resampling in the bootstrap. These are quite specialized uses of the package and so the user is advised to read the relevant sections of Davison and Hinkley (1997) before using these functions.

Acknowledgments

I would like to thank A. C. Davison and V. Ventura for their many helpful suggestions in the development of this library. Thanks are also due to B. D. Ripley for a great deal of help and for porting the code to R. Any bugs in the code, however, are my responsibility and should be reported to me in the first instance.

Bibliography

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993), *Statistical Models Based on Counting Processes*, New York: Springer. 5

Braun, W. J. and Kulperger, R. J. (1997), "Properties of a Fourier bootstrap method for time series," *Communications in Statistics — Theory and Methods*, **26**, 1329–1327. 6

Canty, A. J. and Davison, A. C. (1999), "Implementation of saddlepoint approximations in resampling problems," *Statistics and Computing*, **9**, 9–15. 6

Canty, A. J., Davison, A. C., and Hinkley, D. V. (1996), "Reliable confidence intervals. Discussion of "Bootstrap confidence intervals", by T. J. DiCiccio and B. Efron," *Statistical Science*, **11**, 214–219. 4

Cox, D. R. (1972), "Regression Models and Life Tables" (with discussion), *Journal of the Royal Statistical Society series B*, **34**, 187–220. 5

Davison, A. C. (1988), "Discussion of the Royal Statistical Society meeting on the bootstrap," *Journal of the Royal Statistical Society series B*, **50**, 356–357. 3

Davison, A. C. and Hinkley, D. V. (1997), *Bootstrap Methods and Their Application*, Cambridge: Cambridge University Press. 2, 6, 7

Davison, A. C., Hinkley, D. V., and Schechtman, E. (1986), "Efficient bootstrap simulation," *Biometrika*, **73**, 555–566. 3

Efron, B. (1981), "Censored data and the bootstrap," *Journal of the American Statistical Association*, **76**, 312–319. 4

—— (1992), "Jackknife-after-bootstrap standard errors and influence functions " (with discussion), *Journal of the Royal Statistical Society series B*, **54**, 83–127. 4

Efron, B. and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, New York: Chapman & Hall. 2

Hall, P. (1989), "On efficient bootstrap simulation," *Biometrika*, **76**, 613–617. 3

Johns, M. V. (1988), "Importance sampling for bootstrap confidence intervals," *Journal of the American Statistical Association*, **83**, 709–714. 3

Künsch, H. R. (1989), "The jackknife and bootstrap for general stationary observations," *Annals of Statistics*, **17**, 1217–1241. 5

Politis, D. N. and Romano, J. P. (1994), "The stationary bootstrap," *Journal of the American Statistical Association*, **89**, 1303–1313. 6

Angelo J. Canty
McMaster University, Hamilton, Ont, Canada
cantya@mcmaster.ca

Diagnostic Checking in Regression Relationships

by Achim Zeileis and Torsten Hothorn

Introduction

The classical linear regression model

$$y_i = x_i^\top \beta + u_i \qquad (i = 1, \ldots, n) \qquad (1)$$

is still one of the most popular tools for data analysis despite (or due to) its simple structure. Although it is appropriate in many situations, there are many pitfalls that might affect the quality of conclusions drawn from fitted models or might even lead to uninterpretable results. Some of these pitfalls that are considered especially important in applied econometrics are heteroskedasticity or serial correlation of the error terms, structural changes in the regression coefficients, nonlinearities, functional misspecification or omitted variables. Therefore, a rich variety of diagnostic tests for these situations have been developed in the econometrics community, a collection of which has been implemented in the packages **lmtest**

and **strucchange** covering the problems mentioned above.

