Econometric Computing with HC and HAC Covariance Matrix Estimators

Achim Zeileis
Wirtschaftsuniversität Wien

Abstract

sandwich

Keywords: covariance matrix estimator, heteroskedasticity, autocorrelation, estimating functions, econometric computing, R.

1. Introduction

R Development Core Team (2004) Cribari-Neto and Zarkos (2003)

stress econometric computing

reusable components

covariance matrices not only as options to certain test but as stand-alone functions which can be plugged into various inference procedures

2. The linear regression model

To fix notations, we consider the linear regression model

$$y_i = x_i^\top \beta + u_i \qquad (i = 1, \dots, n), \tag{1}$$

with dependent variable y_i, k-dimensional regressor x_i with coefficient vector β and error term u_i. In the usual matrix notation comprising all n observations this can be formulated as $y = X\beta + u$.

In the general linear model, it is typically assumed that the errors have zero mean and variance $\mathsf{VAR}[u] = \Omega$. Under suitable regularity conditions (see e.g., Greene 1993), the coefficients β can be consistently estimated by ordinary least squares (OLS) giving the well-known OLS estimator $\hat{\beta}$ with corresponding OLS residuals $\hat{u}_i$:

$$\hat{\beta} = \left(X^\top X\right)^{-1} X^\top y \tag{2}$$

$$\hat{u} = (I_n - H)\, u = \left(I_n - X\left(X^\top X\right)^{-1} X^\top\right) u \tag{3}$$

where I_n is the n-dimensional identity matrix and H is usually called hat matrix. The estimates $\hat{\beta}$ are unbiased and are asymptotically normal with covariance matrix Ψ which is usally denoted in one of the two forms given below.

$$\Psi = \mathsf{VAR}[\hat{\beta}] = \left(X^\top X\right)^{-1} X^\top \Omega X \left(X^\top X\right)^{-1} \tag{4}$$

$$= \frac{1}{n}\left(\frac{1}{n} X^\top X\right)^{-1} \Phi \left(\frac{1}{n} X^\top X\right)^{-1} \tag{5}$$

where $\Phi = n^{-1} X^\top \Omega X^\top$ is essiantlly the covariance matrix of the estimating functions $V_i(\beta) = x_i(y_i - x_i^\top \beta)$. The estimating functions evaluated at the parameter estimates $\hat{V}_i = V_i(\hat{\beta})$ have then sum zero.

For doing inference in the linear regression model, it is essential to have a consistent estimator for Ψ—the most common inferential task being to test whether one of the coefficients β_j is zero which is is usually assessed using the t ratio $\beta_j / \sqrt{\hat{\Psi}_{jj}}$. What kind of estimator is used for Ψ depends on the assumptions about Ω that are used: in the classical linear model independent and homoskedastic errors with variance σ^2 are assumed yielding $\Omega = \sigma^2 I_n$ and $\Psi = \sigma^2 (X^\top X)^{-1}$ which can be easily estimated by plugging in the usual OLS estimate $\hat{\sigma}^2 = (n-k)^{-1} \sum_{i=1}^n \hat{u}_i^2$. But if the independence and/or homoskedasticity assumption is violated, inference based on the estimator for squerical errors $\hat{\Psi}_{\mathrm{const}} = \hat{\sigma}(X^\top X)^{-1}$ will be biased. Heteroskedasticity consistent (HC) estimators tackle this problem by plugging an estimate $\hat{\Omega}$ into (4) and heteroskedasiticy and autocorrelation consistent (HAC) estimators by plugging an estimate $\hat{\Phi}$ into (5). These classes of estimators and their implementation are described in the following section.

