The Grunfeld Data at 50

Kleiber, Christian and Zeileis, Achim

Universität Basel, WU Wirtschaftsuniversität Wien

16 February 2010

The Grunfeld Data at 50

Christian Kleiber
Universität Basel

Achim Zeileis
WU Wirtschaftsuniversität Wien

Abstract

This paper revisits Grunfeld's well-known investment data, one of the most widely used data sets in all of applied econometrics, on the occasion of their 50th anniversary. It presents, apparently for the first time after the publication of the original Chicago Ph.D. thesis, the full data set, points out errors and inconsistencies in several currently available versions, and also revisits a number of empirical studies from the literature of the last five decades. Our findings provide a cautionary tale on the use of widely known data and underline the need for mandatory data and code archives.

Keywords: multiple-equation models, panel data, reproducibility.
JEL classification: C80, C23, C30.

1. Introduction

Yehuda Grunfeld – or rather Grünfeld, as the signature on the page of his Ph.D. thesis (Grunfeld 1958) pertaining to reproduction rights reveals – was an exceptionally promising applied econometrician in the second half of the 1950s who died in a drowning accident at the age of 30 (Patinkin 1961; Goodman and Grunfeld 1961). His thesis at the University of Chicago, entitled "The Determinants of Corporate Investment", contains, in an appendix, panel data on a selected set of large US corporations for the period 1935–1954. After his untimely death, these data have been used for illustrating multiple-equation and panel data methodology in research and teaching. In fact, as noted by Greene (2003, p. 329, fn. 39),

> [a]lthough admittedly not current, these data are unusually cooperative for illustrating the different aspects of estimating systems of regression equations.

This paper traces the history of the Grunfeld data over the last five decades and points out errors and inconsistencies in the available variants. It emerges that none of the previously available versions is both complete and correct. An extensive replication exercise reveals that many empirical results are reproducible, at least to a reasonable degree of approximation, once the appropriate version of the data is identified.

Our findings provide a cautionary tale on the use of widely known data and would seem to underline the need for wider adoption of data and code archives, an issue that has gained

great significance in the wake of McCullough and Vinod (2003). As a result of the substantial problems with replicating a single issue of the *American Economic Review* reported in that paper, several leading economics journals, among them the *American Economic Review*, *Econometrica*, the *Journal of Political Economy*, the *Review of Economic Studies* and, more recently, the *Review of Economics and Statistics*, introduced mandatory data and code archives for all new submissions. Previous investigations into the current status of reproducibility in economics have uncovered problems with data construction (Antonovics and Goldberger 2005), model specification and identification issues (McCullough and Vinod 2003) and numerical problems (Zeileis and Kleiber 2005). A more disturbing perspective is fraud (Tödter 2009). Here we consider a hitherto unexplored issue, the effect of widely used but flawed data. Our attempts at replication described below illustrate the widespread impact that bad data can have and underline that data-alone archives will not be able to solve all ensuing problems, instead mandatory data and code archives are required.

To briefly illustrate the problems with widely used data in the virtual absence of data archives, we know of an author who tried to implement a number of diagnostic tests for panel regressions, namely those of Bera *et al.* (2001), and was unable to replicate the Bera *et al.* (2001) numerical illustrations when testing his code. It emerged that his code was correct, but that he unknowingly used the wrong version of the Grunfeld data. Specifically, his version was error-free while Bera *et al.* (2001) used a version containing three errors; once these errors are reintroduced, their computations are reproducible. We provide code for this in the online supplements to this paper, described in greater detail in Appendix A. In addition, these online complements provide the complete original data set along with previously available versions as well as replication files for more than a dozen publications.

2. The story of a data set

What are the Grunfeld data? Our interest was sparked, while working on Kleiber and Zeileis (2008), by the fact that there exist two widely used versions, a 10-firm version popularized by Maddala (1977) and a five-firm subset version popularized by Greene (2003)[1]. Both authors refer to Boot and de Wit (1960) as their source, a paper that provides data for 10 firms. Interestingly, none of the two versions is error-free, nor does use of the relevant subset of the larger data set lead to the same estimates as the smaller one. The fact that both versions are not error-free is known from a suite of TSP benchmarks provided by Cummins (2002); however, there are more errors than previously noted and also a further widely known version with a different set of errors, on which more below. Perusal of some of the references given by Maddala and Greene, among them Grunfeld and Griliches (1960) and Griliches and Wallace (1965), suggested that Boot and de Wit only provide a subset of the original data and that data for at least one further corporation, namely for American Steel Foundries, had been available to Grunfeld. A key reference was Swamy

[1] When referring to textbooks with several editions, we either cite the most recent edition or, if there are substantial differences, the edition that contains the most extensive analysis using the Grunfeld data. In the case of Greene's text, this is the 5th edition (Greene 2003).

