Board of Governors of the Federal Reserve System

International Finance Discussion Papers

Number 663

April 2000

# CONSTRUCTIVE DATA MINING: MODELING CONSUMERS' EXPENDITURE IN VENEZUELA

Julia Campos and Neil R. Ericsson

NOTE: International Finance Discussion Papers are preliminary materials circulated to stimulate discussion and critical comment. References to International Finance Discussion Papers (other than an acknowledgment that the writer has had access to unpublished material) should be cleared with the author or authors. Recent IFDPs are available on the Web at www.bog.frb.fed.us.

# CONSTRUCTIVE DATA MINING: MODELING CONSUMERS' EXPENDITURE IN VENEZUELA

Julia Campos and Neil R. Ericsson*

*Abstract*: Hoover and Perez (1999) advocate a constructive approach to data mining. The current paper identifies four pejorative senses of data mining and shows how Hoover and Perez's approach counters each. To assess the benefits of constructive data mining, the current paper applies a data-mining algorithm similar to Hoover and Perez's to a dataset for Venezuelan consumers' expenditure. The selected model is economically sensible and statistically satisfactory; and it illustrates how data can be highly informative, even with relatively few observations. Limitations to algorithmically based data mining provide opportunities for the researcher to contribute value added in the empirical analysis.

*Keywords*: Dynamics, Encompassing, General-to-specific modeling, Hoover and Perez (1999), Model design, PcGets.

*JEL classifications*: C5, E21.

*Forthcoming in the *Econometrics Journal* (1999) 2, 2. The *Econometrics Journal* is published by the Royal Economic Society both electronically on the WorldWide Web at www.blackwellpublishers.co.uk/ectj/ and in printed format. The first author is a professor of econometrics in the Departamento de Economía e Historia Económica, Facultad de Economía y Empresa, Universidad de Salamanca, Salamanca 37008 España (Spain). The second author is a staff economist in the Division of International Finance, Board of Governors of the Federal Reserve System, Washington, D.C. 20551 U.S.A. The authors may be reached on the Internet at jcampos@gugu.usal.es and ericsson@frb.gov respectively. The views in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System, the Banco Central de Venezuela, or of any other person associated with the Federal Reserve System or the Banco Central de Venezuela. The first author undertook some of the research described herein while employed at the Banco Central de Venezuela. The authors are indebted to Angus Deaton, Juan Dolado, David Hendry, Jeroen Kremers, Jaime Marquez, Adrian Pagan, Neil Shephard, David Wilcox, and an anonymous referee for helpful comments; to Heidi Lyss and Hayden Smith for research assistance; and to David Hendry and Hans-Martin Krolzig for providing us with a copy of PcGets Version 0.99d; see Hendry and Krolzig (1999). All reported regressions were obtained using PcGive Professional Version 9.2: see Doornik and Hendry (1996).

# 1 Introduction

In economics, data mining is commonly viewed as a necessary evil. From this perspective, data mining is needed to construct reasonable empirical models, but the nominal critical levels for test statistics are affected. Hoover and Perez (1999) challenge that view, showing in a mixed empirical-Monte Carlo framework that a certain form of data mining ("general-to-specific modeling" à la LSE methodology) does not appear to suffer from this problem in practice.

Hoover and Perez (1999) extend Lovell's (1983) study of data mining, applying his framework to typical quarterly U.S. macroeconomic data. From that data and a set of pseudo-random errors, Hoover and Perez (1999) generate the dependent variable for several different model specifications. A mechanistic version of general-to-specific modeling is applied to the dependent variable and a much larger set of regressors (but a set including the correct regressors), and the simplified models are examined. Very frequently, the simplified model is either the correct specification or is close to it on a reasonable metric. Furthermore, test statistics such as $t$-ratios in the simplified model are well-behaved.

The results in Hoover and Perez (1999) are surprising and controversial, and they have broad ramifications for empirical modeling in the profession. The current paper complements the work in Hoover and Perez (1999) by clarifying what is meant by data mining, and by applying a variant of their data-mining algorithm empirically — to a dataset on consumers' expenditure in Venezuela. The empirical application provides some practical experience with such algorithms.

This paper is organized as follows. Section 2 distinguishes between four distinct pejorative senses of data mining in the literature and shows how Hoover and Perez's modified general-to-specific modeling strategy can counter each of these senses in practice. Data mining, in each of its pejorative senses, is empirically detectable. Section 3 models consumers' expenditure on non-durables and services in Venezuela over 1970–1985, applying the variant of Hoover and Perez's simplification algorithm implemented in the computer program PcGets; see Hendry and Krolzig (1999). In the selected model, income, liquidity, and inflation determine expenditure in an economically sensible fashion; and that model is robust and has constant, well-determined parameter estimates. This section shows that, even with relatively few observations, high information content in the data can help counter claims of pejorative data mining. It also identifies two limitations to algorithmically based data mining — the initial general model, and data transformations — and demonstrates how they are opportunities for the researcher *qua* economist to contribute value added to the empirical analysis. Section 4 concludes, offering some prospective thoughts for algorithmic constructive data mining. The Appendix provides details on the data.

**Table 1.** Pejorative data mining: four senses, four refutations.

| Sense of data mining<br>Operational procedure and implications | Counter-evidence<br>Operational procedure and refutations |
|---|---|
| 1a. *Repeated testing*<br>Select regressors to maximize $t$-ratios.<br>Empirically, $t$-ratios decline in magnitude and $\hat{\sigma}$ increases as $T$ increases. | 1b. *Recursive estimation, additional data*<br>Use larger critical values.<br>Empirically, $t$-ratios increase in magnitude and $\hat{\sigma}$ is constant as $T$ increases. |
| 2a. *Data interdependence*<br>The dependent variable $y$ is related to $x$; and $\text{corr}(x, z) \neq 0$, implying that $y$ and $z$ are correlated, even while they have no fundamental relationship. | 2b. *Super exogeneity, encompassing*<br>Show empirically that the parameters relating $y$ to $z$ are constant while $\text{corr}(x, z)$ changes over time.<br>Encompass the rival model. |
| 3a. *Corroboration*<br>The regressors are chosen for "sensible" coefficient estimates.<br>There may be omitted variables. | 3b. *General-to-specific modeling, encompassing*<br>Adopt general-to-specific modeling.<br>Demonstrate that the selected model has innovation errors and encompasses other models. |
| 4a. *Over-parameterization*<br>The model is "over-fitted".<br>Very few degrees of freedom remain. | 4b. *High informational content of the data*<br>Adopt general-to-specific modeling.<br>Show that the data have high information content. |

## 2 Pejorative and Constructive Data Mining

This section delineates four senses of pejorative data mining. Each sense may be either confirmed or refuted empirically, and the modified general-to-specific approach of Hoover and Perez embodies techniques for doing so. This section then turns to constructive data mining, which seeks to design an empirical model congruent with the data. For excellent discussions on both senses of data mining, see Leamer (1978) and Hendry (1995, pp. 544–546) *inter alia*.

One common concern with the general-to-specific modeling strategy is its potential to mine the data pejoratively. Data mining might be a problem in any of four distinct senses. For each sense, there is a potential for refuting that allegation of data mining. Table 1 lists each sense in terms of its operational implementation, its empirical consequences, and the techniques and counter-evidence for refutation. Each sense of

data mining is now discussed in turn. For ease of discussion, the dependent variable is denoted $y$, the estimated equation standard error $\hat{\sigma}$, and the sample size $T$.

1. *Repeated testing.* In this sense of data mining, regressors are selected in an attempt to maximize $t$-ratios. This form of data mining has implications for how the $t$-ratios on spuriously selected regressors ought to behave as the sample period is extended after having selected those regressors. Specifically, $t$-ratios on spuriously selected regressors ought to become smaller in absolute value as $T$ increases. Likewise, because the initial model is over-fitted by having selected those spurious regressors, $\hat{\sigma}$ ought to increase as $T$ increases.

   One solution is to use larger critical values, as discussed in Sargan (1981), Lovell (1983), and Denton (1985) *inter alia*. Recursive estimation and additional data also provide mechanisms for confirming or refuting such data mining. Empirically, recursive $t$-ratios might drift away from zero, and recursive estimates of $\hat{\sigma}$ might be statistically and numerically relatively constant, thus refuting this sense of data mining.

2. *Data interdependence.* A second form of pejorative data mining might arise from data interdependence. Suppose that $y$ depends upon some variable $x$, that $x$ is correlated with some other variable $z$, and that $y$ is modeled as a function of $z$. The recursive $t$-ratios on $z$ typically will increase with $T$, due to the correlation between $x$ and $z$; but $z$ is simply the wrong variable.

   Super exogeneity of $z$ and encompassing refute this form of data mining. On the former, if $y$ depends on $x$ alone, and if the correlation between $x$ and $z$ changes over time due to regime changes in the system generating $\{y, x, z\}$, then the coefficient on $z$ in the model of $y$ ought to be nonconstant over time. An empirically constant coefficient on $z$ in the presence of the changing data correlations implies super exogeneity of $z$ and contradicts data mining due to data interdependence. Encompassing tests can also counter this sense of data mining.

3. *Corroboration.* A third sense of data mining might be called "corroboration". The regressors $z$ are chosen according to a criterion such as having sensible coefficient estimates. However, there may still be important omitted variables.

