CORNELL
AGRICULTURAL ECONOMICS
STAFF PAPER

92-1

SOME THOUGHTS ON REPLICATION IN EMPIRICAL ECONOMETRICS

William G. Tomek

Department of Agricultural Economics
Cornell University Agricultural Experiment Station
New York State College of Agriculture and Life Sciences
A Statutory College of the State University
Cornell University, Ithaca, New York, 14853

SOME THOUGHTS ON

REPLICATION IN EMPIRICAL ECONOMETRICS

William G. Tomek*

> *"The art of the econometrician consists as much in defining a good model as in finding an efficient statistical procedure. ... Finally, we must never forget that our progress in understanding economic laws depends strictly on the quality and abundance of statistical data. Nothing can take the place of the painstaking work of objective observation of the facts." E. Malinvaud, Statistical Methods of Econometrics, p. 614.*

It has become increasingly clear that econometric results are often poor. In Leamer's terms, results are fragile: small changes in the model or in the data cause large changes in results. But, surprisingly, this issue continues to be ignored by many applied economists. Obtaining high quality empirical results is exceedingly difficult, and the intensity of effort required to obtain truly useful results appears not to be fully appreciated.

One aid in improving quality is to build more carefully on prior research. Researchers should demonstrate precisely how their work improves upon earlier research, and this requires replication of the earlier work. Replication also can help determine the robustness of prior estimates.

This paper reviews the benefits and difficulties of replicating published studies. The difficulties have implications for how results are published and how research is done. If research and publication protocols in applied economics are improved, then the costs of confirming prior results can be reduced. Incidental to the main point, confirmation studies suggest that errors in published results are common. Error-free publications are impossible, but greater care is warranted.

---

* William G. Tomek is professor of agricultural economics at Cornell University. This staff paper formalizes seminar notes; comments are welcome. The empirical examples of replication attempts are derived from several (cited) sources, but I want to especially acknowledge the research of Douglas J. Miller, whose MS thesis was completed at Cornell in 1991. The articles subjected to confirmation analysis are not cited. If interested, readers can consult the confirmation studies for citations of the original articles, but these articles should not be considered as bad examples. Indeed, in Miller's work, articles were selected, in part, for their clarity, and authors of the articles were very cooperative in assisting with replication attempts.

Replication: Meaning and Benefits

In an experiment, the levels of explanatory variables can, in principle, be controlled and/or repeated. Thus, the dependent variable is observed for repeated, independent samples. The estimated coefficients vary from sample to sample, of course, because of sampling error.

When non-experimental data are used, as is common in economics, it is perhaps more accurate to use the terms "confirmation" or "duplication," rather than "replication." The intent is to duplicate or confirm the published result; it is not to replicate an experiment. Thus, one can ask, why confirm prior results? Why duplicate known coefficients? It turns out, however, that most published results are not easily duplicated and often contain errors. Therefore, an important reason for confirmation is precisely to provide the basis for using and improving results. If a model is to be used for important decisions, it is essential that the results be as robust and error-free as possible.

Moreover, if a researcher is trying to improve upon published results, a correct comparison of the new with old requires confirmation of the old results. Comparisons, using a more recent sample or an altered model, are not valid unless the published results are confirmed. Are changes actually the result of a change in model specification, or are they merely a consequence of data revisions (or other errors in variables), or are they the result of misinterpretation of the procedures used in earlier work?

Confirmation studies also should encourage greater care and honesty in publication. Deliberate dishonesty in research probably is uncommon. But, few incentives exist for careful checking of data compilation and input, documentation of procedures (keeping the equivalent of a lab record), and explanation of these procedures. Indeed, pressures exist for rushing to publication. Thus, I suspect that much careless work has been published.

Replication studies are relatively uncommon in empirical econometrics. (One well-known exception is Dewald, Thursby, and Anderson.) Few professional rewards have existed for confirming prior research; replication research is not seen as path-breaking. As outlined above, however, confirmation often is an important initial step in empirical research. Confirmation can provide an indepth understanding of the strengths and weaknesses of prior work. In Hendry and Richard's terms, a researcher should demonstrate that his or her results encompass the work of others. This can only be done if the previous research is confirmed.