Table 1: The Grunfeld data.

Variable	Description
firm	General Motors (GM), US Steel (US), General Electric (GE), Chrysler (CH), Atlantic Refining (AR), IBM, Union Oil (UO), Westinghouse (WH), Goodyear (GY), Diamond Match (DM), American Steel (AS).
investment	Gross investment, defined as additions to plant and equipment plus maintenance and repairs in millions of dollars deflated by the implicit price deflator of producers' durable equipment (base 1947).
value	Market value of the firm, defined as the price of common shares at December 31 (or, for WH, IBM and CH, the average price of December 31 and January 31 of the following year) times the number of common shares outstanding plus price of preferred shares at December 31 (or average price of December 31 and January 31 of the following year) times number of preferred shares plus total book value of debt at December 31 in millions of dollars deflated by the implicit GNP price deflator (base 1947).
capital	Stock of plant and equipment, defined as the accumulated sum of net additions to plant and equipment deflated by the implicit price deflator for producers' durable equipment (base 1947) minus depreciation allowance deflated by depreciation expense deflator (10 years moving average of wholesale price index of metals and metal products, base 1947).

(1970), who claims (p. 320) to use all firms but also claims to have taken the data from Boot and de Wit (1960), but this is clearly at variance with the fact that the latter paper only has 10 firms. How many firms were there originally, and what were the reasons for using the various subsets? Only the original source could shed light on these issues.

Grunfeld's Ph.D. thesis

Grunfeld's 1958 thesis reveals that beyond the corporations considered by Boot and de Wit there is indeed only American Steel Foundries, thus he worked with 11 firms in total. The full data set is given in the appendix of the thesis (Grunfeld 1958, Appendix, Tables 2–9 and 11–13). Table 1 provides the complete list of firms as well as descriptions of all variables (taken from Grunfeld 1958, pp. 153–154, and Griliches and Wallace 1965).

However, Grunfeld was mainly concerned with eight corporations, namely AR, US, UO, GY, DM, AS, GM and GE. The selection of these firms is described in Chapter 2 of the thesis and was guided by consistency requirements. The data are analyzed in Chapters 3 and 4. The fifth and final chapter performs various robustness checks, among them an extension of the sample period to the years 1955 (for all eight previously considered corpo-

rations) and 1956 (for all but GE, for which these data were unavailable) and also to three further large corporations. It is here where the remaining firms (WH, IBM, CH) appear for the first time. Grunfeld also notes (pp. 147–148) that the definition of the variables is not fully consistent, in that he employs "the 'best' methods at [his] disposal" for the three new firms, namely various corrections developed during the course of the preceding chapters. Also, in measuring the value of the firm he now uses "average stock quotations of December 31 and January 31 instead of the single quotations of December 31 used for seven of the eight corporations analyzed previously". This suggests that his measurements are of varying quality.

Of all these firms, DM and AS are somewhat smaller than the others, one reason for their inclusion was to see whether certain hypotheses also hold true for these smaller corporations. The sample period 1935–1954 is a result of the facts that from the year 1935 on expenditures on gross investment and on maintenance and repairs were collected by the Securities and Exchange Commission for all corporations and that Grunfeld's study was started in 1955 (p. 12). Our version, therefore, presents data on all 11 firms for the longest period for which all observations are available, that is, for 1935–1954. For completeness, we also provide the remaining data for the years 1955 and 1956, where available, in a separate file.

Early journal publications

The data were published in a journal for the first time by Boot and de Wit (1960), who provide, as noted above, data for 10 firms (all but AS). These authors are also responsible for the commonly-used ordering of the firms, which is by decreasing mean investment. Interestingly, AS is the second smallest according to this definition; it remains unclear why it was excluded by Boot and de Wit. This abridged version has been used widely, and many subsequent authors are under the impression that it represents the full data set. A printing glitch in this paper is also responsible for one of the errors that occurs in later publications: investment for AR for the year 1953 is correctly given as 91.90, but the first 9 is difficult to read and might be taken for an 8.

In the 1960s, subsets of the data are also used by Grunfeld and Griliches (1960), Griliches and Wallace (1965) and, perhaps most notably, in the course of the development of seemingly unrelated regression (SUR) methodology by Zellner (1962) and later by Kmenta and Gilbert (1968, 1970). Griliches and Wallace note that "only six of the original eleven firms are used in th[eir] study, because [they] did not succeed in reproducing and extending the original set of data for the others" (fn. 8, p. 313). The SUR papers employ just two firms, GE and WH.