These diagnostic tests are not only useful in econometrics but also in many other fields where linear regression is used, which we will demonstrate with an application from biostatistics. As Breiman (2001) argues it is important to assess the goodness-of-fit of data models, in particular not only using omnibus tests but tests designed for a certain direction of the alternative. These diagnostic checks do not have to be seen as pure significance procedures but also as an explorative tool to extract information about the structure of the data, especially in connection with residual plots or other diagnostic plots. As Brown et al. (1975) argue for the recursive CUSUM test, these procedures can "be regarded as yardsticks for the interpretation of data rather than leading to hard and fast decisions." Moreover, we will always be able to reject the null-hypothesis provided we have enough data at hand. The question is not whether the model is wrong (it always is!) but if the irregularities are serious.

The package **strucchange** implements a variety of procedures related to structural change of the regression coefficients and was already introduced in R news by Zeileis (2001) and described in more detail in Zeileis et al. (2002). Therefore, we will focus on the package **lmtest** in the following. Most of the tests and the datasets contained in the package are taken from the book of Krämer and Sonnberger (1986), which originally inspired us to write the package. Compared to the book, we implemented later versions of some tests and modern flexible interfaces for the procedures. Most of the tests are based on the OLS residuals of a linear model, which is specified by a formula argument. Instead of a formula a fitted model of class `"lm"` can also be supplied, which should work if the data are either contained in the object or still present in the workspace—however this is not encouraged. The full references for the tests can be found on the help pages of the respective function.

We present applications of the tests contained in **lmtest** to two different data sets: the first is a macroeconomic time series from the U.S. analysed by Stock and Watson (1996) and the second is data from a study on measurments of fetal mandible length discussed by Royston and Altman (1994).

U.S. macroeconomic data

Stock and Watson (1996) investigate the stability of 76 monthly macroeconomic time series from 1959 to 1993, of which we choose the department of commerce commodity price index time series `jocci` to illustrate the tests for heteroskedasticity and serial correlation. The data is treated with the same methodology as all other series considered by Stock and Watson (1996): they were transformed suitably (here by log first differences) and then an AR(6) model was fitted and analysed. The transformed series is denoted `dy` and is depicted together with a residual plot of the AR(6) model in Figure 1.

Not surprisingly, an autoregressive model is necessary as the series itself contains serial correlation, which can be shown by the Durbin-Watson test

```
R> data(jocci)
R> dwtest(dy ~ 1, data = jocci)

        Durbin-Watson test

data:  dy ~ 1
DW = 1.0581, p-value = < 2.2e-16
alternative hypothesis:
  true autocorrelation is greater than 0
```

or the Breusch-Godfrey test which also leads to a highly significant result. In the AR(6) model given by

```
R> ar6.model <-
     dy ~ dy1 + dy2 + dy3 + dy4 + dy5 + dy6
```

where the variables on the right hand side denote the lagged variables, there is no remaining serial correlation in the residuals:

```
R> bgtest(ar6.model, data = jocci)

        Breusch-Godfrey test for
        serial correlation of order 1

data:  ar6.model
LM test = 0.2, df = 1, p-value = 0.6547
```

The Durbin-Watson test is biased in dynamic models and should therefore not be applied.

The residual plot suggests that the variance of the error component increases over time, which is emphasized by all three tests for heteroskedasticity implemented in **lmtest**: the Breusch-Pagan test fits a linear regression model to the residuals and rejects if too much of the variance is explained by the auxiliary explanatory variables, which are here the squared lagged values:

```
R> var.model <-
     ~ I(dy1^2) + I(dy2^2) + I(dy3^2) +
       I(dy4^2) + I(dy5^2) + I(dy6^2)
R> bptest(ar6.model, var.model, data = jocci)

        studentized Breusch-Pagan test

data:  ar6.model
BP = 22.3771, df = 6, p-value = 0.001034
```

The Goldfeld-Quandt test `gqtest()` and the Harrison-McCabe test `hmctest()` also give highly significant p values. Whereas the Breusch-Pagan test and the Harrison-McCabe test do not assume a particular timing of the change of variance, the Goldfeld-Quandt test suffers from the same problem as the Chow test for a change of the regression

