COMMENTS ON A PAPER BY MARKUS

BY

DAVID A. FREEDMAN

TECHNICAL REPORT NO. 25
MAY 1983

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY

RESEARCH PARTIALLY SUPPORTED BY
NATIONAL SCIENCE FOUNDATION
GRANT MCS-80-02535

May, 1983

Comments on a paper by Markus

by

D. A. Freedman[1]
Statistics Department
University of California, Berkeley

Abstract. Markus gives a regression model for the development of partisanship. A point-by-point critique of the model is offered here, with a view to suggesting that regression models may not be so useful in analyzing complex social phenomena.

Key words and phrases. Regression models, autoregression, time series

[1]Research partially supported by National Science Foundation Grant MCS-80-02535.

Introduction

Regression models are widely used in the social sciences, to describe and explicate complex relationships. Despite their popularity, I do not believe these models have in fact created much new understanding of the phenomena they are intended to illuminate. The main difficulty is that investigators tend to ignore the stochastic assumptions behind the models, perhaps because these assumptions would in practice be hard to validate.

From this point of view, it may be useful to review in detail one example: the model for partisanship developed by Markus in the American Political Science Review. The main intuitive idea behind the model seems to be that partisanship develops in a dynamic way. The basic equation is a first-order autoregression:

$$S_t = \alpha + \beta S_{t-1} + u_t \tag{1}$$

where S_t measures partisanship in period t and u_t is a disturbance.

The justification for (1) is that under certain circumstances, a first-order autogression will have some qualitative properties in common with the data. However, many other equations would satisfy the same criteria. One such is

$$S_t = \alpha + \sum_n \beta^n S_{t-n} + u_t \tag{2}$$

Another is

$$S_t = \alpha(S_{t-1})^\beta u_t \tag{3}$$

These alternative specifications have quite different dynamic properties from (1), and might lead to different substantive conclusions. However, no effort is made to justify the choice of specification (1), rather than (2) or (3) or the infinite number of other possibilities. Indeed, in the present state of the art, it is very unlikely that any such effort could succeed. The theory is too crude to make such discriminations, and so are the data. This point is conceded, a bit obliquely, in footnote 3 to the paper.

Period effects are crucial to the argument in the paper. These are incorporated into (1) by allowing α to depend on t:

$$S_t = \alpha_t + \beta S_{t-1} + u_t \tag{4}$$

At time t, data is available for many different cohorts, which will be indexed by i. So the model becomes

$$S_{i,t} = \alpha_t + \beta S_{i,t-1} + u_{i,t} \tag{5}$$

Here, $S_{i,t}$ is the average partisanship score of cohort i at time t.

The parameters in equation (5) are

- the period effects α_t
- the coefficient β --"stability"

These parameters are unknown, and must be estimated from the data.

No justification is given for the functional form in (5). Indeed, the paper does not derive the autoregressive model from theory, or test it against the data: instead, the model is simply assumed. This is a recurrent problem in social-science uses of regression techniques.

Stochastic assumptions

I turn now to the disturbance terms u_t in (5). In the paper, almost nothing is said about them. However, they are crucial to the statistical argument, so a close look is in order. The parameters α_t and β in (5) can be estimated by a procedure called "least squares;" estimates will be denoted by corresponding roman letters a_t and b. These estimates are found by minimizing the sum of squares

(6) $$\Sigma_{i,t}(S_{i,t} - a_t - bS_{i,t-1})^2$$

Once a_t, b have been computed, $S_{i,t}$ can be "predicted" from the independent variables. The predicted value is denoted with a hat:

(7) $$\hat{S}_{i,t} = a_t + bS_{i,t-1}$$

There is a discrepancy between the observed value $S_{i,t}$ and the predicted value $\hat{S}_{i,t}$. This discrepancy is called a residual, and denoted $\hat{u}_{i,t}$: so

(8) $$S_{i,t} = a_t + bS_{i,t-1} + \hat{u}_{i,t}$$

Equation (8) involves perfectly definite, tangible numbers, which can all be computed from the data. I will call this equation the <u>computer model</u> for the data. The computer model is different from the <u>stochastic model</u> (5). In particular, the hats in (8) signal that the residuals in the computer model are different from the disturbance terms in the stochastic model.