It is also true that replication can be interpreted as a lack of trust in the integrity and ability of colleagues (Dewald, Thursby, and Anderson). In a relatively small sub-discipline like the econometric study of agricultural markets, this is a potential problem. We don't like to offend friends or colleagues. But, if confirmation of prior work came to be seen as a routine part of a research program, then those whose work is replicated would not be offended. Indeed, they should be pleased that their research is the basis for additional work by others.

Another problem is that confirmation is difficult and time-consuming. This is the subject of the next section.

Difficulties in Replication

The first, and principal, reason for the inability to duplicate previous work is that the actual data used in the analysis are not available. Secondary data from governmental sources are subject to frequent revision, and authors usually do not keep data files. Moreover, citations to data sources frequently are vague. The foregoing is compounded by two other problems. Errors can be made in data input, and sometimes authors are obscure about how they have used or transformed the data.

In considering the data problem, it is useful to think in terms of four different definitions of a particular concept: (1) the theoretical concept; (2) the true values of an observable variable which is used to measure the concept; (3) the observations actually available at time t on the observable variable; and (4) the data input by the researcher at time t. Usually, the theoretical concept in economics is unobservable, and the true values of the observable variable are unknown. Hopefully the researcher has correctly input the data available at time t (case 3), but data are often revised at subsequent dates. Revisions presumably move one toward the so-called true values of the observable variable. Of course, the researcher may not actually use the data which are available at time t, either because of negligence or because of errors in input (case 4).

Clearly many possible reasons exist for variation in the data set, particularly when the frequency of revisions of time series is considered, and the person attempting the confirmation is not sure about which situation is being faced. If the researcher has not kept the data file, then it is virtually impossible to duplicate the research. Indeed, the original researcher probably cannot duplicate his or her own work, if the data have not been kept (e.g., see Miller).

A second difficulty in replication, compounded by the first, is ambiguity in published results about the actual model fitted, including precisely how data were transformed. The specification of the full model may be unclear, say, because only selected coefficients are published. A problem also may arise when models are fitted subject to restrictions. Confirmation depends on having the precise definition of the restrictions.

A third difficulty in replication relates to differences in computer codes. Generalized Least Squares, for instance, is a generic estimator, and specific feasible GLS procedures can vary from one econometric package to another. In one example in Dewald, Thursby, and Anderson, results for a GLS estimator could not be duplicated even though the same data file was used in the replication attempt. Computer programs also may have errors. For example, in an early version of SAS the Durbin Watson statistic was erroneously computed. Dorfman and McIntosh provide another example of an error in (privately written) software.

The potential problem of computing errors is complicated by large data sets and complex estimators. In my opinion, a researcher should duplicate results using an alternative econometrics package, if their is the slightest concern about possible computational problems.

Illustrations of Problems

It is perhaps useful to illustrate the foregoing points. In replicating one study of beef and pork demand, Miller first attempted to duplicate the results by using only the published article. His independent attempts gave generally similar results to the published results, although one coefficient differed by over 40 percent from the published number. When the author provided the data file, the exact published coefficients were obtained; the nearest independent attempt is compared with the published result in Table 1. However, the reported F test for structural change in beef demand was not exactly duplicated (Table 2).[1]

The model was then fitted using data available in 1991. Since consistent revisions are not available back to the beginning of the sample period, it was necessary to join unrevised to revised data. This was done in two ways. Method (1) merely joined the two series end-to-end; method (2) used simple regression procedures in an attempt to make the unrevised portion of the series more consistent with the revisions (Miller). The results from the two alternatives are reported in Table 1. Two coefficients (for beef quantity and income) are nearly identical under the two procedures, but the others are not.

The large changes in the coefficients of the quantity variables from the old to the revised data, are related to the change in the base of the CPI. The original research used a CPI with 1967 = 1.0; the revised data used 1982-84 = 1.0. Thus, the deflated prices--the dependent variable--are larger after revisions, and the effect of a one unit change in quantity is thereby increased. (Deflating does not affect the income series, because both price and income are divided by the same deflator.)