3. Estimating covariance matrices

3.1. Dealing with heteroskedasticity

$\hat{\Psi}_{\mathrm{HC}}$ plugging in a $\hat{\Omega} = \mathrm{diag}(\omega_1, \dots, \omega_n)$

$$\begin{aligned}
\text{const}: \quad \omega_i &= \hat{\sigma}^2 \\
\text{HC0}: \quad \omega_i &= \hat{u}_i^2 \\
\text{HC1}: \quad \omega_i &= \frac{n}{n-k}\,\hat{u}_i^2 \\
\text{HC2}: \quad \omega_i &= \frac{\hat{u}_i^2}{1-h_i} \\
\text{HC3}: \quad \omega_i &= \frac{\hat{u}_i^2}{(1-h_i)^2} \\
\text{HC4}: \quad \omega_i &= \frac{\hat{u}_i^2}{(1-h_i)^{\delta_i}}
\end{aligned}$$

where $h_i = H_{ii}$ are the diagonal elements of the hat matrix and $\delta_i = \min\{4, h_i/\bar{h}\}$.

```
vcovHC(lmobject, omega = NULL, type = "HC3",
  order.by = NULL)
```

```
"HC3", "const", "HC", "HC0", "HC1", "HC2", "HC4"
omega(residuals, diaghat, df)
```

White (1980) MacKinnon and White (1985) Long and Ervin (2000) Cribari-Neto (2004)

3.2. Dealing with autocorrelation

$\hat{\Psi}_{\mathrm{HAC}}$ plugging in a $\hat{\Phi}$ with

$$\hat{\Phi} = \frac{1}{n} \sum_{i,j=1}^{n} w_{|i-j|} \hat{V}_i \hat{V}_j^\top \qquad (6)$$

weights w_k $(k = 0, \dots, n-1)$

finite sample correction $n/(n-k)$

Newey-West or Bartlett kernel

$$w_i = 1 - \frac{i}{L+1} \qquad (7)$$

where L is the maximum lag, other weights are zero. In terms of a generic bandwidth B usually formulated as $B = L + 1$.

Quadratic spectral kernel

$$w_i = \frac{3}{x^2}\left(\frac{\sin(x)}{x} - \cos(x)\right) \tag{8}$$

where $x = 6/5\pi \cdot i/B$ and B is again a bandwidth parameter.

```
vcovHAC(lmobject, weights,
  order.by = NULL, prewhite = FALSE, adjust = TRUE, sandwich = TRUE)
```

```
weights(x, order.by, prewhite, ar.method, data)
```

Newey and West (1987) Andrews (1991) Andrews and Monahan (1992) Lumley and Heagerty (1999)

4. Applications and illustrations

t ratio $\beta_j/\sqrt{\hat{\Psi}_{jj}}$

For computing p values the asymptotic normal distribution or the t distribution with $n-k$ degrees of freedom are used.

4.1. Testing coefficients in cross-sectional data

Greene (1993) Cribari-Neto (2004) Zeileis and Hothorn (2002) Fox (2002)

```
R> data(PublicSchools)
R> ps <- na.omit(PublicSchools)
R> ps$Income <- ps$Income * 1e-04
```

```
R> fm.ps <- lm(Expenditure ~ Income + I(Income^2), data = ps)
```

```
R> coeftest(fm.ps, df = Inf, vcov = vcovHC(fm.ps, type = "HC0"))

z test of coefficients of "lm" object 'fm.ps':

             Estimate Std. Error z value Pr(>|z|)
(Intercept)   832.91     460.89  1.8072  0.07073 .
Income      -1834.20    1243.04 -1.4756  0.14006
I(Income^2)  1587.04     829.99  1.9121  0.05586 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
R> coeftest(fm.ps, df = Inf, vcov = vcovHC(fm.ps, type = "HC4"))

z test of coefficients of "lm" object 'fm.ps':

             Estimate Std. Error z value Pr(>|z|)
(Intercept)   832.91    3008.01  0.2769   0.7819
Income      -1834.20    8183.19 -0.2241   0.8226
I(Income^2)  1587.04    5488.93  0.2891   0.7725
```