The arithmetic behind (8), namely the minimization of the sum of squares (6), can always be done. A computer neither knows nor cares about the stochastic model (5). However, with some assumptions about the u's, minimizing (6) gives sensible parameter estimates.[1] With other assumptions, the same arithmetic leads to nonsense. Furthermore, the arithmetic does not check itself to make sure it applies. The rule is <u>caveat emptor</u>. In order for the regression to make sense, something must be assumed about the stochastic disturbance terms $u_{i,t}$ in (5). Unless these assumptions are made explicit, the stochastic model must be regarded as incompletely specified. The stochastic

---

[1]Markus uses the more complex Wiley procedure, designed to compensate for measurement errors in the data. I will focus on ordinary least squares, to simplify the exposition. Too, for Markus different cohorts may be in the study for different periods.

model in the paper is incompletely specified. So are the stochastic models in many other social-science papers.

There is a standard set of assumptions to make about the stochastic disturbance terms $u_{i,t}$, which justify the least squares estimation procedures. In statistical jargon, these assumptions can be stated as follows: the errors $u_{i,t}$ are independent of one another, and are identically distributed with mean 0. This translates into quite a definite story: a story not about the data, but about the mechanism which generated the data.

This mechanism is assumed to be like the following hypothetical procedure, in which:

- The variables $S_{i,t}$ are public--in the data base.
- The parameters α_t, β are hidden, not known to the investigator.
- There is still another hidden object: a box of tickets, each ticket bearing a number; these numbers average out to 0.

To make the stochastic assumptions as vivid as possible, I will introduce a fictitious character called "the MC" (for master of ceremonies). The MC generates the data base, one period at a time. Suppose we have gotten through period t-1. Focus on one cohort, say cohort #1. Now the average partisanship score for this cohort is period t-1 has some value, $S_{1,t-1}$. To generate $S_{1,t}$, the MC draws a ticket at random from the box, and makes a note of the number on it. Then the ticket goes back in the box, for the future use, and $S_{1,t}$ is generated by the rule

$$S_{1,t} = \alpha_t + \beta_{1,t-1} + u_{1,t}$$

Exactly the same procedure is followed for the other cohorts: the errors $u_{i,t}$ are always drawn (at random, with replacement) from the same box: so they are independent of one another, and are identically distributed. In particular, the u's show no trend or pattern. These are the crucial assumptions about the mechanism for generating the data.

The MC and the box of tickets are purely fictitious--as are the parameters α_t, β. All the investigator gets to see are the data: the $S_{i,t}$. If the process which generated the data is like the one just described, least squares is a good way to estimate the parameters α_t, β. If the assumptions are wrong, a computer package can still be used to generate the least-squares estimates a_t, b. But these estimates may be biased, e.g., if there is serial correlation in the u's. Or the "standard erros" computed by the package may be off, if e.g. there is correlation in the u's across cohorts. Or there may not be any parameters around to estimate. So if the assumptions are wrong, least squares can be an intellectual disaster. Do the standard assumptions hold for the process in question? The paper does not face this issue squarely. Footnote 12 does acknowledge that the assumptions are open to some question.[1]

The stochastic disturbance terms are unobservable. When you come right down to it, there is only one way to show that the stochastic disturbance terms satisfy the standard conditions: by argument a priori. This involves developing some theory to show where the disturbance terms come from.[2]

One response to this kind of criticism is crudely empirical: "But the model fits." So let us consider that fit. The model is autoregressive: partisanship this time is explained in part by partisanship last time. Autoregressions are expected to fit quite well, because most times series evolve

---

[1]Markus does not discuss the assumptions behind the Wiley estimation procedure used for β; nor the legitimacy of using that procedure for β followed by ols for the α's, with β held fast at its estimated value: this is not standard procedure. The ols estimate for β is .71: see footnote 11. The Wiley estimate is .95: the choice of statistical procedure matters a lot--and therefore assumptions are crucial to the analysis.

[2]For technical discussions of the impact of assumptions on performance of statistical procedures, see e.g. Breiman and Freedman (1983), Freedman (1983), Freedman and Peters (1983).

smoothly: ordinarily, tomorrow will be rather like today. The model in the paper includes almost 20 period effects, one for each election: even if some of the elections depart from the autoregression, the "period effects" should bring them back into line. The model is run on aggregate data; this too usually promotes very high correlations. Even so, the r^2 is only 0.7. Waiving all questions about the assumptions, surely the crude empirical conclusion is that equation (5) does not fit: lots of variance gets away.[1]

Aggregation

Another issue to consider is aggregation. Let us suppose, for the sake of argument, that the model (5) is correct--for individuals:

$$S_{n,t} = \alpha_{n,t} + \beta_n S_{n,t-1} + u_{n,t} \tag{11}$$

Here, $S_{n,t}$ is the n'th person's partisanship score in period t. For the sake of argument, let us even stipulate to the standard assumptions about the $u_{n,t}$: they are independent of one another and have a common distribution with mean 0 and finite standard deviation.