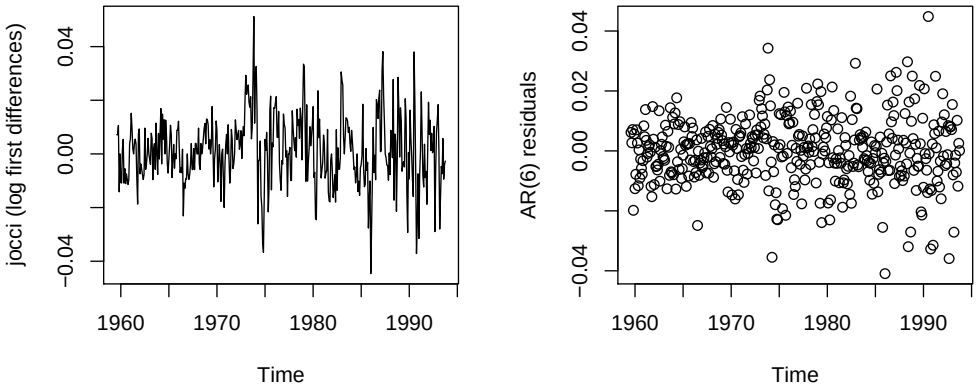

Figure 1: The jocci series and AR(6) residual plot

coefficients: the breakpoint has to be known in advance. By default it is taken to be after 50% of the observations, which leads to a significant result for the present series.

The mandible data

Royston and Altman (1994) discuss a linear regression model for data taken from a study of fetal mandible length by Chitty et al. (1993). The data comprises measurements of mandible `length` (in mm) and gestational `age` (in weeks) in 158 fetuses. The data (after log transformation) is depicted in Figure 2 together with the fitted values of a linear model `length ~ age` and a quadratic model `length ~ age + I(age^2)`.

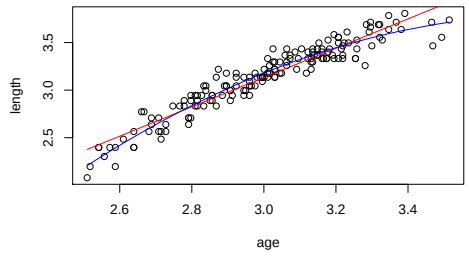

Figure 2: The mandible data

Although by merely visually inspecting the raw data or the residual plots in Figure 3 a quadratic model seems to be more appropriate, we will first fit a linear model for illustrating some tests for nonlinearity and misspecified functional form.

The suitable tests in **lmtest** are the Harvey-Collier test, which is essentially a t test of the recursive residuals (standardized one step prediction errors), and the Rainbow test. Both try to detect nonlinearities when the data is ordered with respect to a specific variable.

```
R> data(Mandible)
R> mandible <- log(Mandible)
R> harvtest(length ~ age, order.by = ~ age,
            data = mandible)
R> raintest(length ~ age, order.by = ~ age,
            data = mandible)
```

Both lead to highly significant results, suggesting that the model is not linear in `age`. Another appropriate procedure is the RESET test, which tests whether some auxiliary variables improve the fit significantly. By default the second and third powers of the fitted values are chosen:

```
R> reset(length ~ age, data = mandible)

        RESET test

data:  length ~ age
RESET = 26.1288, df1 = 2, df2 = 163,
  p-value = 1.436e-10
```

In our situation it would also be natural to consider powers of the regressor `age` as auxiliary variables

```
R> reset(length ~ age, power = 2,
         type = "regressor", data = mandible)

        RESET test

data:  length ~ age
RESET = 52.5486, df1 = 1, df2 = 164,
  p-value = 1.567e-11
```

which also gives a highly significant p value (higher powers do not have a significant influence). These results correspond to the better fit of the quadratic model which can both be seen in Figure 2 and 3. Although its residual plot does not look too suspicious several tests are able to reveal irregularities in this model as well. The Breusch-Pagan tests gives a p value of 0.043 and the Rainbow test gives