   General-to-specific modeling helps address this concern by starting with a general model including many variables — specifically, including the potential omitted variables — and demonstrating that the potential omitted variables do not matter in the final selected model. Encompassing tests of alternative models provide an additional tool for countering the claim of data mining as corroboration.

4. *Over-parameterization.* A fourth sense of data mining is "over-parameterization", in which the model is over-fitted, thereby using up many degrees of freedom. Whether or not such over-parameterization matters depends in fair part on the informational content of the data, as Section 3.3 below discusses in greater detail.

Data mining can have a positive and constructive sense, in which an empirical model is built to satisfy a range of economic and statistical criteria; see Hendry (1995, pp. 67ff, 361ff, 544ff) and the references cited therein. Diagnostic test statistics are interpreted as design criteria, rather than as formal classical test statistics. In principle, the data generation process (DGP) as a model satisfies the design criteria, so the modeling process aims to create a model mimicking properties of the DGP. Under certain conditions, this approach implies selection of the DGP (or a model isomorphic to it) as the final model in large samples; see White (1990, especially p. 381). Put somewhat differently, constructive data mining aims to separate all models into two sets: those that satisfy the design criteria and those that do not. The focus is on the first set because the DGP lies in it; and this partitioning of models markedly shrinks the set of models under serious consideration, thus aiding model design generally. Furthermore, entire classes of models may be incompatible with observed data properties; see Ericsson and Hendry (1999). Using a dataset on consumers' expenditure in Venezuela, the next section empirically demonstrates the benefits and limitations of general-to-specific modeling *qua* constructive data mining.

# 3 Empirical Modeling of Consumers' Expenditure in Venezuela

This section models consumers' expenditure on non-durables and services in Venezuela over 1970–1985, applying the variant of Hoover and Perez's simplification algorithm implemented in the computer program PcGets; see Hendry and Krolzig (1999). This dataset was initially studied in Campos (1984) and Campos and Ericsson (1988); and Ericsson, Campos, and Tran (1990, Section 4.D) summarize those papers' results, including the final model and tests of super exogeneity. The dataset is of particular interest in the context of data mining because there are very few observations for estimation ($T = 16$), yet each observation is very informative. Furthermore, these two characteristics typify datasets for many other developing and emerging-market economies.

Section 3.1 reviews the underlying economic theory and discusses the data. Section 3.2 obtains the final model from a general specification, and Section 3.3 documents the final model's empirical properties in the context of data mining. Section 3.4 considers empirical ramifications of two important choices by the modeler: the initial general model, and data transformations. Capital letters denote both the generic

name and the level of a variable, logarithms are in lowercase, $\Delta$ is the difference operator, $t$ is the time subscript, and OLS standard errors are in parentheses $(\cdot)$. Figures appear as panels of graphs, with each graph designated by a suffix $a, b, c, \ldots$, row by row.

## 3.1 Economic Theory and the Data

Consumers are hypothesized to keep expenditure ($C$) and assets ($W$, for "wealth") proportional to income ($I$) in the long run, as motivated by the permanent income and life-cycle hypotheses *inter alia*. In logarithms, the resulting long-run solution is:

$$c - i = k + \phi \cdot (w - i), \tag{1}$$

for coefficients $k$ and $\phi$. A solved dynamic form of (1) is an error correction model (ECM) for consumers' expenditure, with feedback from both expenditure relative to income $(c - i)$ and assets relative to income $(w - i)$. An unrestricted version of such an ECM is the starting point for empirical modeling in Section 3.2.

The data are annual values over 1968–1985 of Venezuelan consumers' expenditure on non-durables and services ($C$), national disposable income ($Y$), end-of-year $M_2$ ($W$), and the end-of-year price index for all consumers' expenditure ($P$, 1968 = 1.0, for Caracas). The series $C$, $Y$, and $W$ are real 1968 Bolívares per capita. To account for the sometimes substantial "inflation tax" from holding liquid assets, an adjusted income series $I_t$ is derived from $Y_t^n - (\Delta p_t) W_{t-1}^n$, where a superscript $n$ denotes the nominal total value; cf. Hendry and von Ungern-Sternberg (1981). See Campos and Ericsson (1988, Appendices A and B) and the Appendix below for details on data construction, sources, and caveats.

Consider some basic properties of the data themselves. Figure 1*a* plots the logs of expenditure and income, and Figure 1*b* their growth rates. Three distinct episodes are evident. Before 1974, both series grew at relatively moderate rates. Because of dramatically increased petroleum revenues, income increased by over 35% in 1974 and, through 1981, remained on a plateau at 155%–170% of the level of 1968 income. From 1974, expenditure grew rapidly (but less so than income), leveling off in the late 1970s, with the expenditure-to-income ratio in 1981 being virtually the same as in 1968. From 1981 to 1985, real per capita income plummeted at 7% per annum, but expenditure remained relatively constant. One possible explanation for the differing responses of expenditure to income is the change in liquidity. Figure 1*c* graphs the log of liquidity, and Figure 1*d* the logs of the expenditure-income and liquidity-income ratios. Those ratios rose by 27% and 135% over the period. The substantial increase of the latter ratio could account for the constancy of expenditure in the 1980s, even while income fell. Both ratios fell markedly in 1974 and rose in 1982–1983: large changes in income were primarily responsible, rather than changes in expenditure or liquidity; cf. Figure 1*a*. Figure 1*e* plots the implied disequilibrium from (1), assuming

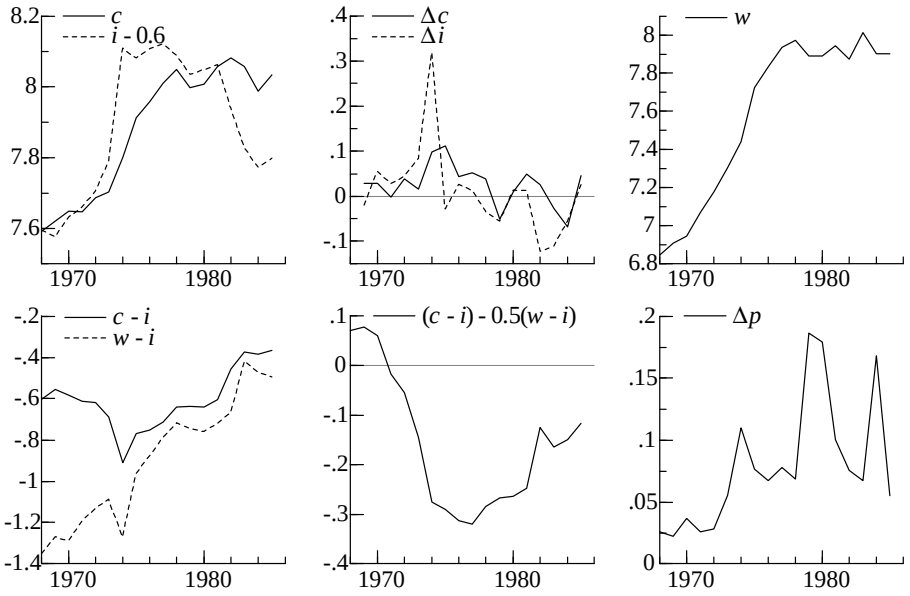


Figure 1: Expenditure and income, their growth rates, liquidity, expenditure and liquidity relative to income, a potential disequilibrium measure $(c-i)_t - \frac{1}{2}(w-i)_t$, and inflation.

that $\phi = 0.5$; and Figure $1f$ plots inflation ($\Delta p$). With the second oil-price shock in 1979, Venezuelan inflation jumped to nearly 20%, falling during the world-wide recession in the early 1980s, and sharply increasing again in 1984. The higher inflation rates during the second half of the sample, combined with greater liquidity, imply substantial discrepancies between $Y$ and $I$. In summary, even while the dataset has relatively few observations, the data movements are large relative to those typical for industrialized countries. These movements will be central to Section 3.3's analysis of the data's information content.

## 3.2 General-to-specific Modeling

This subsection develops a conditional ECM for Venezuelan consumers' expenditure. Initially, a non-algorithmic simplification is applied to the general model. Then, that simplification is shown to result from an algorithmic general-to-specific simplification using PcGets.