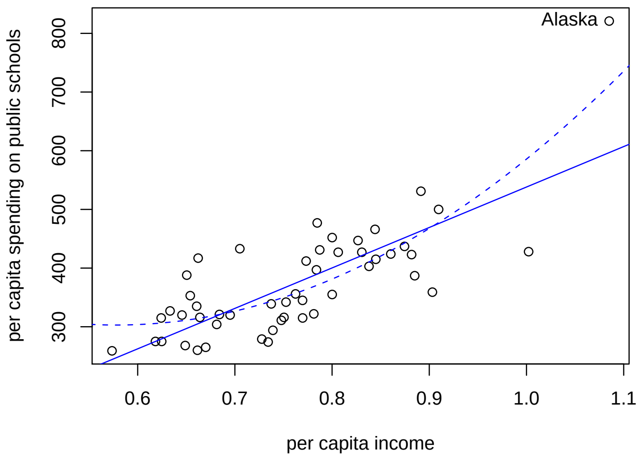

Figure 1: Expenditure on public schools and income with fitted models

```
vcovHC(fm.ps, type = "HC0")
```

4.2. Testing coefficients in time-series data

Greene (1993)

```
R> data(Investment)

R> fm.inv <- lm(RealInv ~ RealGNP + RealInt, data = Investment)

R> coeftest(fm.inv, df = Inf, vcov = NeweyWest(fm.inv, lag = 4))

z test of coefficients of "lm" object 'fm.inv':

              Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.533601  18.958298 -0.6611   0.5085
RealGNP       0.169136   0.016751 10.0972   <2e-16 ***
RealInt      -1.001438   3.342375 -0.2996   0.7645
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.3. Testing and dating structural changes in the presence of heteroskedasticity and autocorrelation

If the first component if x_i is equal to unity

$$\sup_{j=1,\dots,n} \left| \frac{1}{\sqrt{n\,\hat{\Phi}_{11}}} \sum_{i=1}^{j} \hat{u}_i \right|. \tag{9}$$

Bai and Perron (2003) Andrews (1993) Ploberger and Krämer (1992)
Zeileis, Leisch, Hornik, and Kleiber (2002) Zeileis (2004)

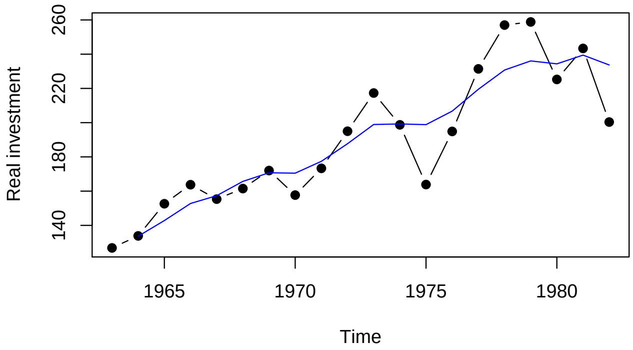

Figure 2: Investment equation data with fitted model

```
R> data(RealInt)
```

```
R> ocus <- gefp(RealInt ~ 1, fit = lm, vcov = kernHAC)
```

`plot(ocus)`, `sctest(ocus)`

```
R> fs <- Fstats(RealInt ~ 1, vcov = kernHAC)
```

`sctest(fs)`, `plot(fs)`

```
R> bp <- breakpoints(RealInt ~ 1)
R> confint(bp, vcov = kernHAC)

         Confidence intervals for breakpoints
         of optimal 3-segment partition:

Call:
confint.breakpointsfull(object = bp, vcov = kernHAC)

Breakpoints at observation number:
  2.5 % breakpoints 97.5 %
1    37          47     48
2    77          79     81