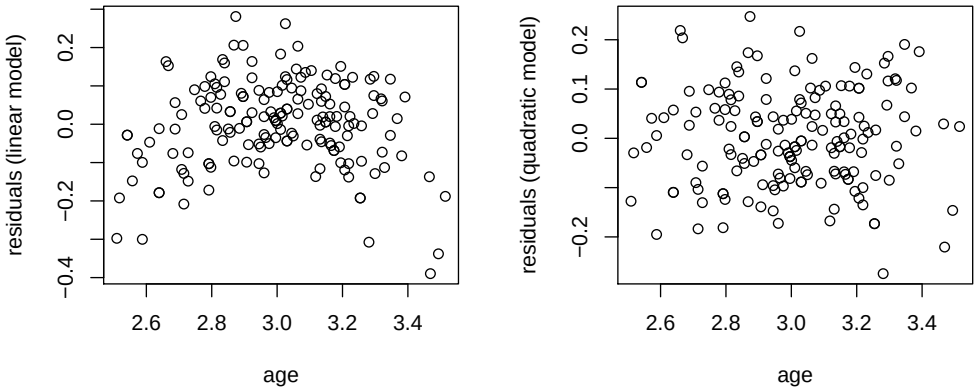

Figure 3: Residual plots for mandible models

```
R> raintest(length ~ age + I(age^2),
            order.by = ~ age, data = mandible)

        Rainbow test

data:  length ~ age + I(age^2)
Rain = 1.5818, df1 = 84, df2 = 80,
  p-value = 0.01995
```

and finally an sup$F$ test from the **strucchange** package would also reject the null hypothesis of stability at 10% level ($p = 0.064$) in favour of a breakpoint after about 90% of the observations. All three tests probably reflect that there is more variability in the edges (especially the right one) than in the middle which the model does not describe sufficiently.

Conclusions

We illustrated the usefulness of a collection of diagnostic tests for various situations of deviations from the assumptions of the classical linear regression model. We chose two fairly simple data sets—an econometric and a biometric application—to demonstrate how the tests work, but they are also particularly helpful to detect irregularities in regressions with a larger number of regressors.

Bibliography

L. Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16:199–231, 2001. 8

R. L. Brown, J. Durbin, and J. M. Evans. Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society*, B 37:149–163, 1975. 8

L. S. Chitty, S. Campbell, and D. G. Altman. Measurement of the fetal mandible – feasibility and construction of a centile chart. *Prenatal Diagnosis*, 13: 749–756, 1993. 9

W. Krämer and H. Sonnberger. *The Linear Regression Model Under Test*. Physica-Verlag, Heidelberg, 1986. 8

P. Royston and D. G. Altman. Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Applied Statistics*, 43:429–453, 1994. 8, 9

J. H. Stock and M. W. Watson. Evidence on structural instability in macroeconomic time series relations. *Journal of Business & Economic Statistics*, 14:11–30, 1996. 8

A. Zeileis. `strucchange`: Testing for structural change in linear regression relationships. *R News*, 1(3):8–11, September 2001. URL `http://cran.R-project.org/doc/Rnews/`. 8

A. Zeileis, F. Leisch, K. Hornik, and C. Kleiber. `strucchange`: An R package for testing for structural change in linear regression models. *Journal of Statistical Software*, 7(2):1–38, 2002. URL `http://www.jstatsoft.org/v07/i02/`. 8

Achim Zeileis
Institut für Statistik & Wahrscheinlichkeitstheorie,
Technische Universität Wien, Austria
`zeileis@ci.tuwien.ac.at`

Torsten Hothorn
Institut für Medizininformatik, Biometrie und Epidemiologie
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
`Torsten.Hothorn@rzmail.uni-erlangen.de`