Applied Econometrics with R: Package Vignette and Errata

Christian Kleiber
Universität Basel

Achim Zeileis
Universität Innsbruck

Abstract

"Applied Econometrics with R" (Kleiber and Zeileis 2008, Springer-Verlag, ISBN 978-0-387-77316-2, pp. vii+222) is the first book on applied econometrics using the R system for statistical computing and graphics (R Core Team 2019). It presents hands-on examples for a wide range of econometric models, from classical linear regression models for cross-section, time series or panel data and the common non-linear models of microeconometrics, such as logit, probit, tobit models as well as regression models for count data, to recent semiparametric extensions. In addition, it provides a chapter on programming, including simulations, optimization and an introduction to R tools enabling reproducible econometric research. The methods are presented by illustrating, among other things, the fitting of wage equations, growth regressions, dynamic regressions and time series models as well as various models of microeconometrics.

The book is accompanied by the R package **AER** (Kleiber and Zeileis 2019) which contains some new R functionality, some 100 data sets taken from a wide variety of sources, the full source code for all examples used in the book, as well as further worked examples, e.g., from popular textbooks.

This vignette provides an overview of the package contents and contains a list of errata for the book.

Keywords: econometrics, statistical software, R.

1. Package overview

1.1. R code from the book

The full R code from the book is provided in the demos for the package **AER**. The source scripts can be found in the `demo` directory of the package and executed interactively by calling `demo()`, as in

```
R> demo("Ch-Intro", package = "AER")
```

One demo per chapter is provided:

- `Ch-Intro` (Chapter 1: Introduction),
- `Ch-Basics` (Chapter 2: Basics),
- `Ch-LinearRegression` (Chapter 3: Linear Regression),

- `Ch-Validation` (Chapter 4: Diagnostics and Alternative Methods of Regression),
- `Ch-Microeconometrics` (Chapter 5: Models of Microeconometrics),
- `Ch-TimeSeries` (Chapter 6: Time Series),
- `Ch-Programming` (Chapter 7: Programming Your Own Analysis).

This list of demos is also shown by `demo(package = "AER")`.

The same scripts are contained in the `tests` directory of the package so that they are automatically checked and compared with the desired output provided in `.Rout.save` files.

To make the code fully reproducible and to avoid some lengthy computations in the daily checks, a few selected code chunks are commented out in the scripts. Also, for technical reasons, some graphics code chunks are repeated, once commented out and once without comments.

1.2. Data sets

The **AER** package includes some 100 data sets from leading applied econometrics journals and popular econometrics textbooks. Many data sets have been obtained from the data archive of the *Journal of Applied Econometrics* and the (now defunct) data archive of the *Journal of Business & Economic Statistics* (see note below). Some of these are used in recent textbooks, among them Baltagi (2002), Davidson and MacKinnon (2004), Greene (2003), Stock and Watson (2007), and Verbeek (2004). In addition, we provide all further data sets from Baltagi (2002), Franses (1998), Greene (2003), Stock and Watson (2007), and Winkelmann and Boes (2009). Selected data sets from Franses, van Dijk, and Opschoor (2014) are also included.

Detailed information about the source of each data set, descriptions of the variables included, and usually also examples for typical analyses are provided on the respective manual pages. A full list of all data sets in **AER** can be obtained via

```
R> data(package = "AER")
```

In addition, manual pages corresponding to selected textbooks are available. They list all data sets from the respective book and provide extensive code for replicating many of the empirical examples. See, for example,

```
R> help("Greene2003", package = "AER")
```

for data sets and code for Greene (2003). Currently available manual pages are:

- `Baltagi2002` for Baltagi (2002),
- `CameronTrivedi1998` for Cameron and Trivedi (1998),
- `Franses1998` for Franses (1998),
- `Greene2003` for Greene (2003),
- `StockWatson2007` for Stock and Watson (2007).

- `WinkelmannBoes2009` for Winkelmann and Boes (2009).