[1] The model specifications were relatively simple and the article clearly written. The author also cited selected data points, which served as benchmarks. Nonetheless, potential for confusion still existed, including the issue of whether the author shifted to the CPI-U from the CPI-W series when CPI-U became available in 1978. The independently collected income observations were somewhat smaller than those in the author's file for the years 1970-82, and other small differences existed for some data points. The F tests for structural change were verified by computing the sum of squared residuals for the range of possible partitions. As noted in Table 2, the 1950-70, 1971-82 partition for beef did not appear to be the one which minimized the sum of squared errors.

Table 1. Inverse Demand Function, Beef, 1950-82

Variables[a]	Reported & Confirmed	Independent Attempt	Revised data[c] (1)	(2)
Intercept	80.25 (5.33)[b]	79.01 (4.17)	214.5 (3.65)	267.3 (2.86)
QBF	-0.836 (7.00)	-0.814 (6.02)	-2.093 (6.00)	-2.030 (3.95)
QPK	0.165 (0.92)	0.091 (0.40)	0.346 (0.69)	-0.040 (0.05)
QCH	-0.647 (2.57)	-0.753 (2.72)	-3.161 (3.26)	-3.770 (3.41)
INC	0.037 (5.43)	0.040 5.08)	0.041 (4.96)	0.041 (4.038)
R^2	0.716	0.677	0.615	.548
d	0.99	0.99	0.92	1.25

[a] Dependent variable is retail price beef, deflated by CPI, cents per lb.; QBF, QPK, QCH are consumption of beef, pork and broilers, lb. per capita; INC is disposable personal income, delated by CPI, $ per capita. d is Durbin-Watson statistic.

[b] t-ratios given in parentheses

[c] Two ways of using revised data, see text. Both shift CPI base from 1967 = 1.0 to 1982-4 = 1.0.

Source: Miller

Of course, the coefficients also change because of other revisions.[2] If the coefficients of the quantity variables are adjusted by the ratio of the two indexes, then one can obtain an estimate of the effects of other revisions. This ratio is 0.334 in 1967, and the adjusted coefficient of beef quantity for revised data set (2) is -0.678 (-2.030 x 0.334) rather than the -0.836 for unrevised data (Table 1). Thus, revisions in nominal prices and per capita consumption resulted in almost a 20 percent change in the coefficient of beef consumption.[3]

The data revisions result in vastly different conclusions about the timing of structural change (Table 2). It should be noted that all of the fitted equations have autocorrelated residuals, thus casting doubt on the validity of all the F tests.

Another study of meat demands used a quantity dependent specification, quarterly observations, and a spline specification to analyze structural change. The authors did not save the data file, and Miller could not confirm the results even with an author's assistance. Since some ambiguity existed about the exact spline restrictions, Miller tried variants of the restrictions, and selected coefficients for two of these alternatives are reported in Table 3. It is clear that the attempted replications differ from the published results and from each

---

[2] Meat consumption series are revised frequently. Consumption is estimated as disappearance into the marketing system using balance sheet components, such as production and beginning and ending inventories, which are subject to revision. In addition, farm-level data are converted to retail-level weights, and the conversion factors are revised from time to time (e.g., to reflect closer trimming of fat). Moreover, the definition of the population used to compute per capita values has changed; at one time, the military population was excluded; it is now included. Note, the use of a larger population and larger trim of carcass weights has had the effect of reducing the published per capita red meat series.

[3] Empirical results from different data sets are tricky to compare, and alternative ways of making comparison each have potential limitations. My conversion of one coefficient to account for the units-of-measure problem is merely to illustrate a point. The effect of changing the base of the price deflator could be examined by comparing results using both the old and the new deflator; i.e., the results would include the old sample and deflator, the recent sample and old deflator, and the recent sample and deflator. The price index with the old base will, however, not always be available for recent years, and splicing of the two series may be necessary. Another alternative is to standardize the variables for the comparisons, but this makes the economic interpretation of the coefficients more difficult. Still another alternative is to compare elasticities, since they do not depend on the units of measure of the variables. But, for most functional forms, the elasticity must be computed for a particular point on the function, and as the sample changes, the computed elasticity will change because the point at which the elasticity is computed shifts (even if the demand structure has not changed). A rather common error is to assume that changes in elasticities are synonymous with changes in structure; they may or may not be.