In order to establish a baseline innovation variance, consider the following general autoregressive distributed lag relationship for consumers' expenditure, conditional upon liquidity, income, and prices.

$$c_t = \sum_{j=1}^{2}(a_{1j}c_{t-j} + a_{2j}w_{t-j}) + \sum_{j=0}^{2}(a_{3j}i_{t-j} + a_{4j}p_{t-j}) + a_5 + a_6 D_t + u_t, \qquad (2)$$

**Table 2.** An unrestricted ECM of consumers' expenditure in Venezuela.

| lag $j$ | Variable | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $\Delta c_{t-j}$ | $\Delta i_{t-j}$ | $\Delta p_{t-j}$ | $\Delta w_{t-j}$ | $(c-i)_{t-j}$ | $(w-i)_{t-j}$ | $D_{t-j}$ | Constant |
| 0 | $-1$ | 0.224 | $-0.553$ | – | – | – | 0.0260 | 0.006 |
| | ( – ) | (0.045) | (0.086) | | | | (0.0087) | (0.033) |
| 1 | $-0.081$ | 0.211 | 0.260 | 0.004 | $-0.169$ | 0.070 | – | – |
| | (0.163) | (0.056) | (0.118) | (0.057) | (0.062) | (0.027) | | |

$T = 16$ [1970–1985] $\quad R^2 = 0.97 \quad \hat{\sigma} = 1.1993\% \quad dw = 2.59 \quad AR : F(1,5) = 1.13$
$ARCH : F(1,4) = 0.12 \quad Normality : \chi^2(2) = 2.14 \quad RESET : F(1,5) = 0.52$

where $a_{1j}$, $a_{2j}$, $a_{3j}$, $a_{4j}$, $a_5$, and $a_6$ are unknown coefficients; $D_t$ is a +1/–1 dummy for 1970–1971 to account for apparent measurement errors in consumers' expenditure for those years (see Campos and Ericsson (1988, Appendix B)); and $u_t$ is the residual. *With* loss of generality, we assume that $c$ is long-run homogeneous in $i$ and $w$ (i.e., $1 - \sum a_{1j} = \sum(a_{2j} + a_{3j})$, as implied by (1)) and that prices enter only as inflation (i.e., $\sum a_{4j} = 0$). Even with these two restrictions, the equation — to be estimated on only 16 observations — has 10 unrestricted coefficients.

$$\begin{aligned} \Delta c_t &= b_1 \Delta c_{t-1} + \sum_{j=0}^{1} (b_{2j} \Delta i_{t-j} + b_{3j} \Delta p_{t-j}) + b_4 \Delta w_{t-1} \\ &\quad + b_5 (c-i)_{t-1} + b_6 (w-i)_{t-1} + b_7 + b_8 D_t + u_t, \end{aligned} \tag{3}$$

where $b_1$, $b_{2j}$, $b_{3j}$, $b_4$, $b_5$, $b_6$, $b_7$, and $b_8$ are transformations of the coefficients in (2). Table 2 lists the least squares coefficient estimates and standard errors for (3), and the diagnostic statistics available for such an over-parameterized model. The residual standard error is slightly above 1%: any new model will require a similar or smaller equation standard error as a necessary condition for encompassing the model in Table 2.

Four economically sensible and statistically acceptable parametric restrictions are apparent for the model in Table 2. First, the lagged rates of change of expenditure and of liquidity are insignificant: a single lag on each log-level is sufficient to capture those aspects of dynamics.

Second, the coefficients on the current and lagged growth rates of income are nearly equal: a restriction of equality can be interpreted as a statistical smoothing of income in order to extract changes that are more permanent. Changes in the growth rate of income also immediately affect the budget constraint and liquidity, so giving the coefficient on current income (or its growth rate) alternative interpretations.

Third, the coefficient on current inflation is approximately twice the magnitude of that on lagged inflation, and opposite in sign. Imposing that restriction implies the term $\Delta p_t + \Delta^2 p_t$, which is a data-based predictor of next period's inflation, optimal if

prices vary quadratically. Economically, $\Delta p_t + \Delta^2 p_t$ might capture a smaller desired $k$ in (1) in the face of higher inflation, or (e.g.) the aim of consumers to save *more* now in anticipation of higher inflation later so as to be able to consume more closely the same amount in real terms when that higher inflation arrives. See Flemming (1976, Chapter 7) for a formal justification of data-based predictors like $\Delta p_t + \Delta^2 p_t$, and Cochrane (1989) for implications of such rule-of-thumb behavior. Additional complementary justifications for $\Delta p_t + \Delta^2 p_t$ turn on "saving" for future consumption through current purchases of consumer *durables*, adjustments to liquidity through saving, difficulties in distinguishing between relative and aggregate changes in prices, differences between short- and long-run income elasticities, and the adequacy of the measure $I$ in capturing the inflation tax.

Fourth, the implied estimate of the long-run elasticity $\phi$ in equation (1) is 0.41 (= 0.070/0.169), very close to Hendry and von Ungern-Sternberg's estimate of 0.44 for the United Kingdom. These estimates are also close to one half, which would imply a long-run solution of $c = k + (i + w)/2$, in which income and wealth have equal effects on expenditure. Because of that simple solution, and in an effort to design as parsimonious a model as possible, we consider the restriction that $\phi = 0.5$.

Re-estimating (3) with those four restrictions imposed obtains (4).

$$\begin{aligned} \Delta c_t \;=\; & \underset{(0.0042)}{0.0193} \;+\; \underset{(0.030)}{0.457}\, \Delta_2 i_t/2 \;-\; \underset{(0.026)}{0.270}\, (\Delta p_t + \Delta^2 p_t) \\ & -\; \underset{(0.019)}{0.142}\, [(c-i) - \tfrac{1}{2}(w-i)]_{t-1} \;+\; \underset{(0.0065)}{0.0263}\, D_t \end{aligned} \tag{4}$$

$$\begin{aligned} & T = 16\ [1970\text{–}1985] \quad R^2 = 0.97 \quad \hat{\sigma} = 0.9160\% \quad dw = 3.05 \\ & AR : F(2,9) = 3.48 \quad ARCH : F(1,9) = 0.61 \quad Normality : \chi^2(2) = 1.78 \\ & RESET : F(1,10) = 0.00 \quad Hetero : F(8,2) = 0.33 \end{aligned}$$

The long-run elasticities of expenditure with respect to both income and liquidity are $\frac{1}{2}$, noting that the error correction term can be rewritten as $[c - \frac{1}{2}(i + w)]_{t-1}$. Short-run (within-year) elasticities are 0.23 for income, zero for liquidity, and $-0.54$ for inflation; and adjustment to disequilibrium is 14% per year. Remarkably, estimates in (4) involving dynamics are close to those obtained by Hendry and von Ungern-Sternberg (1981) with quarterly data for the United Kingdom. Comparable estimates to those for the growth rate of income and the error correction term are 0.50 and $-0.16$ (versus 0.46 and $-0.14$ in (4)). Even so, the time series and data moments of the two countries differ markedly, highlighting how relationships can be similar, even while data properties differ.

Figures 2*a*–2*f* plot actual and fitted values of $\Delta c_t$ from (4), their cross-plot, the residuals from (4), actual and fitted values of $c_t$ from (4), their cross-plot, and the histogram and estimated density of the residuals from (4). These graphs show how well (4) explains the data, and that the residuals are visually serially uncorrelated and approximately normally distributed. From the reported diagnostic statistics, Table 2

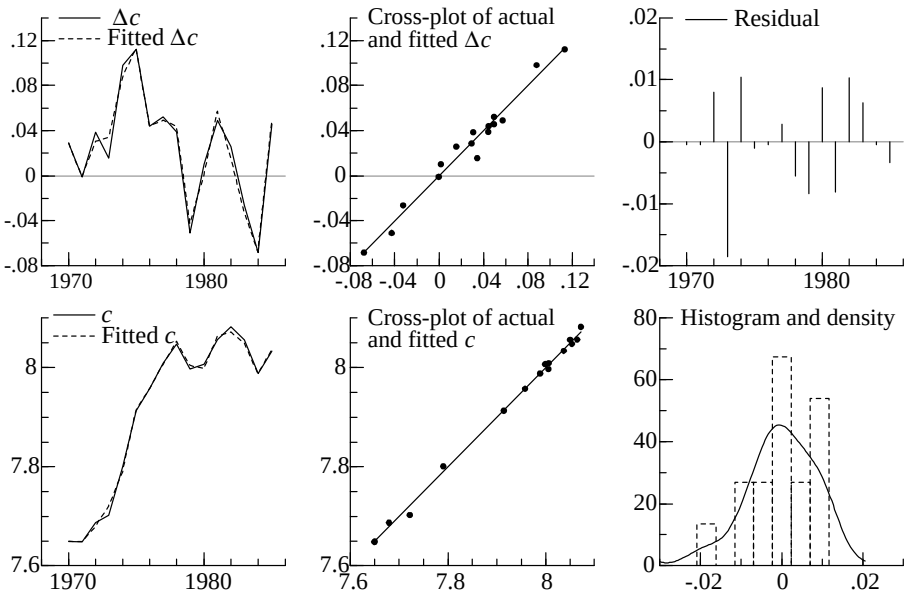


Figure 2: Actual and fitted values of $\Delta c_t$ from (4), their cross-plot, the residuals from (4), actual and fitted values of $c_t$ from (4), their cross-plot, and the histogram and estimated density of the residuals from (4).

and (4) appear well-specified; see Doornik and Hendry (1996) for definitions of and references for those statistics. Equation (4) also is an adequate simplification of (3), i.e., the model in Table 2: the corresponding $F$-statistic is $F(5,6) = 0.08$, with a $p$-value of 0.99.