Textbook publications

In the 1970s, textbook authors begin to use Grunfeld's data for illustrating multiple equation and panel data models. An early example is Theil (1971) who, following the emerging literature on SUR methodology, employs the GE as well as the WH data when illustrating

Table 2: Versions of the Grunfeld data.

Source	Firms used	Errors
Grunfeld (1958, Tables 2–9, 11–13)	all 11	none (by definition)
Boot and de Wit (1960, Table 10)	10 (all but AS)	none (with printing glitch for $\text{invest}_{\text{AR},1953}$)
Theil (1971, Table 7.1)	2 (GE, WH)	none
Maddala (1977, Table 10-4)	10 (all but AS)	$\text{capital}_{\text{US},1946} = 232.60$, $\text{invest}_{\text{AR},1953} = 81.90$ AR mislabeled "Atlantic Richfield"
Vinod and Ullah (1981, Table 10.1)	10 (all but AS)	see Maddala (1977)
Fomby *et al.* (1984, p. 167–168)	3 (WH, GE, GM)	none
Griffiths *et al.* (1993, Table 17B.1)	10 (all but AS)	see Vinod and Ullah (1981), plus $\text{invest}_{\text{US},1952} = 645.2$
Baltagi (2002, `grunfeld.dat.txt`)	10 (all but AS)	none
Greene (2003, `TableF13-1.txt`)	5 (GM, US, GE, CH, WH)	$\text{capital}_{\text{US},1946} = 232.6$, $\text{invest}_{\text{US},1940} = 261.6$, $\text{invest}_{\text{US},1952} = 645.2$
Baltagi (2005, `Grunfeld.fil`)	10 (all but AS)	none
Greene (2008, `Grunfeld.txt`)	10 (all but AS)	none
Hill *et al.* (2008, `grunfeld.dat`)	10 (all but AS)	see Griffiths *et al.* (1993)

SUR and aggregation issues. Maddala (1977) provides further examples of these techniques but considers all the firms previously used by Boot and de Wit (1960). However, a few errors have crept in there and these were propagated to later versions, for example, the widely used text by Greene (2003). Specifically, Maddala has investment for the year 1953 as 81.90 for AR, while the correct value is 91.90 (the printing glitch mentioned above). Also, capital for the year 1946 is given as 232.6 for US, while the correct value is 132.60. These two errors are mentioned by Cummins (2002). We also note that Maddala uses the label "Atlantic Richfield" instead of "Atlantic Refining". Atlantic Richfield Company (ARCO) was formed by the merger of East Coast-based Atlantic Refining and California-based Richfield Petroleum in 1966 (Encyclopedia Britannica 2009). Maddala states (p. 216) that he reproduces data for 10 firms and that "some of these are different from the ones considered in the Grunfeld-Griliches paper", without further explanation (he does not use AS while Grunfeld and Griliches exclude WH, IBM and CH). Interestingly, the error for $\text{capital}_{\text{US},1946}$ appears in the printed data but is apparently not used in his computations.