Table 2. Estimated Dates of Structural Change in Beef Demand, U.S.

Alternatives	Dates[a]	F-statistic
Reported	1950-70, 71-82[b]	11.12
Confirmation[c]	1950-68, 1969-73, 1974-82	11.57
Revised Data		
(1)	1950-72, 1973-82	18.13
(2)	1950-58, 1959-82	8.96

[a] Regimes determined by partition that minimized total sum of squared errors.

[b] Reported result could not be confirmed in sense that reported partition did not minimize total sum of squared errors. However, alternative partitions give only slightly different results.

[c] Confirmation result based on original author's data file.

Source: Miller

Table 3. Selected Coefficients, Pork Demand, 1960I-1979III

Selected Variables[a]	Reported	Replications[c] (1)	(2)
PPK	-0.123	-0.093	-0.125
	(3.62)[b]	(1.77)	(2.85)
PPK 1	-0.337	-0.100	0.005
	(1.43)	(1.08)	(0.08)
PPK 2	0.763	0.079	-0.015
	(1.50)	(1.20)	(0.26)
PCH	0.078	0.214	0.353
	(1.10)	(2.62)	(3.88)
PCH 1	0.208	-0.974	-7.118
	(0.90)	(2.73)	(3.95)
PCH 2	-0.247	0.813	6.800
	(0.95)	(2.52)	(3.88)
R^2	.74	.76	.80

[a] Dependent variable is consumption pork, lb. per capita; PPK and PCH are retail prices of pork and chicken, respectively, cents per lb., deflated on a 1971 base. PPK1, PPK2, PCH1, PCH2, pertain to prices for spline regimes.

[b] t-ratios in parentheses

[c] Replication (1) uses spline knot locations defined in article. Replication (2) uses knot locations suggested by author of article.

Source: Miller

other. Some coefficients have different signs. The large changes are probably related to the collinearity inherent in spline specifications. Interestingly, the attempted replications have better statistical fits than the reported results.

The logic of the spline restrictions used in the original article is also questionable (see the appendix), and it probably requires the indepth analysis of a replication study to unearth the logic underlying a particular specification. At least, the explanation of the model and its actual specification are inconsistent, and this apparently was not noted by the referees for the paper.

The effects of using a new sample period and an alternative model are illustrated in Table 4. Some systematic changes in coefficients appear to be occurring. The effect of a unit change in beef quantity on beef price seems to be declining (in absolute value), while the effect of a unit change in chicken quantity is growing. However, if one subjects the revised model for the most recent sample period (Model 2, 1958-90) to tests of model adequacy, the model cannot pass all of the tests (Miller). Thus, doubt exists about the validity of the results from the revised data and model, illustrating the difficulty of obtaining useful empirical results.

In 1983, Shonkwiler and Spreen commented on their inability to replicate the results in a published study of farm-level supply and demand for fed beef. The original paper suggested that this market was in disequilibrium. Ferguson discovered that two observations on income were erroneously entered in the data file of the original study. Specifically, the observations for two quarters were approximately 20% too large, and while the income series could not be reconstructed from cited sources, an approximate correction of the two data points provided coefficients which roughly approximate those of Shonkwiler and Spreen (Table 5). If the erroneous data points are used, the original results can be exactly duplicated. Clearly, two data-entry errors for one variable had a large impact on the coefficients. (Ferguson further pointed out that the residuals in the demand equation are highly autocorrelated for both data sets, and it should be noted that these results were obtained only with the cooperation of an author.)