The parameters in (11) were deliberately subscripted: person "n" has his or her own personal "period effect" $\alpha_{n,t}$ and "stability coefficient" β_n. If these parameters are in fact constant across people, then they can be estimated either from aggregate data or from panel data. But the basic assumption seems quite unlikely. To test it, separate regressions would have to be run for each person in the panel study; and the variability in the resulting parameter estimates would have to be analyzed. The paper does not test this basic assumption, or even mention it.[2] If this assumption is false, the specification

---

[1] A draft version of the paper gave the r^2 on microdata as 0.3.

[2] A draft version of the paper argued the weaker assertion, that estimates from aggregate data and panel data are in reasonably good agreement.

(5) is likely to be quite wrong, and serious bias in the estimate of the stability coefficient is likely; see Freedman (1981).

The sampling model

There is another way of looking at regressions. Suppose, for example, that there was a stable population of voters in the United States for the period of the Markus study. Index these voters by n: so $S_{n,t}$ represents the partisanship score of individual n at time t. If $S_{n,t}$ was known for all periods and voters, it would be possible to run the regression

(12) $$S_{n,t} = \alpha_t + \beta S_{n,t-1} + \delta_{n,t}$$

That is, the parameters α_t, β can be defined by the requirement that they minimize the sum of squares

(13) $$\Sigma_{n,t}(S_{n,t} - \alpha_t - \beta S_{n,t-1})^2$$

This definition shortcircuits any question about the errors $\delta_{n,t}$: they are defined by (12). The coefficients α_t, β may be construed as statistics describing the population of voters.[1] So far, so good. The catch is that we cannot do the minimization in (13), because we do not have data on all the voters, but only on the ones in the sample. So it is impossible to compute the parameters α_t, β. However, it is possible to run the regression on the sample, hoping the coefficients a_t, b obtained that way will be good estimates of the population values α_t, β, and hoping too that the standard errors printed out by the computer package will indicate the accuracy of these estimates.[2]

---

[1]Are these descriptive statistics any good? To find out, we have to consider the same issues of linearity and homogeneity raised before, so this reconstruction may only move the difficulty to another place.

[2]Of course, some of the basic assumptions needed to bring off this sampling-theory justification are quite wrong. The population of voters over the study period was not stable. Most of the regressions were run not on individuals but on cohorts.

Unfortunately, it all depends on how the sample was chosen and how the residuals in (12) are related to the variables. The conventional assumptions of simple random sampling and homoscedasticity are embedded in all the computer packages. But the design of the sample used involved a series of repeated cross-sections. And each cross-section presumably involved a multi-stage cluster sample. This design is far from a simple random sample, so the standard errors computed by the conventional formula can be off by a large factor. The issues created by sample design or heteroscedasticity are not considered in the paper, and are glossed over in many other papers too.

Summary and conclusions

To summarize, the form of relationship among the basic variables in the partisanship model is unknown, so there is no basis for the proposed model or the estimation procedure. Therefore, the conclusions drawn from the model do not have adequate scientific foundation.

What are the alternatives? There is no easy, mechanical answer. In particular, I think it would be wrong to introduce still more technique (causal modeling, latent variables, two-stage least squares). I believe it is necessary to begin much close to the beginning. This means figuring out what the basic variables are, and how to measure them. It means collecting good data. It means developing some theory and some ways of looking at the data which will bring out the fundamental laws connecting the variables. Finally, it is necessary to test the theory by making nontrivial predictions about the future, and seeing whether they come true.

I have argued my views at greater length elsewhere, in Freedman (1981, 1983b, c), Freedman-Rothenberg-Sutch (1983a, b). Also see Baumrind (1983), Brown-Koziol (1983), Christ (1975), Hausman-Wise (1982), Hendry (1980), Karlin (1979), Kiefer (1979), Leamer (1983), Ling (1983), Lucas (1975), Lucas and Sargent (1978), Zarnowitz (1979), Zeisel (1982), Pratt and Schlaifer (1981). For some complementary views, see Fienberg (1983), Sims (1980, 1982).