That said, *algorithmic* simplification from (3) to (4) is still of interest: (4) involves nonzero restrictions on the coefficients in (3), and statistics for intermediate models might reject. To address both issues, the regressors in (3) are transformed without loss of generality to those in the following equation, which explicitly includes the regressors from the final model (4).

$$\begin{aligned}\Delta c_t = {} & d_1 \Delta c_{t-1} + d_2(\Delta_2 i_t/2) + d_3 \Delta i_{t-1} + d_4(\Delta p_t + \Delta^2 p_t) + d_5 \Delta p_{t-1} + d_6 \Delta w_{t-1} \\ & + d_7[(c-i) - \tfrac{1}{2}(w-i)]_{t-1} + d_8(w-i)_{t-1} + d_9 + d_{10} D_t + u_t, \end{aligned} \tag{5}$$

where $d_1, \ldots, d_{10}$ are transformations of the coefficients in (3). Applying the variant of Hoover and Perez's simplification algorithm implemented in PcGets, (5) simplifies to (4).[1] *Inter alia*, that implies that intermediate models in simplification paths from (5) to (4) also appear well-specified. Section 3.3 documents additional empirical

[1]For all empirical results herein, simplifications in PcGets use the following settings. The model selection criterion is the Schwarz criterion. Significance levels are 1% ($t$- and $F$-tests), 2.5% (split-sample tests), 1% (diagnostics (high)), and 0.5% (diagnostics (low)). No $F$ pre-testing is used. In the split-sample analysis, the subsample is 75%; and the penalty is 25% for a failed $t$-test in the

properties of (4) in the context of data mining, and Section 3.4 re-examines the issue of data transformations.

## 3.3 Data Mining, and Properties of the Final Model

This subsection analyzes the final model (4) in terms of the four pejorative senses of data mining discussed in Section 2. The data provide evidence countering each of those senses of data mining.

1. *Repeated testing.* Figure 3 plots the recursive estimates for the coefficients on the constant term, $\Delta_2 i_t/2$, $\Delta p_t + \Delta^2 p_t$, and $[(c-i) - \frac{1}{2}(w-i)]_{t-1}$ (denoted $ecm_1$ in the graphs); their respective $t$-ratios; and the recursive residual sum of squares, one-step residuals, one-step Chow statistics, and breakpoint Chow statistics. For all coefficients, the recursive $t$-ratios (in the second row of graphs: Figures 3*e*–3*h*) increase in absolute value as the sample size increases, countering this sense of data mining. Even with only 10 observations, the $t$-ratios on $\Delta_2 i_t/2$, $\Delta p_t + \Delta^2 p_t$, and $[(c-i) - \frac{1}{2}(w-i)]_{t-1}$ are all greater than six in magnitude, reflecting high information content in the data. Figure 3*j* plots the one-step residuals and $0 \pm 2\hat{\sigma}_t$, where $\hat{\sigma}_t$ is the recursive estimate of $\sigma$. The standard error $\hat{\sigma}_t$ is constant or declines slightly over time, rather than increases.

2. *Data interdependence.* The recursive estimates in Figures 3*a*–3*d* and the Chow statistics in Figures 3*k* and 3*l* all point to the empirical constancy of (4). The data graphed in Figure 1 document large changes in the Venezuelan economy, and Campos and Ericsson (1988) and Ericsson, Campos, and Tran (1990, Section 4.D) show that marginal equations for income and inflation are nonconstant, implying super exogeneity of those variables for the parameters in (4).

   Figure 4 presents the corresponding "backward" recursive estimates, $t$-ratios, residual sum of squares, one-step residuals, and Chow statistics. These are of particular interest, as income, expenditure, and inflation all increase markedly in 1974. Forward recursive estimation must always include 1974 in the estimation period, whereas backward recursive estimation need not. The backward recursive estimates also support the constancy of (4); and the corresponding recursive $t$-ratios generally increase as the estimation period is extended backwards, countering the first sense of data mining.

3. *Corroboration.* Equation (4) is a statistically satisfactory simplification of (3), as Section 3.2 showed, thus countering the third sense of data mining.

full sample or in either subsample. The diagnostic statistics test for constancy (i.e., with the Chow statistic for each of the two subsamples — the first 75% and the last 75%), normality, residual autocorrelation, and heteroscedasticity.

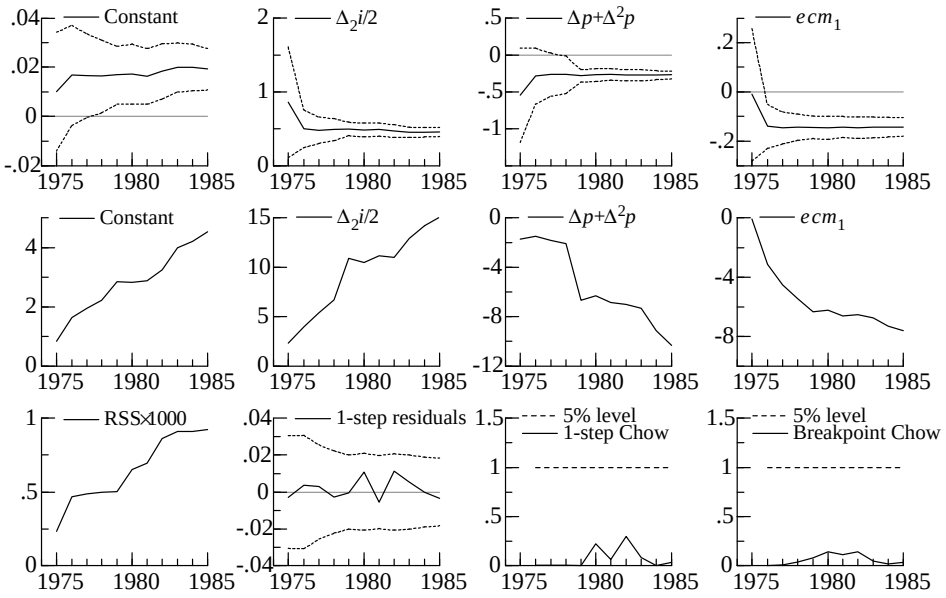


Figure 3: Recursive calculations: four estimates and $\pm 2$ estimated standard errors, four $t$-ratios, the residual sum of squares (RSS), one-step residuals and $0 \pm 2\hat{\sigma}_t$, one-step Chow statistics, and breakpoint Chow statistics.

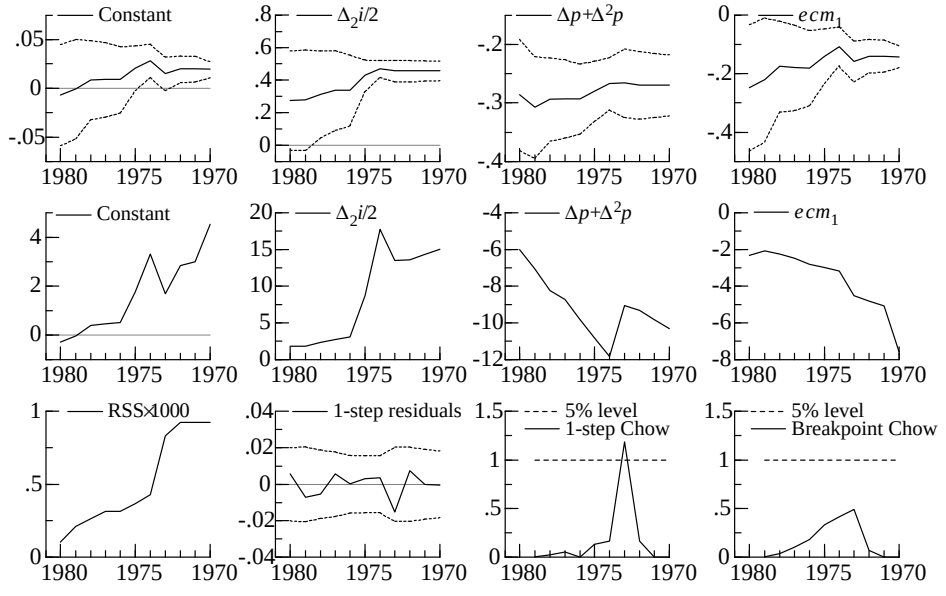


Figure 4: Backward recursive calculations: four estimates and $\pm 2$ estimated standard errors, four $t$-ratios, the residual sum of squares (RSS), one-step residuals and $0 \pm 2\hat{\sigma}_t$, one-step Chow statistics, and breakpoint Chow statistics.

4. *Over-parameterization.* The initial model (3) uses 10 degrees of freedom in estimation, relative to 16 observations total, which, by many accounts, represents extreme over-parameterization. Even the final model (4) has five coefficients, nearly one third the number of observations in the full sample. However, the sample size is only one of three factors that determine how much information is in the sample, and hence how much can be gleaned from it. The three factors follow from rewriting the inverse of the information matrix for the coefficient estimate in the standard linear regression of $y_t$ on $z_t$:

$$\hat{\sigma}^2(Z'Z)^{-1} \;=\; \hat{\sigma}^2\left(\sum z_t z_t'\right)^{-1} \;=\; \hat{\sigma}^2\left(\left[\frac{\sum z_t z_t'}{T}\right]\cdot T\right)^{-1}, \tag{6}$$

where $Z' = (z_1, \ldots, z_T)$. The square root of (6), or the square root of the respective diagonal element, is the estimated standard error of the coefficient estimate. The smaller the estimated standard error, the more information present on that estimated coefficient. From the final expression in (6), three factors determine the standard error on a given coefficient: the estimated equation error variance ($\hat{\sigma}^2$), the estimated average variance of a single observation of the regressor $z_t$ ($\sum z_t z_t'/T$), and the sample size ($T$). Thus, information on a given coefficient could increase for three different reasons. The model could fit better, entailing a smaller value of $\hat{\sigma}^2$; the per-observation second moment $\sum z_t z_t'/T$ could be larger; or $T$ could be larger. Typically, researchers think only of the last — a larger sample size — yet effects from the other two factors can easily outweigh the effects of a *smaller* sample size. That said, in (4), only 16 observations are available for estimation. That number does restrict the complexity of the empirical model, placing a premium on both good economic theory and effective econometric methodology; and, that number of observations could be impracticably small if the signal-to-noise ratio $\left(\sum z_t z_t'/T\right)/\hat{\sigma}^2$ were relatively low.