Corresponding to breakdates:
  2.5 %   breakpoints 97.5 %
1 1970(1) 1972(3)     1972(4)
2 1980(1) 1980(3)     1981(1)
```

5. Summary

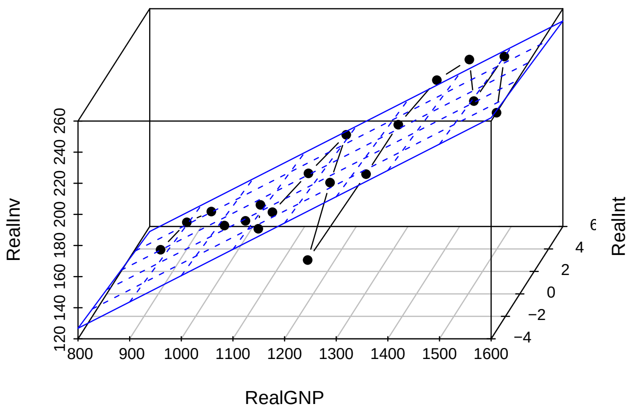

Figure 3: Investment equation data with fitted model (3D)

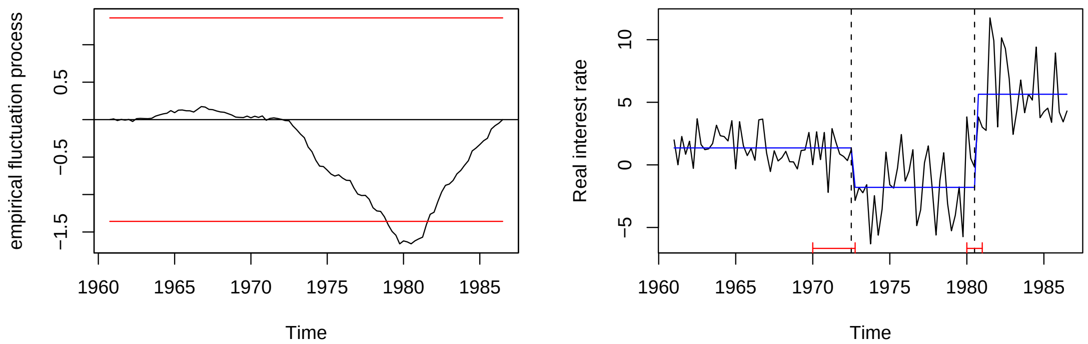

Figure 4: OLS-based CUSUM test (left) and fitted model (right) for real interest data

A. R code

A.1. Testing coefficients in cross-sectional data

Load public schools data, omit `NA` in Wisconsin and scale income:

```
data(PublicSchools)
ps <- na.omit(PublicSchools)
ps$Income <- ps$Income * 0.0001
```

Fit quadratic regression model:

```
fm.ps <- lm(Expenditure ~ Income + I(Income^2), data = ps)
```

Compare standard errors:

```
sqrt(diag(vcov(fm.ps)))
sqrt(diag(vcovHC(fm.ps, type = "const")))
sqrt(diag(vcovHC(fm.ps, type = "HC0")))
sqrt(diag(vcovHC(fm.ps, type = "HC3")))
sqrt(diag(vcovHC(fm.ps, type = "HC4")))
```

Test coefficient of quadratic term:

```
coeftest(fm.ps, df = Inf, vcov = vcovHC(fm.ps, type = "HC0"))
coeftest(fm.ps, df = Inf, vcov = vcovHC(fm.ps, type = "HC4"))
```

Visualization:

```
plot(Expenditure ~ Income, data = ps,
  xlab = "per capita income",
  ylab = "per capita spending on public schools")
inc <- seq(0.5, 1.2, by = 0.001)
lines(inc, predict(fm.ps, data.frame(Income = inc)), col = 4, lty = 2)
fm.ps2 <- lm(Expenditure ~ Income, data = ps)
abline(fm.ps2, col = 4)
text(ps[2,2], ps[2,1], rownames(ps)[2], pos = 2)
```

A.2. Testing coefficients in time-series data

Load investment equation data:

```
data(Investment)
```

Fit regression model:

```
fm.inv <- lm(RealInv ~ RealGNP + RealInt, data = Investment)
```

Test coefficients using Newey-West HAC estimator:

```
coeftest(fm.inv, df = Inf, vcov = NeweyWest(fm.inv, lag = 4))
```

Visualization:

```
plot(Investment[, "RealInv"], type = "b", pch = 19, ylab = "Real investment")
lines(ts(fitted(fm.inv), start = 1964), col = 4)
```

3-d visualization:

```
library(scatterplot3d)
s3d <- scatterplot3d(Investment[,c(5,7,6)],
  type = "b", angle = 65, scale.y = 1, pch = 16)
s3d$plane3d(fm.inv, lty.box = "solid", col = 4)
```

A.3. Testing and dating structural changes in the presence of heteroskedasticity and autocorrelation

Load real interest series:

```
data(RealInt)
```

OLS-based CUSUM test with quadratic spectral kernel HAC estimate:

```
ocus <- gefp(RealInt ~ 1, fit = lm, vcov = kernHAC)
plot(ocus, aggregate = FALSE)
sctest(ocus)
```

sup$F$ test with quadratic spectral kernel HAC estimate:

```
fs <- Fstats(RealInt ~ 1, vcov = kernHAC)
plot(fs)
sctest(fs)
```

Breakpoint estimation and confidence intervals with quadratic spectral kernel HAC estimate:

```
bp <- breakpoints(RealInt ~ 1)
confint(bp, vcov = kernHAC)
```

Visualization:

```
plot(RealInt, ylab = "Real interest rate")
lines(ts(fitted(bp), start = start(RealInt), freq = 4), col = 4)
lines(confint(bp, vcov = kernHAC))
```

References

Andrews DWK (1991). "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation." *Econometrica*, **59**, 817–858.

Andrews DWK (1993). "Tests for Parameter Instability and Structural Change With Unknown Change Point." *Econometrica*, **61**, 821–856.

Andrews DWK, Monahan JC (1992). "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator." *Econometrica*, **60**(4), 953–966.

Bai J, Perron P (2003). "Computation and Analysis of Multiple Structural Change Models." *Journal of Applied Econometrics*, **18**, 1–22.

Cribari-Neto F (2004). "Asymptotic Inference Under Heteroskedasticity of Unknown Form." *Computational Statistics & Data Analysis*, **45**, 215–233.

Cribari-Neto F, Zarkos SG (2003). "Econometric and Statistical Computing Using Ox." *Computational Economics*, **21**, 277–295.

Fox J (2002). *An R and S-PLUS Companion to Applied Regression*. Sage Publications, Thousand Oaks, CA.

Greene WH (1993). *Econometric Analysis*. Macmillan Publishing Company, New York, 2nd edition.

Long JS, Ervin LH (2000). "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model." *The American Statistician*, **54**, 217–224.

Lumley T, Heagerty P (1999). "Weighted Empirical Adaptive Variance Estimators for Correlated Data Regression." *Journal of the Royal Statistical Society B*, **61**, 459–477.

MacKinnon JG, White H (1985). "Some Heteroskedasticity-consistent Covariance Matrix Estimators with Improved Finite Sample Properties." *Journal of Econometrics*, **29**, 305–325.

Newey WK, West KD (1987). "A Simple, Positive-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica*, **55**, 703–708.

Ploberger W, Krämer W (1992). "The CUSUM Test With OLS Residuals." *Econometrica*, **60**, 271–285.

R Development Core Team (2004). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL http://www.R-project.org/.

White H (1980). "A Heteroskedasticity-consistent Covariance Matrix and a Direct Test for Heteroskedasticity." *Econometrica*, **48**, 817–838.

Zeileis A (2004). "Implementing a Class of Structural Change Tests: An Econometric Computing Approach." *Report 7*, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Research Report Series. URL http://epub.wu-wien.ac.at/.

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, Leisch F, Hornik K, Kleiber C (2002). "**strucchange**: An R Package for Testing for Structural Change in Linear Regression Models." *Journal of Statistical Software*, **7**(2), 1–38. URL http://www.jstatsoft.org/v07/i02/.