1.3. New R functions

AER provides a few new R functions extending or complementing methods previously available in R:

- `tobit()` is a convenience interface to `survreg()` from package **survival** for fitting tobit regressions to censored data. In addition to the fitting function itself, the usual set of accessor and extractor functions is provided, e.g., `print()`, `summary()`, `logLik()`, etc. For more details see `?tobit`.
- `ivreg()` fits instrumental-variable regressions via two-stage least squares. It provides a formula interface and calls the workhorse function `ivreg.fit()` which in turn calls `lm.fit()` twice. In addition to the fitting functions, the usual set of accessor and extractor functions is provided, e.g., `print()`, `summary()`, `anova()`, etc. For more details see `?ivreg`, `?ivreg.fit`, and `?summary.ivreg`, respectively.
- `dispersiontest()` tests the null hypothesis of equidispersion in Poisson regressions against the alternative of overdispersion and/or underdispersion. For more details see `?dispersiontest`.

2. Errata and comments

Below we list the errors that have been found in the book so far. Please report any further errors you find to us.

We also provide some comments, for example on functions whose interface has changed.

- p. 5–9, 46–53: There are now very minor differences in the plots pertaining to Example 2 (Determinants of wages) in Chapter 1.1 and Chapter 2.8 (Exploratory Data Analysis with R) due to a missing observation. Specifically, the version of the `CPS1985` data used for the book contained only 533 observations, the original observation 1 had been omitted inadvertently.
- p. 38, 48, 85: By default there is less rounding in calls to `summary()` starting from R 3.4.0.
- p. 63–65, 130, 143: The function `linear.hypothesis()` from the **car** package is now defunct, it has been replaced by `linearHypothesis()` starting from **car** 2.0-0.
- p. 85–86: Due to a bug in the `summary()` method for "`plm`" objects, the degrees of freedom reported for the F statistics were interchanged and thus the p values were not correct. Therefore, the p values printed in the book at the end of `summary(gr_fe)` and `summary(gr_re)` are not correct, they should both be `< 2.22e-16`. Using **plm** 1.1-1 or higher, the code produces the correct output. Also the degrees-of-freedom adjustment in the p values for the coefficient tests in `summary(gr_re)` were corrected.

- pp. 88–89: As of version 1.3-1 of the **plm** package, summaries of "`pgmm`" objects provide robust standard errors by default. The output presented on pp. 88–89 is still available, but now requires `summary(empl_ab, robust = FALSE)`.

  Also, the formula interface for `pgmm()` has changed: as of version 1.7-0 of the **plm** package, the function `dynformula()` is deprecated. Instead, lags should now be specified via the package's `lag()` function. In addition, instruments should now be specified via a two-part formula.

  Using the new interface, the function call for the Arellano-Bond example is

```
R> empl_ab <- pgmm(log(emp) ~ lag(log(emp), 1:2) + lag(log(wage), 0:1)
+      + log(capital) + lag(log(output), 0:1) | lag(log(emp), 2:99),
+      data = EmplUK, index = c("firm", "year"),
+      effect = "twoways", model = "twosteps")
```

- p. 92: Exercise 6 cannot be solved using `PSID1982` since that data set only contains a cross-section while Hausman-Taylor requires panel data. A panel version has been available in the **plm** package under the name `Wages`; we have now added `PSID7682` to **AER** for completeness (and consistent naming conventions). Use `PSID7682` for the exercise.

- pp. 98–100: R only provides a function `dffits()` but not `dffit()` as claimed on p. 99. Somewhat confusingly the corresponding column in the output of `influence.measures()` (as shown on p. 100) is called `dffit` by R (rather than `dffits`).

- p. 124: The argument `ylevels = 2:1` in the spinogram is no longer needed because the defaul ordering of the y-levels changed in R 4.0.0.

- p. 141: The log-likelihood for the tobit model lacked a minus sign. The correct version is

$$\ell(\beta, \sigma^2) = \sum_{y_i > 0} \Big(\log \phi\{(y_i - x_i^\top \beta)/\sigma\} - \log \sigma \Big) + \sum_{y_i = 0} \log \Phi(-x_i^\top \beta/\sigma).$$

- p. 149: The standard error (and hence the corresponding z test) of `admin|manage` in the output of `coeftest(bank_polr)` is wrong, it should be `1.4744`. This was caused by an inconsistency between `polr()` and its `vcov()` method which has now been improved in the **MASS** package ($\geq$ 7.3-6).