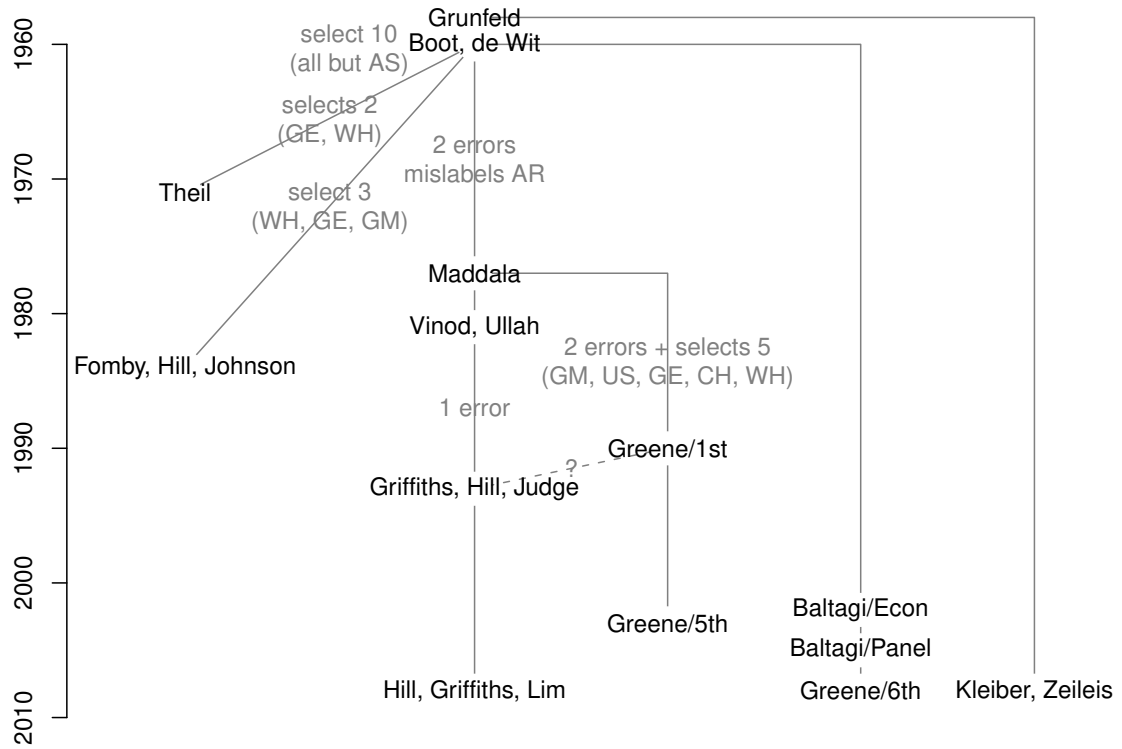

Figure 1: History of the Grunfeld data.

A few years later, Vinod and Ullah (1981), in their monograph on regression methods, also present the 10-firm version which they claim to have taken from Boot and de Wit. However, a closer look reveals that they have the same errors as Maddala. Like Maddala, they also call AR "Atlantic Richfield".

In the mid-1980s, Fomby *et al.* (1984, Chapter 8.4) use the Grunfeld data when discussing FGLS/SUR methods. Following Zellner (1962) and Theil (1971), they consider GE and WH for illustration but do not present any empirical results. In the exercises (pp. 167–168), they provide data for three firms (WH, GE, GM), cite Grunfeld (1958) and ask the reader to compute various estimators.

In textbooks, the correct version for the 10-firm data set resurfaces only in the 1990s in the panel data text by Baltagi (1995). Meanwhile, Greene (1990) opted for a smaller data set comprising five firms that were taken from Maddala. Fortunately, these did not include the data for AR hence he did not inherit the error pertaining to that firm; unfortunately, they did include the data for US with the error mentioned above, namely $\text{capital}_{\text{US},1946} = 232.6$. In addition, two further errors for US were introduced there, namely investment for the year 1940 is given as 261.6 (the correct value is 361.6) and investment for the year 1952 as 645.2 (the correct value is 645.5). The former two errors are noted by Cummins (2002),

the third appears to have gone unnoticed so far.

To muddy the water even further, Griffiths *et al.* (1993) state that their 10-firm version of the Grunfeld data is from "a recent book by Vinod and Ullah" but give no exact reference. They have one additional error compared to Vinod and Ullah, namely $\text{invest}_{\text{US},1952} = 645.2$. Astonishingly, this is one of the errors also introduced by Greene, while the other error introduced there – for the same firm! – does not figure here. In our correspondence with Professors Greene, Griffiths and Hill we have been unable to shed light on this unlikely coincidence. These errors reappear in the still more recent Hill *et al.* (2008) text and its online supplements.

In the most recent edition of Greene's text (Greene 2008) there is a further variation on the theme, in that now a four-firm subset (namely GM, US, GE, and CH) is used for illustrating SUR estimation. The online complements to Greene's text now provide the (error-free) 10-firm version of the data, which seem to have been taken from Baltagi (2005) as the entries are physically identical (except for a mal-formatted header line that is comma-separated instead of space-separated).

Table 2 provides an overview of the various available versions and the errors contained therein, all pertaining to either US or AR. Figure 1 visualizes the history of the data set, and Appendix A presents some further details.

3. Replication

The online supplements to this paper provide replication files for selected results from various papers and textbooks as well as for Grunfeld's thesis. Our collection is by no means exhaustive. In this section, we confine ourselves to contributions published in leading economics and statistics journals. The papers considered, the relevant subsets of the corporations and the methods employed are given in Table 3. The main obstacle to successful replication is the identification of the correct – or rather the appropriate damaged – version of the data. Once this version is identified many results are reproducible, albeit with some effort. We note in passing that R^2s from old papers are not always reproducible although OLS estimates often are, perhaps a result of the fact that there are many ways to compute R^2 which might give different answers, at least in single-precision computations.