To illustrate the importance of (6), Table 3 compares these three components for the growth rate of real per capita income in Venezuela and the United States. The overall result is as follows.

$$\frac{\text{Information content (Venezuela)}}{\text{Information content (United States)}} = \frac{1781.}{965.5} = 1.84 \tag{7}$$

That is, the 16 years of annual Venezuelan data have nearly *twice* the information of that in over four decades of quarterly U.S. data. Even though the Venezuelan sample size is one tenth that for the United States, and $\hat{\sigma}$ for Venezuela is about three times that for the United States (taken from Harnett (1988)), the Venezuelan per-observation data variance is over *one hundred* times that for the United States. The 38% increase in Venezuelan income in 1974 and

**Table 3.** The informational content of data:
a comparison of income growth for Venezuela and the United States ($z_t = \Delta i_t$).

| Statistic | Country | |
|---|---|---|
| | Venezuela | United States |
| Sample | 1970–1985 | 1959Q2–1999Q3 |
| Frequency | annual | quarterly |
| $T$ | 16 | 162 |
| $std.dev.(z_t)$ | 0.09982 | 0.008718 |
| $\sum z_t z_t'$ | 0.1495 | 0.01224 |
| $\hat{\sigma}$ | 0.9160% | 0.356% |
| $(\sum z_t z_t')/\hat{\sigma}^2$ | 1781. | 965.5 |

its 9% per annum fall for 1981–1984 are the main sources of this high information content per observation. By comparison, a 38% increase in U.S. real per capita income typically occurs only over a decade or more on post-war data. Granted, the comparison using income is a simple (albeit likelihood-based) measure of the information provided by a single variable, but the picture is quite similar for the model as a whole because the error correction representation involves relatively orthogonal explanatory variables. These arguments also apply to the power of tests.

Thus, even with a small sample size, coefficient estimates can be well-determined and corresponding tests powerful, provided that the per-observation variance of the data is large relative to the innovation error variance. Estimates of (4) are often well-determined even over subsamples, as the recursive estimates and $t$-ratios in Figures 3 and 4 show. Whether or not over-parameterization is a concern depends in part on the nature of the data, and not just on the number of coefficients estimated and the sample size.

## 3.4 Choice of the General Model, and Data Transformations

Even with constructive, algorithmic data mining, two important choices face the empirical modeler: the initial general model, and isomorphic transformations of that model. This subsection discusses the importance of each, beginning with the second.

Section 3.2 showed that the final model (4) is a statistically satisfactory simplification of the ECM (3), and that (4) can be obtained by algorithmic data mining of (5) using PcGets, where (5) is a nonsingular linear transformation of (3). However, while (3) and (5) are equivalent specifications, algorithmic simplification from each

yields different models. In particular, applying PcGets to (3) obtains (8).

$$\begin{aligned}\Delta c_t \;=\; & \underset{(0.030)}{0.232}\,\Delta i_t \;+\; \underset{(0.033)}{0.196}\,\Delta i_{t-1} \;-\; \underset{(0.064)}{0.527}\,\Delta p_t \;+\; \underset{(0.058)}{0.306}\,\Delta p_{t-1} \\ & -\; \underset{(0.029)}{0.152}\,(c-i)_{t-1} \;+\; \underset{(0.015)}{0.060}\,(w-i)_{t-1} \;+\; \underset{(0.0073)}{0.0254}\,D_t \qquad (8)\end{aligned}$$

$T = 16$ [1970–1985] $\quad R^2 = 0.98 \quad \hat{\sigma} = 1.0138\% \quad dw = 2.84$
$AR : F(1,8) = 2.93 \quad ARCH : F(1,7) = 0.48 \quad Normality : \chi^2(2) = 1.37$
$RESET : F(1,8) = 0.43$

Except for the inclusion of a constant term, the final model (4) is nested within (8), and (4) is a statistically satisfactory simplification of the union model nesting (4) and (8): $F(3,8) = 0.07$. Thus, the data transformations that convert (3) into (5) represent the modeler's value added by achieving a more parsimonious simplification.

Because the estimated coefficients in the unrestricted model entail a linear combination of the data, it is always feasible to obtain a completely parsimonious, statistically satisfactory simplification. However, that simplification is not usually of interest as a parametric restriction, whereas economically motivated simplifications are: hence the consideration of the three transformations $\Delta_2 i_t/2$, $\Delta p_t + \Delta^2 p_t$, and $[(c-i) - \frac{1}{2}(w-i)]_{t-1}$. This discussion of parsimony also reflects the arbitrariness of zero restrictions in linear models. A nonzero restriction in one model is a zero restriction in an isomorphic model: equations (3) and (5) show that for the restrictions on income growth, inflation, and the lagged feedback terms.

Models other than (3) might be employed as initial general models. Specifically, while (3) has a large number of parameters relative to the sample size, (3) still imposes two restrictions on the second-order autoregressive distributed lag model (2). Section 3.3 showed that high information content in the data may permit working on a short sample period with models having many parameters, so the remainder of the current subsection considers general-to-specific modeling on that unrestricted autoregressive distributed lag (2). Two parameterizations of (2) are entertained to underscore the importance of data transformations.

The first parameterization is (2) itself, and PcGets is applied directly to that equation. The resulting simplification is (9).

$$\begin{aligned}c_t \;=\; & \underset{(0.085)}{0.409}\,c_{t-1} \;+\; \underset{(0.063)}{0.258}\,c_{t-2} \;+\; \underset{(0.023)}{0.246}\,i_t \;+\; \underset{(0.036)}{0.195}\,i_{t-1} \;-\; \underset{(0.029)}{0.129}\,i_{t-2} \\ & -\; \underset{(0.040)}{0.519}\,p_t \;+\; \underset{(0.043)}{0.597}\,p_{t-1} \;+\; \underset{(0.0046)}{0.0245}\,D_t \qquad (9)\end{aligned}$$

$T = 16$ [1970–1985] $\quad R^2 = 1.00 \quad \hat{\sigma} = 0.6424\%$

While the coefficients in (9) are close to the solved values from (4), (9) is not particularly parsimonious, and its representation in log-levels implies highly correlated regressors and a parameterization that is not very convenient for economic interpretation.

Those features of (9) lead to the second parameterization of (2), which is (5) without the long-run homogeneity and price restrictions:

$$\begin{aligned}\Delta c_t = {} & d_1\Delta c_{t-1} + d_2(\Delta_2 i_t/2) + d_3\Delta i_{t-1} + d_4(\Delta p_t + \Delta^2 p_t) + d_5\Delta p_{t-1} \\ & + d_6\Delta w_{t-1} + d_7[(c-i) - \tfrac{1}{2}(w-i)]_{t-1} + d_8(w-i)_{t-1} \\ & + d_9 + d_{10}D_t + d_{11}i_{t-1} + d_{12}p_{t-1} + u_t, \end{aligned} \tag{10}$$

where $d_{11}$ and $d_{12}$ are coefficients. Thus, (10) is a parameterization from which the final model (4) can be obtained by imposing only zero restrictions. Simplification of (10) by PcGets obtains (11).

$$\begin{aligned}\Delta c_t \; = \; & \underset{(0.030)}{0.458}\ \Delta_2 i_t/2 \; - \; \underset{(0.026)}{0.269}\ (\Delta p_t + \Delta^2 p_t) \\ & - \underset{(0.019)}{0.139}\ [(c-i) - \tfrac{1}{2}(w-i)]_{t-1} \; + \; \underset{(0.0065)}{0.0263}\ D_t \; + \; \underset{(0.00051)}{0.00233}\ i_{t-1} \end{aligned} \tag{11}$$

$$T = 16\ [1970\text{–}1985] \qquad R^2 = 0.98 \qquad \hat{\sigma} = 0.9147\%$$

Equation (11) is identical to the final model (4), except that (11) includes lagged income and excludes the constant term, whereas (4) excludes lagged income and includes the constant term. Interestingly, the coefficients on the overlapping variables are identical to two decimals. Also, the coefficient on $i_{t-1}$ in (11) implies a long-run income elasticity of 0.517, differing only slightly from the (imposed) elasticity of exactly 0.5 in (4). Encompassing tests of (11) and (4) fail to reject either equation, as the nesting model reveals.

$$\begin{aligned}\Delta c_t \; = \; & - \underset{(0.403)}{0.115} \; + \; \underset{(0.033)}{0.460}\ \Delta_2 i_t/2 \; - \; \underset{(0.028)}{0.268}\ (\Delta p_t + \Delta^2 p_t) \\ & - \underset{(0.073)}{0.119}\ [(c-i) - \tfrac{1}{2}(w-i)]_{t-1} \; + \; \underset{(0.0068)}{0.0265}\ D_t \; + \; \underset{(0.049)}{0.016}\ i_{t-1} \end{aligned} \tag{12}$$

$$T = 16\ [1970\text{–}1985] \qquad R^2 = 0.97 \qquad \hat{\sigma} = 0.9554\%$$

The coefficients in (12) are virtually unchanged relative to those in (4) and (11); and the $t$-ratios on the constant term and $i_{t-1}$ are both very small ($-0.28$ and 0.33), reflecting the inability of this data to distinguish between (4) and (11). Economically and statistically, little justification exists for excluding the constant term, whereas long-run homogeneity in income and liquidity has some motivation, and the corresponding restriction is virtually satisfied numerically in (11). Thus, (4) appears to be a reasonable model selection, both economically and statistically.