- p. 167: The truncation lag parameter in the output of `kpss.test(log(PepperPrice[, "white"]))` is wrong, it should be `5` instead of `3`, also leading to a somewhat smaller test statistic and larger p value. This has now been corrected in the **tseries** package ($\geq$ 0.10-46).

- p. 169: The comment regarding the output from the Johansen test is in error. The null hypothesis of no cointegration is not rejected at the 10% level. Nonetheless, the table corresponding to Case 2 in Juselius (2006, p. 420) reveals that the trace statistic is significant at the 15% level, thus the Johansen test weakly confirms the initial two-step approach.

- p. 179: For consistency, the GARCH code should be preceded by `data("MarkPound")`.

- p. 192: The likelihood for the generalized production function was in error (code and computations were correct though).

  The correct likelihood for the model is

$$\mathcal{L} = \prod_{i=1}^{n} \left\{ \frac{1}{\sigma} \phi\left(\frac{\varepsilon_i}{\sigma}\right) \cdot \frac{1 + \theta Y_i}{Y_i} \right\}.$$

  giving the log-likelihood

$$\ell = \sum_{i=1}^{n} \{\log(1 + \theta Y_i) - \log Y_i\} - n \log \sigma + \sum_{i=1}^{n} \log \phi(\varepsilon_i/\sigma).$$

- p. 205: The reference for Henningsen (2008) should be:

  > Henningsen A (2008). "Demand Analysis with the Almost Ideal Demand System in R: Package **micEcon**," Unpublished. URL http://CRAN.R-project.org/package=micEcon.

Note: Currently, all links on manual pages corresponding to data sets taken from the Journal of Business & Economic Statistics (JBES) archive are broken (data sets `MarkPound`, and `RecreationDemand`). The reason is the redesign of the American Statistical Association (ASA) website, rendering the old ASA data archive nonfunctional. The ASA journals manager currently appears to supply data on a case-by-case basis. The problem awaits a more permanent solution.

References

Baltagi BH (2002). *Econometrics*. 3rd edition. Springer-Verlag, New York. URL https://www.springer.com/us/book/9783662046937.

Cameron AC, Trivedi PK (1998). *Regression Analysis of Count Data*. Cambridge University Press, Cambridge.

Davidson R, MacKinnon JG (2004). *Econometric Theory and Methods*. Oxford University Press, Oxford.

Franses PH (1998). *Time Series Models for Business and Economic Forecasting*. Cambridge University Press, Cambridge.

Franses PH, van Dijk D, Opschoor A (2014). *Time Series Models for Business and Economic Forecasting*. 2nd edition. Cambridge University Press, Cambridge. URL http://www.cambridge.org/us/academic/subjects/economics/econometrics-statistics-and-mathematical-economics/time-series-models-business-and-economic-forecasting-2nd-edition.

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

Juselius K (2006). *The Cointegrated VAR Model*. Oxford University Press, Oxford.

Kleiber C, Zeileis A (2008). *Applied Econometrics with R*. Springer-Verlag, New York. ISBN 978-0-387-77316-2.

Kleiber C, Zeileis A (2019). ***AER**: Applied Econometrics with R*. R package version 1.2-7, URL https://CRAN.R-project.org/package=AER.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Stock JH, Watson MW (2007). *Introduction to Econometrics*. 2nd edition. Addison-Wesley, Reading, MA.

Verbeek M (2004). *A Guide to Modern Econometrics*. 2nd edition. John Wiley & Sons, Hoboken, NJ.

Winkelmann R, Boes S (2009). *Analysis of Microdata*. 2nd edition. Springer-Verlag, Berlin and Heidelberg.

Affiliation:

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

Achim Zeileis
Department of Statistics
Faculty of Economics and Statistics
Universität Innsbruck
Universitätsstr. 15
6020 Innsbruck, Austria
E-mail: Achim.Zeileis@R-project.org
URL: https://eeecon.uibk.ac.at/~zeileis/