These examples, plus others, suggest the following. First, published results usually can not be duplicated without the assistance of the original author, and independent attempts to confirm prior estimates often result in large differences in magnitudes of coefficients. Second, data revisions constitute an important issue in model specification; put another way, initial observations often can be viewed as seriously in error relative to subsequent revisions. Conclusions may be changed by revisions; e.g., whether or not structural change occurred and at what point in time may depend on the magnitude of errors in variables (or errors in specification).

Third, at least 75 percent of published empirical studies contain errors of varying kinds. (This is a subjective estimate based partly on other published studies, e.g., Dewald, Thursby and Anderson). Fourth, perhaps half of these errors are serious in the sense that, if corrected, they change the conclusions

Table 4. Inverse Demand Functions, Beef, Alternate Sample Periods and Models, Revised Data

Variables[a]	Model 1 1950-82	Model 1 1958-90	Model 2 1958-90
Intercept	267.3 (2.86)[b]	176.3 (2.14)	79.87 (1.12)
QBF	-2.030 (3.95)	-1.649 (2.81)	-1.550 (3.23)
QPK	-0.040 (0.05)	0.797 (1.11)	0.605 (1.03)
QCH	-3.770 (3.41)	-7.138 (5.43)	-5.339 (4.58)
INC	0.040 (4.38)	0.052 (4.40)	0.042 (4.15)
P lagged	-	-	0.481 (3.89)
R^2	.548	.686	.799
d	1.25	1.11	-
h	-	-	.212[c]

[a] See Table 1

[b] t-ratios in parentheses

[c] Durbin's h-statistic

Source: Miller

Table 5. Fed Beef Demand, Farm Level, U.S., 1965-79

Variables[a]	Reported & Confirmed	Replications[c]	
		S - S	F
Intercept	588.0	-4708	-3888
	(0.8)[b]	(4.70)	(4.99)
Price fed beef	-72.01	-157.8	-132.1
	(1.78)	(4.50)	(4.26)
Price utility cows	89.75	147.8	119.8
	(2.32)	(4.65)	(4.18)
Price hogs	-9.15	-16.94	-12.84
	(0.64)	(1.50)	(1.24)
Real income	172.2	455.8	344.8
	(6.38)	(10.16)	(11.60)
Durbin-Watson	-d	-	0.66

[a] Dependent variable is marketings of fed cattle.

[b] Coefficients in parentheses are ratios of coefficients to standard deviations; all equations estimated by two-stage least squares.

[c] Independent replications: S-S = Shonkwiler and Spreen; F = Ferguson. S-S also compare alternative specifications and point out the residuals are autocorrelated. F confirmed original result with authors' data file; her replication uses corrected income series (see text), but otherwise retains original data file.

[d] Not reported

of the publication. Thus, when these problems are combined, one can hardly be sanguine about the usefulness of empirical research in economics.

Some Concluding Remarks

Obtaining high quality, useful empirical results is difficult, probably far more difficult than commonly understood by most applied economists. Errors in variables and in specifications are compounded by the carelessness of analysts. Confirmation of published research can help ameliorate these problems.

Confirmation requires that original data files and precise definitions of model specification and restrictions be available. As others have pointed out, professional journals must play a role in assuring that this information is available to those wishing to confirm results.

Researchers also have an obligation to demonstrate that their results improve upon previous work and that their results can "pass" tests of model adequacy (e.g., see Godfrey). These tests, it should be noted, must go well beyond conventional appeals to R^2 and the Durbin-Watson statistic. Tests of model adequacy cannot tell the researcher what precisely is wrong, but can help alert the researcher to the fact that a problem exists.

The problem of errors in variables and data revisions is even less tractable, but at a minimum researchers should determine whether influential observations and collinearity are important. Computer software is available for such analyses, and it should be used. Analysts also must accept the possibility that secondary, time-series data cannot answer some of the research questions which we would like to ask. In such situations, high quality answers require original data.

In sum, high quality empirical results require high quality inputs. These inputs include the model specification, data, and their management by the analyst. In my view, confirmation of prior results can improve the depth and quality of the knowledge of the analyst, and thereby contribute to the quality of new results.