For brevity, we do not comment on the numerous textbook examples. Instead, we refer the interested reader to the online supplements, and, in the case of Greene (2003), to the extensive online discussion and errata for that text. For the most recent edition (Greene 2008), we just note that the results using the four-firm subset mentioned above are reproducible. Now there is only one example (pertaining to SUR), but several exercises also make use of the data.

We now briefly discuss selected results from our replications:

- Grunfeld (1958): Despite dating back half a century, the OLS estimates presented in Grunfeld's Ph.D. thesis are almost perfectly reproducible. For 10 out of 11 firms,

Table 3: Selected papers using subsets of the Grunfeld data.

Source	Firms	Methods
Grunfeld and Griliches (1960)	8 (GM, GE, US, AR, UO, DM, GY, AS)	OLS for individual firms and aggregate
Boot and de Wit (1960)	10 (all but AS)	OLS for individual firms and aggregate
Zellner (1962)	2 (GE, WH)	OLS, SUR
Griliches and Wallace (1965)	6 (CH, GY, AR, WH, UO, GM)	OLS for individual firms and aggregate
Kmenta and Gilbert (1968)	2 (GE, WH)	OLS, SUR, iterated SUR/ML
Swamy (1970)	all 11	random coefficient regression
Koenker and Portnoy (1990)	2 (GE, WH)	OLS, SUR, single equation and multivariate M estimation
Bera *et al.* (2001)	5 (GM, US, GE, CH, WH)	diagnostic tests for panel data regressions

there are occasional but very minor variations concerning the last digit given. The only problem is General Motors, for which the coefficients on capital and value are given as 0.4 and 0.116 (we take the former to mean 0.400 as all other values are rounded to three digits by Grunfeld), whereas our computations suggest 0.371 and 0.119. Also, Grunfeld's R^2 is 0.919 while we obtain 0.921. This R^2 for GM is cited by Grunfeld and Griliches (1960) and also reappears in Griliches and Wallace (1965). In view of the excellent agreement of all other estimates numerical problems are unlikely. Instead, the following observation suggests that the problem is more severe.

- Grunfeld also provides, in Table 10 of his appendix, the aggregate data for his main eight corporations. These differ from the aggregate obtained from the individual data for the various firms and, regrettably, also lead to estimates that differ from those presented in his thesis. The fact that his regression for these aggregate data differs from our estimates combined with our earlier observation on problems with the GM regression suggests that copying or typesetting errors are improbable and that, instead, there are differences in the data, be they wrong inputs or unacknowledged data revisions. The explanation for these remaining discrepancies is probably lost to antiquity.

- Boot and de Wit (1960): Interestingly, the estimates of Boot and de Wit are reproducible, including those for GM. The only notable difference pertains to the standard error of the coefficient on capital for US, given as 0.045 whereas 0.142 would seem to

be correct. Also, for the aggregate regression given on p. 10, the standard errors of the slopes appear to have been interchanged. Furthermore, Boot and de Wit appear to use an adjusted R^2 without degrees-of-freedom adjustment for the total sum of squares (TSS), possibly inspired by sources such as Theil (1961).

- Zellner (1962): His OLS estimates are reproducible to no fewer than eight digits. However, his SUR example – the example introducing the widely used SUR estimator! – is not reproducible. Kmenta and Gilbert (1968, p. 1200, fn. 12) maintain that the "slight differences [between their own and Zellner's estimates]... are presumably due to errors of rounding", but this is an unlikely explanation given that Zellner's OLS estimates are impressively accurate even after 46 years. Instead, a closer look at the calculations presented on p. 359 suggests that an algebraic error is responsible for the discrepancies. Specifically, the moment matrices are correct, but the alleged RSS $\hat{U}^\top \hat{U}$ does not correspond to a multivariate linear regression.

- Griliches and Wallace (1965) reproduce Grunfeld's estimates for six firms, they also add an aggregate relation based on these six firms. This aggregate relation is again not reproducible, since their subset includes GM this would seem to support our hypothesis that, for this corporation, Grunfeld used data that differed from those given in his thesis.

- Kmenta and Gilbert (1968) present ML estimates for the SUR model, noting that these coincide with iterated SUR estimates (a fact that was proved only several years later). Their results are reproducible, but for their SUR and ML estimates the coefficients and standard errors for capital and value pertaining to Westinghouse are interchanged.