REFERENCES

BAUMRIND, D. (1983). Specious causal attributions in the social sciences: the reformulated stepping-stone theory of heroin use as exemplar. *Jour. of Personality and Social Psychology*. To appear.

BREIMAN, L. and FREEDMAN, D. (1983). How many variables should be entered in a regression equation? *Jour. of the Amer. Statist. Assoc.*, **78** 131-136.

BROWN, C. C. and KOZIOL, J. A. (1983). Statistical aspects of the estimation of human work for suspected environmental carcinogens. *SIAM Review*, **25** no 2, 151-181.

CHRIST, C. (1975). Judging the performance of econometric models of the United States economy. *International Economic Review*, **16** 54-74.

FIENBERG, S. (1983). Statistics and the scientific method: comments on and reactions to Freedman. Technical report no. 196, Department of Statistics, Carnegie-Mellon University. To appear in *Cohort Aanalysis in Social Research: Beyond the Identification Problem*. Eds. W. Mason and S. Fienberg, Academic Press, New York.

FREEDMAN, D. (1981). Some pitfalls in large econometric models: a case study. *Jour. of Business*, **54** 479-500.

FREEDMAN, D. (1983a). A note on screening regression equations. *Amer. Statistician*, to appear.

FREEDMAN, D. (1983b). Statistics and the scientific method. Technical report no. 19. Department of Statistics, University of California, Berkeley. To appear in *Cohort Analysis in Social Research: Beyond the Identification Problem*. Eds. W. Mason and S. Fienberg, Academic Press, New York.

FREEDMAN, D. (1983c). Structural-equation models: a case study. Technical report no. 22. Department of Statistics, University of California, Berkeley.

FREEDMAN, D. and PETERS, S. (1983). Bootstrapping a regression equation: some empirical results. Technical report no. 10, Department of Statistics, University of California, Berkeley. To appear in *Jour. of Amer. Statist. Assoc.*

FREEDMAN, D., PISANI, R. and PURVES, R. (1978). *Statistics.* Norton, New York.

FREEDMAN, D., ROTHENBERG, T. and SUTCH, R. (1983a). On energy policy models. *Jour. of Business and Economic Statist.*, **1** no. 1, 24-36.

FREEDMAN, D., ROTHENBERG, T. and SUTCH, R. (1983b). A review of a residential energy end use model. Technical report no. 14, Department of Statistics, University of California, Berkeley.

HAUSMAN, J. and WISE, D. (1982). Technical problems in social experimentation: cost versus ease of analysis. N.B.E.R. Working Paper No. 1061.

HENDRY, D. (1980). Econometrics - alchemy or science? Inaugural lecture, London School of Economics. *Economica,* **47** 387-406.

KARLIN, S. (1979). Comments on statistical methodology in medical genetics. *Genetic Analysis of Common Diseases: Application to Predictive Factors in Coronary Disease.* 497-520. Alan R. Liss, Inc., New York.

KIEFER, J. (1979). Comments on taxonomy, independence, and mathematical models (with reference to a methodology of Machol and Singer). *Mycologia,* **LXXI** 343-378.

LEAMER, E. (1983). Taking the 'con' out of econometrics. *Amer. Economic Rev.*

LING, R. F. (1983). Review of *Correlation and Causation* by Kenny. *Jour. Amer. Statist. Assoc.,* **77** 489-491.

LUCAS, R. E. (1975). Macroeconomic policy making: a critique. *Jour. of Monetary Economics*.

LUCAS, R. E. and SARGENT, T. J. (1978). After Keynesian macro-economics. In *After the Phillips curve: persistence of high inflation and high employment*, Conference Series No. 19, 49-72, Federal Reserve Bank of Boston.

PRATT, J. W. and SCHLAIFER, R. (1981). On the nature and discovery of structure. Technical report. Harvard Business School.

SIMS, C. A. (1980). Macro-economics and reality. *Econometrica*, **48** 1-48.

SIMS, C. A. (1982). Scientific standards in econometric modelling. Paper presented at the symposium on the developments in econometrics and related fields. Netherlands Econometric Institute.

ZARNOWITZ, V. (1979). An analysis of annual and multiperiod quarterly forecasts of aggregate income, output, and the price level. *Jour. of Business*, **52** 1-34.

ZEISEL, H. (1982). Disagreement over the evaluation of a controlled experiment. *Amer. Jour. of Sociology*, **88** no. 2, 378-389.