In summary, constructive algorithmic data mining still leaves the empirical modeler with two important choices: the initial general model, and the particular parameterization of that model. Application of PcGets to a Venezuelan dataset highlights the potential value added in each choice.

# 4 Conclusions and Prospects

In a mixed empirical-Monte Carlo study, Hoover and Perez (1999) demonstrate how an algorithmic approach to constructive data mining can be successful. We discuss how elements of such constructive data mining can empirically counter pejorative senses of data mining. Using PcGets, we then apply constructive data mining to the modeling of consumers' expenditure in Venezuela and show how such data mining can be successful empirically, even on very short samples. With algorithmic data mining, the empirical modeler still faces the important choices of the general model's specification and of its parametric representation before simplification.

Several projects for further research come to mind. First, in a statistical framework, show analytically why the tests' size and power are so well-behaved when there is a specification search. Part of the answer may turn on the tests having power to detect alternatives for which they were not designed, implying that many of the statistics are not independent under various alternatives. Second, analyze how the simplification algorithms implemented in Hoover and Perez (1999) and Hendry and Krolzig (1999) might be improved. For instance, users might want to specify some variables as being non-excludable, or (as with the Venezuelan data) to specify nonzero restrictions directly. See also White (1999) on data snooping. Third, apply the algorithms to additional datasets, both those previously modeled and new ones, in order to gain practical experience with the strengths and weaknesses of algorithmic simplifications. Such knowledge could in turn help redesign the algorithms themselves. Fourth, develop data-mining algorithms for systems of equations. Fifth, systematize the reporting of results, including the documentation of the criteria used and corresponding critical values, so that others can replicate published results.

We congratulate Hoover and Perez on a creative, novel, and provocative paper.

# Appendix. Data Definitions and Measurement

This appendix defines the data, gives their sources, and notes several caveats about their measurement. The data are annual and the sample period is 1968–1985, unless otherwise noted. The descriptions of the series are in alphabetical order, by series symbol. Table A1 (at the end of the appendix) lists the data. These data, along with PcGive files for transforming the data and constructing the empirical results, are also available in computer-readable form from the authors upon request by email (jcampos@gugu.usal.es and ericsson@frb.gov respectively).

The data sources are the *Anuario de Cuentas Nacionales* (Annual Report of the National Accounts), the *Boletín Mensual* (Monthly Bulletin), and the *International Financial Statistics Yearbook*. The first two publications are produced by the Banco Central de Venezuela (Central Bank of Venezuela, Caracas, Venezuela), and the last is produced by the International Monetary Fund (Washington, D.C.). Data from the *Anuario de Cuentas Nacionales* are from the 1982 volume, with extensions and revisions to the data from subsequent issues; data from the *Boletín Mensual* are from various issues; and data from the *International Financial Statistics Yearbook* are from the 1984 and 1987 issues. In each description, the name of the series as it appears in the source publication is given first, and an English (or Spanish) translation appears in brackets. The table ("cuadro") or the line number in the corresponding country table appears after the source publication. We are grateful to Mireya de Cabré, Angel Lucenti, David Mendoza, and Trino Valerio for their help in finding and interpreting the data.

1. Notation. $ABT$

   Name. Gasto de consumo final de los hogares en el mercado interno: alimentos, bebidas y tabaco [Final consumers' expenditure by households in the domestic market: food, drink, and tobacco].

   Definition. Consumers' expenditure on food, drink, and tobacco.

   Units. Millions of Bolívares.

   Source. *Anuario de Cuentas Nacionales*, Cuadro III–5.

2. Notation. $C$

   Name. Real per capita consumers' expenditure on non-durables and services [Gasto de consumo final de los hogares en bienes no durables y servicios, en términos reales, per capita].

   Definition. Constructed as $C = (ABT + OT + SD + S)/(P \cdot N)$.

   Units. 1968 Bolívares per capita.

   Source. Not applicable.

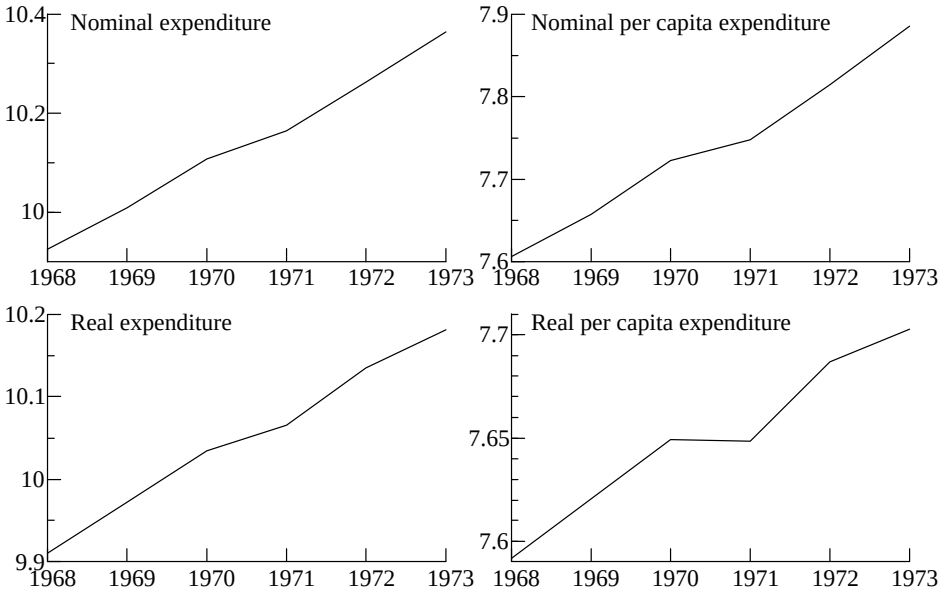


Figure A1: The logs of nominal, nominal per capita, real, and real per capita consumers' expenditure over 1968–1973.

Notes. Consumers' expenditure ($C$) is final expenditure in the domestic market by households and non-profit institutions. It includes purchases in the domestic market by non-residents and excludes purchases by residents in foreign markets. Purchases of each of those two types were estimated to be small relative to measured final expenditure in the domestic market, so no adjustments to $C$ were made.

There appear to be measurement errors in $C$ for 1970 and 1971: Figures A1$a$, A1$b$, A1$c$, and A1$d$ plot the logs of nominal, nominal per capita, real, and real per capita consumers' expenditure for 1968–1973. We have found no economic or institutional explanations of the unusual movements in 1970–1971. Estimation of (4) without the dummy $D_t$ results in virtually identical estimates of the remaining coefficients, but with larger estimated standard errors and $\hat{\sigma}$, as (A1) shows.

$$\begin{aligned}\Delta c_t \;=\; & \underset{(0.0064)}{0.0193} \;+\; \underset{(0.046)}{0.451}\;\Delta_2 i_t/2 \;-\; \underset{(0.039)}{0.262}\;(\Delta p_t + \Delta^2 p_t) \\ & -\; \underset{(0.028)}{0.139}\;[(c-i) - \tfrac{1}{2}(w-i)]_{t-1} \qquad\qquad (A1)\end{aligned}$$

$$T = 16\ [1970\text{–}1985] \qquad R^2 = 0.93 \qquad \hat{\sigma} = 1.3815\% \qquad dw = 2.92$$

These results are consistent with measurement errors on expenditure that are nearly orthogonal to the right-hand side variables in (A1).

Figure A2$a$ graphs the actual and fitted values from (A1), and Figure A2$b$ plots the corresponding residuals. The residuals for 1970 and 1971 are opposite in sign and nearly equal in magnitude, and all other residuals are considerably smaller in magnitude. Figure A3 plots the one-step residuals with bands for plus-or-minus twice the calculated equation standard error. These bands narrow over time, indicative of the relatively large outliers early on in the sample.

One explanation of these anomalies is a minor re-definition of the expenditure series in 1971. Another is an over-estimate of expenditure in 1970 and a compensating under-estimate in 1971. Both explanations are plausible and, without additional evidence, we adopt the latter hypothesis, capturing the effect on $C$ with the dummy variable $D$.

3. Notation. $CTOTAL$

   Name. Gasto de consumo final de los hogares en el mercado interno: total [Final consumers' expenditure by households in the domestic market: total].

   Definition. Total consumers' expenditure. This series is numerically identical to $ABT + DUR + OT + SD + S$.

   Units. Millions of Bolívares.

   Source. *Anuario de Cuentas Nacionales*, Cuadro III–5.

4. Notation. $D$

   Name. A dummy variable for mis-measurement on consumers' expenditure.

   Definition. Constructed as $D_t = +1$ for $t = 1970$, $D_t = -1$ for $t = 1971$, and $D_t = 0$ otherwise.

   Units. Not applicable.

   Source. Not applicable.

5. Notation. $DUR$

   Name. Gasto de consumo final de los hogares en el mercado interno: bienes durables [Final consumers' expenditure by households in the domestic market: durable goods].