- Swamy (1970) claims to use all 11 firms but also claims to have taken the data from Boot and de Wit (1960), a claim that is at variance with the fact that the latter paper only has 10 firms. However, Swamy (1971), an extension of his 1968 thesis at the University of Wisconsin, reveals that indeed all 11 firms were used in these computations. Thus Swamy's 1970 *Econometrica* paper and 1971 book are of special interest in that they are the only publications we are aware of that make use of the full data set. In view of the rather sparse information given in his paper, replication of its random coefficient estimates proved far from trivial. Of course, its main contributions and focus are theoretical; the computations only illustrate the main findings. Numerical experiments with R and Stata, using raw as well as demeaned data, resulted in estimates that are quite close to those reported in Swamy's paper. Advances in numerical algorithms and almost four decades of software development would seem to account for the remaining discrepancies.

- Koenker and Portnoy (1990): Table 1 is not fully reproducible due to insufficient numerical detail, notably the standard errors are unclear. The single-equation L_1 and M estimates (their Table 2) are reproducible though.

- As mentioned in the introduction, Bera *et al.* (2001) provide various diagnostic tests for panel data regressions. Their numerical illustrations are fully reproducible once the Greene five-firm version with its three errors is employed.

4. Conclusion

We have reviewed the long and convoluted history of one of the most widely used data sets in all of econometrics. Evaluation of the various versions suggests that a considerable amount of caution is necessary when working with these data, and that differing estimates may be traced to versions of varying degrees of reliability. In view of Figure 1, all versions derived from Maddala are probably best avoided. Also, all versions derived from Boot and de Wit are incomplete.

Given the sparse information regarding computational aspects provided in many of the older publications, a substantial amount of detective work was often required to reproduce empirical results, not always successfully. Of course it would have been unreasonable to expect perfect reproducibility. However, the rather small number of fully reproducible works is sobering: essentially only some (but not all) papers dealing with SUR methodology using the GE and WH data and some more recent textbook results are fully replicable. Our results would thus seem to underline the recent requests for mandatory data and code archives, see, e.g., Anderson *et al.* (2008), McCullough *et al.* (2006) and McCullough *et al.* (2008) and the references therein.

It should also be emphasized why many publications are only reproducible with some effort: several published versions of the Grunfeld data are plagued by transcription errors that were propagated to estimation results. This is another advantage of electronic data archives, which help to eliminate one source of error that presumably affects a number of works from past decades. Of course such archives did not exist in the early days of econometric computing and have only become available as a result of more recent technological advances such as the Internet. However, the message is clear: the profession needs to make use of these technological advances on a much broader basis; specifically, every journal interested in publishing reproducible research should adopt mandatory data archives. But our results suggest even more than this: Data-alone archives will not help to eliminate problems such as the non-replicable aggregate regression described above. In addition to the data, code is necessary to have any hope of reproducibility. Perhaps this aspect is nowadays more important than it used to be given that current methods of inference are inherently more complex than OLS and its many variations, the prevalent methodology of the 1950s and 1960s. Typically, modern methods cannot be described in all algorithmic detail given the scarcity of available journal space. Code archives will ensure that crucial details of complex algorithms and their implementations are available which alone permit successful replication.

Also, the Grunfeld data have proven immensely useful for the development of econometric methodology as well as for the development of econometric software. We refer again to the

quotation taken from Greene (2003) presented in the introduction, according to which the data are eminently suitable for illustrating multiple-equation methodology. However, in view of the numerous accidents described in this paper it would appear that the profession needs to agree on a benchmark version of the Grunfeld data. The full 11-firm version provided here would seem to be appropriate. No doubt many further studies in the field will make use of these data. It is hoped that future explorations will be better documented than some of those in the past.

Finally, it is also highly unlikely that only the Grunfeld data are plagued by errors of various types. It would be interesting to examine other widely used data sets and estimation results based thereon.

Computational details

Our results were obtained using R 2.10.1 (R Development Core Team 2009) with the packages **plm** 1.2-3 (Croissant and Millo 2008) and **systemfit** 1.1-4 (Henningsen and Hamann 2007) for estimating panel models and multiple-equation models, respectively. Furthermore, the packages **lmtest** 0.9-26 (Zeileis and Hothorn 2002), **MASS** 7.3-5 (Venables and Ripley 2002), **quantreg** 4.44 (Koenker 2009), and **sandwich** 2.2-5 (Zeileis 2004, 2006) were employed for some analyses. R itself and all packages used are freely available under the terms of the General Public License from the Comprehensive R Archive Network at `http://CRAN.R-project.org/`. All versions of the data, replication files and associated R output are available from `http://statmath.wu.ac.at/~zeileis/grunfeld/` (see also Appendix A). All results were identical on various platforms including Debian GNU/Linux (with a 2.6.29 kernel) and Mac OS X, version 10.5.8.

