



Applied and Reproducible Econometrics with R

Achim Zeileis

<http://eeecon.uibk.ac.at/~zeileis/>

Overview

- R and econometrics
- AER: Book and package
- Illustrations
 - Demand for economics journals
 - Mobility in educational attainment
 - Forensic econometrics of growth
- Excursions
 - Object orientation
 - Reproducible research

R and econometrics

- Econometric *theory* always had large impact on statistical research.
- However, econometrics lagged behind in embracing *computational methods* and *software* as an intrinsic part of research.
- Traditionally, rely on software provided by commercial publishers, e.g., Stata, EViews, or programming environments such as GAUSS, Ox, among others.
- Recently, software development/dissemination are increasingly regarded as natural concomitants of econometric research.
- Hence also increasing interest in econometrics with R.

R and econometrics

Question: Why R?

Answers:

- Free and platform independent: Important for teaching.
- Open source: Important for reproducible research.
- Flexible, object-oriented programming environment.
- Superior graphics and extensive methods for (exploratory) data analysis.
- Tools for reproducibility: Packaging of data/code/documentation, Sweave() for “dynamic” documents, . . .

R and econometrics

Challenges:

- Differences in language and terminology, e.g.,
 - factor vs. dummy variable(s),
 - generalized linear model (GLM) vs. logit, probit, Poisson regression.
- Different workflow: Command line interface, functional language, object-oriented approach.
- Some basic econometric methods scattered across various CRAN packages. Some of these still relatively new, e.g., the following have been published in JSS.
 - **gmm**: Generalized method of moments.
 - **np**: Nonparametric kernel methods.
 - **plm**, **splm**: Linear models for (spatial) panel data.
 - **pscl**: Zero-inflated and hurdle models for count data.
 - **vars**: Vector autoregression and error correction models.

AER: Book and package

Book: Kleiber & Zeileis, *Applied Econometrics with R*, Springer-Verlag.

Aims:

- Introduction to econometric computing with R.
- Not an econometrics book, rather “second book” for a course in econometrics.
- Bridge differences in jargon, explain some statistical concepts.
- Provide overview of relevant/useful R packages.

R package: <http://CRAN.R-project.org/package=AER>.

- Demos: Full R code from the book.
- Data: More than 100 data sets from leading applied econometrics journals and popular econometrics books.
- Examples: Replication code for many examples from textbooks of Baltagi, Greene, Stock & Watson, Winkelmann & Boes, . . .

Illustration: Demand for economics journals

Data: From Stock & Watson (2007), originally collected by T. Bergstrom, on subscriptions to 180 economics journals at US libraries, for the year 2000.

10 variables are provided including:

- `subs` – number of library subscriptions,
- `price` – library subscription price,
- `citations` – total number of citations,

and other information such as number of pages, founding year, characters per page, etc.

Of interest: Relation between demand and price for economics journals. Price is measured as price per citation.

Illustration: Demand for economics journals

Load data and obtain basic information:

```
R> library("AER")  
R> data("Journals", package = "AER")  
R> dim(Journals)
```

```
[1] 180  10
```

```
R> names(Journals)
```

```
[1] "title"          "publisher"      "society"        "price"  
[5] "pages"          "charpp"         "citations"     "foundingyear"  
[9] "subs"           "field"
```

Plot variables of interest:

```
R> plot(log(subs) ~ log(price/citations), data = Journals)
```

Fit linear regression model:

```
R> j_lm <- lm(log(subs) ~ log(price/citations), data = Journals)  
R> abline(j_lm)
```