   Definition. Consumers' expenditure on durables.

   Units. Millions of Bolívares.

   Source. *Anuario de Cuentas Nacionales*, Cuadro III–5.

6. Notation. $I$

   Name. Real per capita national disposable income, adjusted for the inflation tax on liquid assets [Ingreso nacional disponible, en términos reales, per capita, corregido por la pérdida de valor adquisitivo de la liquidez monetaria].

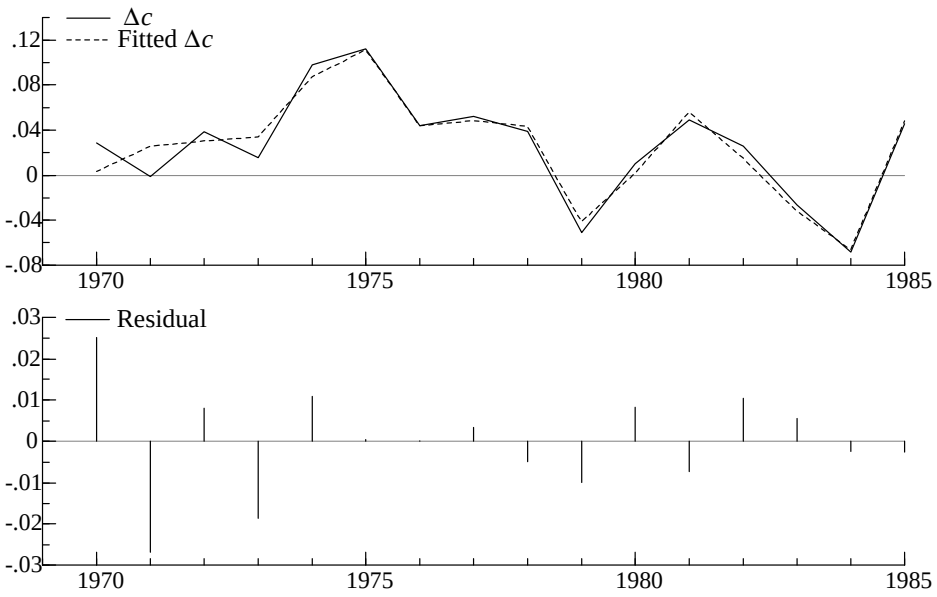


Figure A2: Actual and fitted values of $\Delta c_t$ from (A1), and the corresponding residuals.

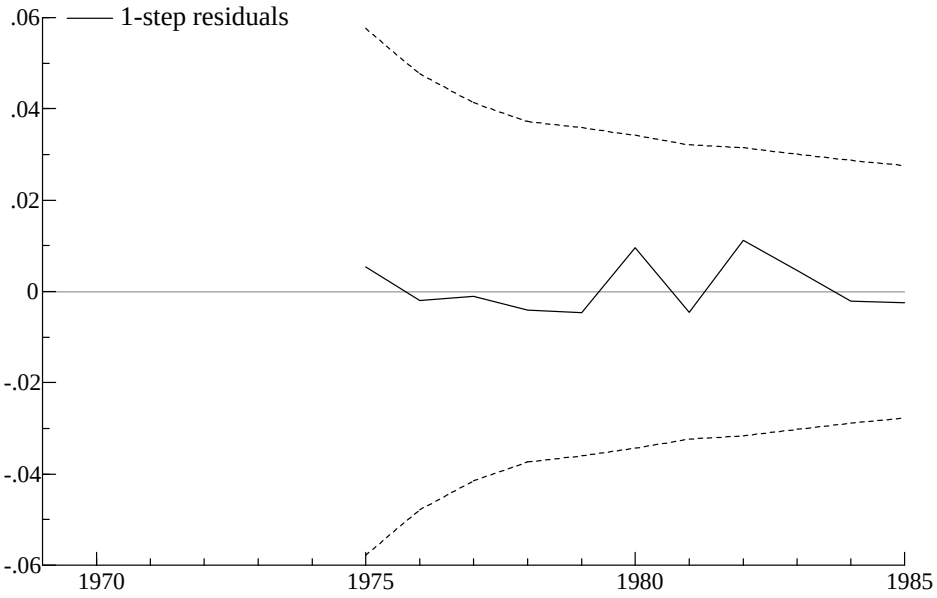


Figure A3: One-step residuals and $0 \pm 2\hat{\sigma}_t$ from (A1).

Definition. Constructed as $I = IN/(P \cdot N)$.

Units. 1968 Bolívares per capita.

Source. Not applicable.

7. Notation. *IN*

   Name. Nominal national disposable income, adjusted for the inflation tax on liquid assets [Ingreso nacional disponible, en términos nominales, corregido por la pérdida de valor adquisitivo de la liquidez monetaria].

   Definition. Constructed as $IN_t = YN_t - (\Delta p_t)\, WN_{t-1}$, i.e., the current year's nominal disposable income, adjusted for the loss in value (due to inflation) of the previous year's end-of-period nominal liquid assets.

   Units. Millions of Bolívares.

   Source. Not applicable.

8. Notation. *N*

   Name. Population [Población].

   Definition. Domestic population of Venezuela.

   Units. Millions of people, mid-year estimate.

   Source. *International Financial Statistics, Yearbook, 1984*, pp. 608–609, line 99z; *International Financial Statistics, Yearbook, 1987*, pp. 712–713, line 99z.

   Notes. A break occurs between 1974 and 1975: overlapping values for 1975 are 11.99 (old), 12.67 (new). We account for that break, proportionately rescaling data before the break to match the post-break value for 1975.

9. Notation. *OT*

   Name. Gasto de consumo final de los hogares en el mercado interno: otros bienes no durables [Final consumers' expenditure by households in the domestic market: other non-durables].

   Definition. Consumers' expenditure on non-durables other than food, drink, and tobacco.

   Units. Millions of Bolívares.

   Source. *Anuario de Cuentas Nacionales*, Cuadro III–5.

10. Notation. *P*

    Name. Indice de precios al consumidor para el área metropolitana de Caracas; índice general [Consumer price index for the Caracas metropolitan area; general index].

Definition. Consumer price index (CPI).

Units. December values, 1968 = 100. The original data are then normalized such that 1968 = 1.00.

Source. *Boletín Mensual*, Cuadro III.4.6.

Notes. The consumer price index ($P$) is for the Caracas metropolitan area only, so it does not reflect regional variations in prices. The index is for all expenditure, both durable and non-durable. Although the index may be well-suited for deflating expenditure of non-durables and services for our econometric analysis, the price of durable goods relative to that of non-durable goods also could be important, especially in light of the multiple interpretations of the term $\Delta p_t + \Delta^2 p_t$ in (4). This index is sensitive to the particular basket of goods used in its calculation and to the presence of price controls, which were common in Venezuela and whose coverage varied over the sample.

11. Notation. $P^*$

    Name. Consumer price index for the United States [Indice de precios al consumidor para los EE.UU.].

    Definition. Consumer price index.

    Units. 1980 = 100, period average.

    Source. *International Financial Statistics, Yearbook, 1987*, pp. 698–699, line 64.

12. Notation. $P^V$

    Name. Consumer price index for Venezuela [Indice de precios al consumidor para Venezuela].

    Definition. Consumer price index.

    Units. 1980 = 100, period average.

    Source. *International Financial Statistics, Yearbook, 1987*, pp. 712–713, line 64.

    Notes. A break occurs between 1983 and 1984, and no overlapping values are available. This series is not used in this study, but it is included for comparison with $P^*$ and $P$.

13. Notation. $S$

    Name. Gasto de consumo final de los hogares en el mercado interno: servicios [Final consumers' expenditure by households in the domestic market: services].

    Definition. Consumers' expenditure on services.

    Units. Millions of Bolívares.

    Source. *Anuario de Cuentas Nacionales*, Cuadro III–5.

14. Notation. *SD*

    Name. Gasto de consumo final de los hogares en el mercado interno: bienes semidurables [Final consumers' expenditure by households in the domestic market: semi-durable goods].

    Definition. Consumers' expenditure on "semi-durables".

    Units. Millions of Bolívares.

    Source. *Anuario de Cuentas Nacionales*, Cuadro III–5.

15. Notation. *VP*

    Name. Volume of petroleum exports [Cantidad de exportaciones de petróleo].

    Definition. Index of the volume of petroleum exports.

    Units. 1980 = 100.

    Source. *International Financial Statistics, Yearbook, 1987*, pp. 712–713, line 72a.

16. Notation. *W*

    Name. Real per capita $M_2$ [Liquidez monetaria $M_2$, en términos reales, per capita].

    Definition. Constructed as $W = WN/(P \cdot N)$.

    Units. 1968 Bolívares per capita.

    Source. Not applicable.

17. Notation. *WN*

    Name. Liquidez monetaria $M_2$ [Monetary aggregate $M_2$].

    Definition. Monedas + billetes + depósitos a la vista + depósitos de ahorro + depósitos a plazo [coins + bills + sight deposits (checking) + savings deposits + time deposits (CDs)].

    Units. Millions of Bolívares, end-of-year.

    Source. *Boletín Mensual*, Cuadro III.2.1.