A. Online complements

All electronic resources accompanying this paper are available online at http://statmath.wu.ac.at/~zeileis/grunfeld/.

- **Data:** We provide a text file `Grunfeld.csv` (comma-separated values) that contains the original and complete 11-firm data set from Grunfeld's Ph.D. thesis (Appendix, Tables 2–9, 11–13). For R users, it is conveniently accessible as `data("Grunfeld", package = "AER")` if the package **AER**, accompanying Kleiber and Zeileis (2008), is installed. In addition to the main 11-firm data set, Grunfeld's aggregate data – his Table 10, which is inconsistent with the aggregate obtained from the 11-firm data set – are available as `Grunfeld-agg.csv`. Furthermore, Grunfeld's additional data for eight firms for the years 1955–1956 are available as `Grunfeld-ext.csv`.

- **Replications:** For each publication considered here, an annotated R script is provided along with its output generated with a current version of R and all required packages (as of 2010-01-29). Replication files are provided for: Grunfeld (1958), Grunfeld and Griliches (1960), Boot and de Wit (1960), Zellner (1962), Griliches and Wallace (1965), Kmenta and Gilbert (1968), Swamy (1970), Theil (1971), Maddala (1977), Koenker and Portnoy (1990), Griffiths *et al.* (1993), Bera *et al.* (2001), Baltagi (2002), Greene (2003), Baltagi (2005), Greene (2008).

- **Further data versions:** For textbooks with online electronic versions of the Grunfeld data, we provide a link to the original resource as well as to a local copy obtained on 2008-12-22. An R script is available that automatically compares these versions against the relevant subsets of the full data set. Textbooks with online electronic versions of the data include: Baltagi (2002), Greene (2003), Baltagi (2005), Greene (2008), Hill *et al.* (2008).

References

Anderson RD, Greene WH, McCullough BD, Vinod HD (2008). "The Role of Data/Code Archives in the Future of Economic Research." *Journal of Economic Methodology*, **15**, 99–119.

Antonovics KL, Goldberger AS (2005). "Does Increasing Women's Schooling Raise the Schooling of the Next Generation? Comment." *American Economic Review*, **95**, 1738–1744.

Baltagi BH (1995). *Econometric Analysis of Panel Data*. John Wiley & Sons, Chichester.

Baltagi BH (2002). *Econometrics*. 3rd edition. Springer-Verlag. URL http://www.springer.com/economics?SGWID=0-165-2-107420-0.

Baltagi BH (2005). *Econometric Analysis of Panel Data*. 3rd edition. John Wiley & Sons, Chichester. URL http://www.wiley.com/legacy/wileychi/baltagi3e/.

Bera AK, Sosa-Escudero W, Yoon M (2001). "Testing for the Error Component Model in the Presence of Local Misspecification." *Journal of Econometrics*, **101**, 1–23.

Boot JCG, de Wit GM (1960). "Investment Demand: An Empirical Contribution to the Aggregation Problem." *International Economic Review*, **1**(1), 3–30.

Croissant Y, Millo G (2008). "Panel Data Econometrics in R: The **plm** Package." *Journal of Statistical Software*, **27**(2), 1–43. URL http://www.jstatsoft.org/v27/i02/.

Cummins C (2002). "Different Versions of Grunfeld Dataset." Web Page (accessed 2009-05-18), URL http://www.stanford.edu/~clint/bench/grunfeld.htm.

Encyclopedia Britannica (2009). "Atlantic Richfield Company (ARCO)." Online version (accessed 2009-05-18), URL http://search.eb.com/eb/article-9010104.

Fomby TB, Hill RC, Johnson SR (1984). *Advanced Econometric Methods*. Springer-Verlag, New York.

Goodman LA, Grunfeld Y (1961). "Some Nonparametric Tests for Comovements Between Time Series." *Journal of the American Statistical Association*, **56**, 11–26.

Greene WH (1990). *Econometric Analysis*. Macmillan, London/New York.

Greene WH (2003). *Econometric Analysis*. 5th edition. Prentice Hall, Upper Saddle River, NJ. URL http://pages.stern.nyu.edu/~wgreene/Text/econometricanalysis.htm.

Greene WH (2008). *Econometric Analysis*. 6th edition. Prentice Hall, Upper Saddle River, NJ. URL http://pages.stern.nyu.edu/~wgreene/Text/econometricanalysis.htm.

Griffiths WE, Hill RC, Judge GG (1993). *Learning and Practicing Econometrics*. John Wiley & Sons, New York.