Appendix
Logic of Spline Restrictions

An unrestricted linear model allows the slope and intercept parameters to change at a point by the use of zero-one variables and the interaction of those variables with other regressors. The simplest example is:

$$Y_t = a_0 + a_1D_t + b_0X_t + b_1(D_tX_t) + e_t$$

where $D_t = 1$ for regime 1 (say, $X_t > X_n$)
$= 0$ for regime 2.

The spline restriction joins the two regimes at a particular point, say, "n." Thus, the restriction is:

$$a_0 + b_0X_n = (a_0 + a_1) + (b_0 + b_1)X_n,$$
$$\text{or } a_1 = -b_1X_n.$$

Clearly, a key issue is the selection of the point "n." In a paper analyzed by Miller, the authors write as if the logic of the restrictions depends on a change in structure at particular time periods. But, their restrictions use price levels (X_n's) for those points in time. Defining the restrictions in this way implies that the regimes depend on price levels (say, a hypothesis that consumption meets a resistance level at a particular price, X_n), and since prices can move above and below the particular level with the passage of time (and indeed did so in the sample), the logic of the restriction is not based on a point in time. In other words, if prices move higher than the specified level, the regime changes, and this can, and does, occur at more than one point in time.

In the article under analysis, three regimes (two knots) were specified for five price variables. Since the five prices varied among the various regimes through time, the actual parameters applicable to any point in time are a complex mixture of regimes.

References

Dewald, W. G., J. G. Thursby, and R. G. Anderson. "Replication in Empirical Econometrics: The Journal of Money, Credit and Banking Project," American Economic Review 76(1986): 587-603.

Dorfman, J. H., and C. S. McIntosh. "Results of a Price-Forecasting Competition: Reply," American Journal of Agricultural Economics 73(1991): 1277-1278.

Ferguson, C. A. An Evaluation of a Disequilibrium Model. Cornell Univ. A. E. Res. 83-27, August 1983.

Godfrey, L. G. Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and Other Approaches. Cambridge University Press, 1988.

Hendry, D. F., and J. F. Richard. "On the Formulation of Empirical Models in Dynamic Econometrics," Journal of Econometrics 20(1982): 3-33.

Leamer, E. E. "Let's Take the Con out of Econometrics," American Economic Review 73(1983): 31-43.

Miller, D. J. Assessing Studies of Structural Change in Meat Demand. MS thesis, Cornell University, August 1991.

Shonkwiler, J. S., and T. H. Spreen. "Disequilibrium Market Analysis: An Application to the U.S. Fed Beef Sector: Comment," American Journal of Agricultural Economics 65(1983): 360-363.

Other Agricultural Economics Staff Papers

No. 91-15	Time-of-Use Pricing for Electric Power: Implications for the New York Dairy Sector--A Preliminary Analysis	Mark C. Middagh
No. 91-16	The Causes of Economic Inefficiencies in New York Dairy Farms	Arthur C. Thomas Loren W. Tauer
No. 91-17	Role of the Non-Profit Private Sector in Rural Land Conservation: Results from a Survey in the Northeastern United States	Nelson L. Bills Stephen Weir
No. 91-18	Concept Maps: A Tool for Teachers and Learners	Deborah H. Streeter
No. 91-19	What Can Be Learned from Calculating Value-Added?	B. F. Stanton
No. 91-20	Northeast Dairy Cooperative Financial Performance 1984-1990	Brian Henehan Bruce Anderson
No. 91-21	Urban Agriculture in the United States	Nelson Bills
No. 91-22	Effects of Housing Costs and Home Sales on Local Government Revenues and Services	David J. Allee
No. 91-23	Current Outlook for Dairy Farming, Dairy Products, and Agricultural Policy in the United States	Andrew Novakovic Nelson L. Bills Kevin Jack
No. 91-24	Government Influence on the Supply of Commercial Inventories of American Cheese	James Pratt Andrew Novakovic
No. 91-25	Role of the Non-Profit Private Sector in Rural Land Conservation: Results from a Survey in the Northeastern United States (Revised)	Nelson L. Bills Stephen Weir