarily on the number of regressors. Therefore, holding the number of regressors constant, the smaller the sample size, the larger the difference between adjusted and unadjusted R^2-values. However, performing a simple regression analysis with the adjusted coefficient of determination as the dependent variable results in a less negative but still highly significant parameter estimate for ln N ($\hat{\beta}_{\ln N} = -0.04$; $p_{\ln N} = 0.01$).

Table 3 summarizes the results of the multiple regression analysis. Starting from Model 1, which incorporates all explanatory variables of Table 1, Model 2 takes only variables with a p-value ≤ 0.3 into account. Model 3 finally shows the outcome of a forward selection procedure of the six remaining variables of Model 2 with a criterion of inclusion of $p \leq 0.1$. In addition to the parameter estimates and corresponding p-values, the reported results of Model 3 include the estimated standard errors and the beta weights of the four remaining variables.

TABLE 3. Results of the multiple regression analysis

Dependent variable: R^2

Explanatory	Model 1		Model 2		Model 3			
variables	$\hat{\beta}$	p-value	$\hat{\beta}$	p-value	$\hat{\beta}$	$\hat{\sigma}_{\hat{\beta}}$	Beta	p-value
D_C	-.13	.22	-.13	.07	-.19	.05	-.32	.00
D_P	.18	.17	.20	.08	–	–	–	–
IJRM	-.07	.34	–	–	–	–	–	–
JMR	-.16	.01	-.16	.00	-.14	.05	-.27	.00
COL	-.07	.32	–	–	–	–	–	–
SOU	.15	.13	.14	.13	–	–	–	–
HOM	.06	.51	–	–	–	–	–	–
COR	.01	.89	–	–	–	–	–	–
K	.01	.00	.01	.00	.01	.00	.35	.00
ln N	-.07	.00	-.07	.00	-.05	.01	-.31	.00
(Constant)	.74	.00	.67	.00	.78	.07	–	.00
	$F_{10,94} = 7.45$ $R^2 = .44$		$F_{6,98} = 12.27$ $R^2 = .43$		$F_{4,100} = 17.01$ $R^2 = .40$			

The results of the multiple analysis are very stable, and they are nearly identical to the ones obtained from the simple analyses. The coefficients of the variables JMR, K and ln N are highly significant in all three models (JMR and ln N having negative signs and K having a positive sign). The different p-values of the variable D_C are due to high correlation between D_C, IJRM and COL ($r_{\text{D\_C,IJRM}} = -0.71$; $r_{\text{D\_C,COL}} = -0.68$; $r_{\text{IJRM,COL}} = -0.57$). The variable COL, which was highly significant in the simple analysis, shows no significant influence on R^2 in the multiple one. Obviously the effects of the variables D_C and COL overlap in the multiple case. However, if COL is included in Model 3 instead of D_C, the corresponding parameter estimate is highly significant. All other variables (IJRM, D_P, SOU, HOM and COR) have no significant influences on R^2. Summarizing, all conclusions of Table 2 are confirmed by the results of the multiple regression analysis.

CONCLUSIONS AND FUTURE WORK

In this paper we have analyzed potential impacts of research designs on the magnitude of the coefficient of determination achieved in empirical studies. We have found a significant relationship between R^2 and the data type, the data collection method, the number of regressors in a study, and the sample size. Additionally, we have found that publications in the *Journal of Marketing Research* show lower average values for R^2 than publications in the *International Journal of Research in Marketing* and *Marketing Letters*.

Numerous extensions of the outlined meta-analysis are possible. Additional explanatory variables could be integrated into the analysis. In order to enlarge the data base, additional marketing journals or earlier volumes of the JMR and IJRM could be used as data sources. Furthermore, the data base does not have to be restricted to marketing studies. Generally speaking, it would be interesting to analyze the influence of research designs on R^2 and other goodness-of-fit criteria in various linear and nonlinear models, and to compare the results obtained. In this respect, the following contributions have already been made by Reisinger (1996): (a) the performance of a similar meta-analysis to the one outlined in this paper, with the adjusted coefficient of determination as the dependent variable; and (b) the analysis of influences on the Likelihood-Ratio-Index in multinomial logit models.

Evaluating the results of the present study, it seems to be potentially important to include a variable in the meta-analysis that captures the effects of aggregation.However, we do not think that it will be easy to compare the different aggregationlevels of time-series and cross-sectional data directly, which would be necessary to define just one "aggregationvariable". The discussion of the differences between the three data types could be augmented as follows. A combination of time-series and cross-sectional data should increase the average R^2-value if at least dummy variables are included to account for averages between the cross-sections. Furthermore, it would be interesting to study in more detail the impact of the sample size on the adjusted and unadjusted coefficient of determination. We do not think that the statistical argument that R^2 is a measure which overstates true explanatory power entirely explains the empirically observed relationship between R^2 and the sample size.

However, what is probably most important is to validate the results obtained. Replication analyses are needed to see if our results do indeed generalize across other data sets.

ACKNOWLEDGEMENTS

The author would like to thank two anonymous referees, and Udo Wagner and Dick R. Wittink, for their helpful comments on an earlier version of the present paper. Dick R. Wittink kindly provided several interpretations of the findings in Table 2.

REFERENCES

Achen, C.H. (1982) *Interpreting and using regression*, Newbury Park: Sage University Papers.

Anderson-Sprecher, R. (1994) Model comparisons and R^2. *The American Statistician* **48**, **2**, 113–117.

Bass, F.M. and Wind, J. (1995) Introduction to the Special Issue: Empirical Generalizations in Marketing. *Marketing Science* **14***,* **3***,* **2***,* G1–G5.

Ehrenberg, A.S.C. (1988) *Repeat-Buying*, 2nd Edn, New York: Oxford University Press.

Mayer, T. (1975) Selecting economic hypotheses by goodness of fit. *Economic Journal* **85***,* 877–883.

Naert, P.A. and Leeflang, P.S.H. (1978) *Building implementable marketing models*, Leiden: Martinus Nijhoff.

Peterson, A.P. (1994) A Meta-analysis of Cronbach's coefficientalpha. *Journal of Consumer Research,* **21**, 381–391.

Pindyck, R.S. and Rubinfeld, D.L. (1991) *Econometric models and economic forecasts*, 3rd Edn, NewYork: McGraw-Hill.

Reisinger, H. (1996) *Goodness-of-Fit-Maße in linearen Regressions- und Logit-Modellen*, Frankfurt am Main: Lang.