    Notes: Liquid assets ($WN$) are measured by the monetary aggregate $M_2$ and include holdings by both the personal and commercial sectors, but not those by the government and financial institutions (las Sociedades Financieras). The aggregate $M_2$ excludes the (very liquid) holdings in savings and loans associations (Cédulas Hipotecarias) and holdings abroad. The latter increased dramatically over the sample because of capital flight. While accounting for capital flight may be important, it is not immediately obvious how (e.g.) Venezuelan assets in dollar accounts in the United States affected expenditure in Venezuela.

18. Notation. $XP$

    Name. Petroleum exports [Exportaciones de petróleo].

    Definition. Nominal value of petroleum exports.

    Units. Millions of Bolívares.

    Source. *International Financial Statistics, Yearbook, 1987*, pp. 712–713, line 70a.

19. Notation. $Y$

    Name. Real per capita national disposable income [Ingreso nacional disponible, en términos reales, per capita].

    Definition. Constructed as $Y = YN/(P \cdot N)$.

    Units. 1968 Bolívares per capita.

    Source. Not applicable.

20. Notation. $YN$

    Name. Ingreso nacional disponible [National disposable income].

    Definition. Nominal national disposable income.

    Units. Millions of Bolívares.

    Source. *Anuario de Cuentas Nacionales*, Cuadro I–1, Cuenta 3.

    Notes. National disposable income ($YN$) is not the best measure of income conceivable; personal disposable income would be better, but it is not available. Notably, national disposable income includes profits of the petroleum industry, but those profits should not affect consumers' expenditure directly. Those profits were unusually high in 1973 and low in 1982–1983: these fluctuations may be responsible for (4) over-predicting in 1973 and under-predicting in 1982–1983; cf. Figure 2*c*.

**Table A1.** A listing of the data series.

| Year | Series | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *ABT* | *OT* | *SD* | *DUR* | *S* | *CTOTAL* | *YN* | *P* | *WN* | *D* |
| 1967 | – | – | – | – | – | – | – | 98.90 | 8687 | 0 |
| 1968 | 9109 | 2287 | 2739 | 1905 | 6307 | 22347 | 37598 | 101.48 | 9703 | 0 |
| 1969 | 10244 | 2266 | 2924 | 2057 | 6796 | 24287 | 38961 | 103.75 | 10905 | 0 |
| 1970 | 10744 | 2400 | 3073 | 2152 | 8325 | 26694 | 44363 | 107.62 | 12121 | 1 |
| 1971 | 11405 | 2468 | 3500 | 2185 | 8601 | 28159 | 48250 | 110.45 | 14571 | –1 |
| 1972 | 12589 | 2796 | 3773 | 2453 | 9468 | 31079 | 53594 | 113.59 | 17204 | 0 |
| 1973 | 14281 | 3196 | 3948 | 2916 | 10269 | 34610 | 63990 | 120.07 | 21284 | 0 |
| 1974 | 17745 | 3689 | 5936 | 3847 | 12865 | 44082 | 102206 | 134.01 | 28047 | 0 |
| 1975 | 22713 | 5010 | 7077 | 4938 | 15304 | 55042 | 110268 | 144.68 | 41406 | 0 |
| 1976 | 26776 | 5336 | 8158 | 6561 | 17727 | 64558 | 125786 | 154.77 | 51187 | 0 |
| 1977 | 32655 | 5594 | 8648 | 7732 | 21514 | 76143 | 143611 | 167.29 | 63535 | 0 |
| 1978 | 37281 | 6036 | 11816 | 9572 | 23731 | 88436 | 154200 | 179.19 | 73180 | 0 |
| 1979 | 45983 | 6971 | 13172 | 9911 | 27267 | 103304 | 190324 | 215.93 | 84043 | 0 |
| 1980 | 59004 | 8986 | 14948 | 10970 | 33566 | 127474 | 236421 | 258.30 | 103744 | 0 |
| 1981 | 74727 | 9499 | 15738 | 11122 | 39440 | 150526 | 265836 | 285.51 | 124691 | 0 |
| 1982 | 85767 | 12184 | 17288 | 11322 | 43621 | 170182 | 260311 | 307.87 | 129126 | 0 |
| 1983 | 97157 | 12506 | 15055 | 6437 | 45493 | 176648 | 255612 | 329.28 | 162998 | 0 |
| 1984 | 111230 | 15097 | 17284 | 6831 | 49711 | 200153 | 311505 | 389.50 | 177329 | 0 |
| 1985 | 121761 | 20436 | 21583 | 8885 | 56070 | 228735 | 326413 | 411.60 | 192838 | 0 |

**Table A1.** (continued).

| Year | Series | | | | |
|---|---|---|---|---|---|
| | *XP* | *VP* | *N* | $P^V$ | $P^*$ |
| 1967 | 10267 | 179.7 | 9.31 | 41.99 | 40.52 |
| 1968 | 10370 | 180.6 | 9.62 | 42.54 | 42.22 |
| 1969 | 10141 | 182.3 | 9.94 | 43.58 | 44.49 |
| 1970 | 10550 | 185.7 | 10.28 | 44.67 | 47.12 |
| 1971 | 12814 | 175.6 | 10.61 | 46.11 | 49.15 |
| 1972 | 12571 | 164.5 | 10.94 | 47.41 | 50.77 |
| 1973 | 18632 | 165.7 | 11.28 | 49.37 | 53.93 |
| 1974 | 45200 | 150.4 | 11.63 | 53.45 | 59.85 |
| 1975 | 35668 | 111.6 | 12.67 | 58.95 | 65.32 |
| 1976 | 37593 | 114.7 | 13.12 | 63.42 | 69.08 |
| 1977 | 39106 | 105.3 | 13.59 | 68.36 | 73.58 |
| 1978 | 37517 | 105.3 | 14.07 | 73.19 | 79.17 |
| 1979 | 58519 | 112.5 | 14.55 | 82.28 | 88.13 |
| 1980 | 78328 | 100.0 | 15.02 | 100.00 | 100.00 |
| 1981 | 81723 | 94.1 | 15.48 | 116.17 | 110.35 |
| 1982 | 67068 | 83.2 | 15.94 | 127.34 | 117.15 |
| 1983 | 59473 | 80.4 | 16.39 | 135.33 | 120.91 |
| 1984 | 85226 | 81.6 | 16.85 | 151.78 | 126.07 |
| 1985 | 77641 | 73.4 | 17.32 | 169.08 | 130.55 |

# References


Campos, J. (1984) "Un Modelo Econometrico para el Consumo en Bienes no Durables y Servicios", mimeo, Banco Central de Venezuela, Caracas, Venezuela, December.

Campos, J., and N. R. Ericsson (1988) "Econometric Modeling of Consumers' Expenditure in Venezuela", International Finance Discussion Paper No. 325, Board of Governors of the Federal Reserve System, Washington, D.C., June.

Cochrane, J. H. (1989) "The Sensitivity of Tests of the Intertemporal Allocation of Consumption to Near-rational Alternatives", *American Economic Review*, 79, 3, 319–337.

Denton, F. T. (1985) "Data Mining as an Industry", *Review of Economics and Statistics*, 67, 1, 124–127.

Doornik, J. A., and D. F. Hendry (1996) *PcGive Professional 9.0 for Windows*, International Thomson Business Press, London.

Ericsson, N. R., J. Campos, and H.-A. Tran (1990) "PC-GIVE and David Hendry's Econometric Methodology", *Revista de Econometria*, 10, 1, 7–117.

Ericsson, N. R., and D. F. Hendry (1999) "Encompassing and Rational Expectations: How Sequential Corroboration Can Imply Refutation", *Empirical Economics*, 24, 1, 1–21.

Flemming, J. S. (1976) *Inflation*, Oxford University Press, Oxford.

Harnett, I. (1988) "An Error Correction Model of United States Consumption Expenditure", mimeo, Bank of England, London.

Hendry, D. F. (1995) *Dynamic Econometrics*, Oxford University Press, Oxford.

Hendry, D. F., and H.-M. Krolzig (1999) *PcGets Version 1.0 for Windows*, Nuffield College, Oxford (in preparation).

Hendry, D. F., and T. von Ungern-Sternberg (1981) "Liquidity and Inflation Effects on Consumers' Expenditure", Chapter 9 in A. S. Deaton (ed.) *Essays in the Theory and Measurement of Consumer Behaviour*, Cambridge University Press, Cambridge, 237–260.

Hoover, K. D., and S. J. Perez (1999) "Data Mining Reconsidered: Encompassing and the General-to-specific Approach to Specification Search", *Econometrics Journal*, 2, 2, 167–191 (with discussion).

Leamer, E. E. (1978) *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, John Wiley, New York.

Lovell, M. C. (1983) "Data Mining", *Review of Economics and Statistics*, 65, 1, 1–12.

Sargan, J. D. (1981) "The Choice Between Sets of Regressors", mimeo, Department of Economics, London School of Economics, London, June.

White, H. (1990) "A Consistent Model Selection Procedure Based on $m$-testing", Chapter 16 in C. W. J. Granger (ed.) *Modelling Economic Series: Readings in Econometric Methodology*, Oxford University Press, Oxford, 369–383.

White, H. (1999) "A Reality Check For Data Snooping", mimeo, Department of Economics, University of California at San Diego, La Jolla, California, August (paper presented at the 1999 Econometric Society European Meeting, Santiago de Compostela, Spain).