Griliches Z, Wallace N (1965). "The Determinants of Investment Revisited." *International Economic Review*, **6**(3), 311–329.

Grunfeld Y (1958). *The Determinants of Corporate Investment*. Ph.D. thesis, Department of Economics, University of Chicago.

Grunfeld Y, Griliches Z (1960). "Is Aggregation Necessarily Bad?" *The Review of Economics and Statistics*, **42**(1), 1–13.

Henningsen A, Hamann JD (2007). "**systemfit**: A Package for Estimating Systems of Simultaneous Equations in R." *Journal of Statistical Software*, **23**(4), 1–40. URL http://www.jstatsoft.org/v23/i04/.

Hill RC, Griffiths WE, Lim GC (2008). *Principles of Econometrics*. 3rd edition. John Wiley & Sons, New York.

Kleiber C, Zeileis A (2008). *Applied Econometrics with R*. Springer-Verlag, New York.

Kmenta J, Gilbert RF (1968). "Small-Sample Properties of Alternative Estimators of Seemingly Unrelated Regressions." *Journal of the American Statistical Association*, **63**(324), 1180–1200.

Kmenta J, Gilbert RF (1970). "Estimation of Seemingly Unrelated Regressions with Autoregressive Disturbances." *Journal of the American Statistical Association*, **65**(329), 186–197.

Koenker R (2009). ***quantreg**: Quantile Regression*. R package version 4.44, URL http://CRAN.R-project.org/package=quantreg.

Koenker R, Portnoy S (1990). "*M* Estimation of Multivariate Regressions." *Journal of the American Statistical Association*, **85**(412), 1060–1068.

Maddala GS (1977). *Econometrics*. McGraw-Hill, New York.

McCullough BD, McGeary KA, Harrison TD (2006). "Lessons from the JMCB Archive." *Journal of Money, Credit and Banking*, **38**(4), 1093–1107.

McCullough BD, McGeary KA, Harrison TD (2008). "Do Economics Journal Archives Promote Replicable Research?" *Canadian Journal of Economics*, **41**(4), 1406–1420.

McCullough BD, Vinod HD (2003). "Verifying the Solution from a Nonlinear Solver: A Case Study." *American Economic Review*, **93**, 873–892.

Patinkin D (1961). "In Memoriam [Yehuda Grunfeld]." *Econometrica*, **29**(3), 404–405.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Swamy PAVB (1970). "Efficient Inference in a Random Coefficient Regression Model." *Econometrica*, **38**(2), 311–323.

Swamy PAVB (1971). *Statistical Inference in Random Coefficient Regression Models*. Springer-Verlag, New York.

Theil H (1961). *Economic Forecasts and Policy*. 2nd edition. North-Holland, Amsterdam.

Theil H (1971). *Principles of Econometrics*. John Wiley & Sons, New York.

Tödter KH (2009). "Benford's Law as an Indicator of Fraud in Economics." *German Economic Review*, **10**(3), 339–351.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. 4th edition. Springer-Verlag, New York.

Vinod HD, Ullah A (1981). *Recent Advances in Regression Methods*. Marcel Dekker, New York.

Zeileis A (2004). "Econometric Computing with HC and HAC Covariance Matrix Estimators." *Journal of Statistical Software*, **11**(10), 1–17. URL http://www.jstatsoft.org/v11/i10/.

Zeileis A (2006). "Object-Oriented Computation of Sandwich Estimators." *Journal of Statistical Software*, **16**(9), 1–16. URL http://www.jstatsoft.org/v16/i09/.

Zeileis A, Hothorn T (2002). "Diagnostic Checking in Regression Relationships." *R News*, **2**(3), 7–10. URL http://CRAN.R-project.org/doc/Rnews/.

Zeileis A, Kleiber C (2005). "Validating Multiple Structural Change Models – A Case Study." *Journal of Applied Econometrics*, **20**, 685–690.

Zellner A (1962). "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias." *Journal of the American Statistical Association*, **57**, 348–368.

Affiliation:

Christian Kleiber
Faculty of Business and Economics (WWZ)
Universität Basel
Peter Merian-Weg 6
4002 Basel, Switzerland
E-mail: Christian.Kleiber@unibas.ch
URL: http://wwz.unibas.ch/kleiber/

Achim Zeileis
Department of Statistics and Mathematics
WU Wirtschaftsuniversität Wien
Augasse 2–6
1090 Wien, Austria
E-mail: Achim.Zeileis@R-project.org
URL: http://statmath.wu.ac.at/~zeileis/