Illustration: Demand for economics journals

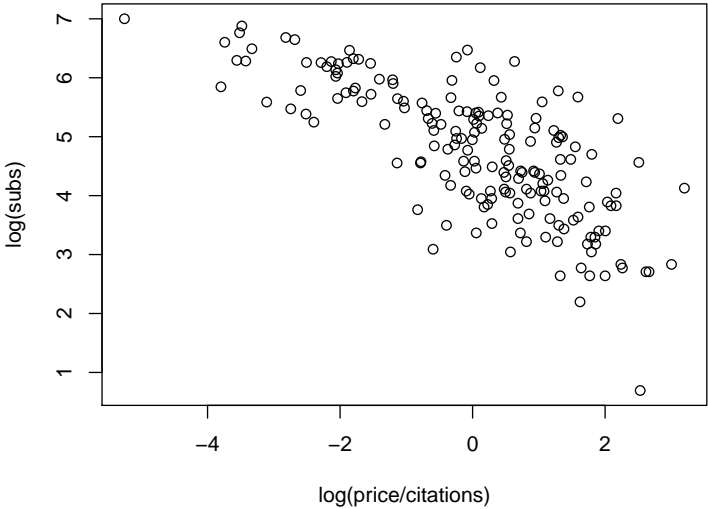


Illustration: Demand for economics journals

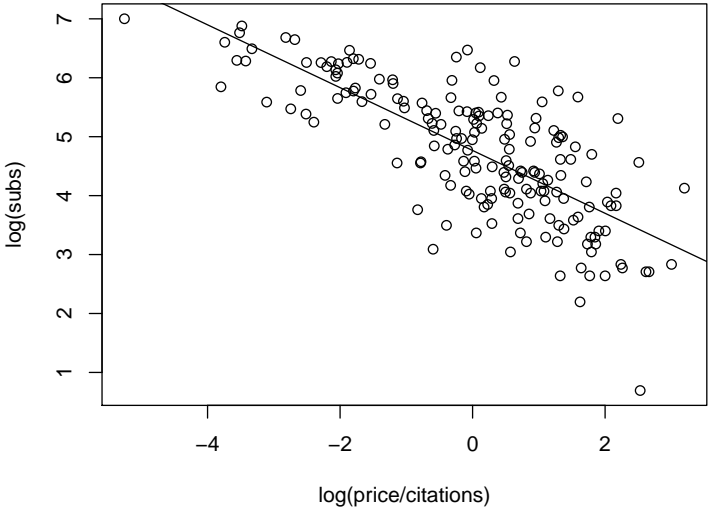


Illustration: Demand for economics journals

```
R> summary(j_lm)
```

```
Call:
```

```
lm(formula = log(subs) ~ log(price/citations), data = Journals)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-2.7248	-0.5361	0.0372	0.4662	1.8481

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.7662	0.0559	85.2	<2e-16
log(price/citations)	-0.5331	0.0356	-15.0	<2e-16

```
Residual standard error: 0.75 on 178 degrees of freedom
```

```
Multiple R-squared: 0.557, Adjusted R-squared: 0.555
```

```
F-statistic: 224 on 1 and 178 DF, p-value: <2e-16
```

Excursion: Object orientation

In most other econometrics packages: An analysis leads to a large amount of output containing information on estimation, model diagnostics, specification tests, etc.

In R:

- Analysis is broken down into a series of steps.
- Intermediate results are stored in *objects*.
- Minimal output at each step (often none).
- Objects can be manipulated and interrogated to obtain the information required (e.g., `print()`, `summary()`, `plot()`).

Fundamental design principle: “Everything is an object.”

Examples: Vectors and matrices are objects, but also fitted model objects, functions, and even function calls \Rightarrow facilitates programming tasks.

Excursion: Object orientation

```
R> coef(j_lm)
```

```
      (Intercept) log(price/citations)
      4.7662          -0.5331
```

```
R> vcov(j_lm)
```

```
              (Intercept) log(price/citations)
(Intercept)      3.126e-03      -6.144e-05
log(price/citations) -6.144e-05      1.268e-03
```

```
R> logLik(j_lm)
```

```
'log Lik.' -202.6 (df=3)
```

Excursion: Object orientation

<code>print()</code>	simple printed display with coefficients
<code>summary()</code>	standard regression summary
<code>plot()</code>	diagnostic plots
<hr/>	
<code>coef()</code>	extract coefficients
<code>vcov()</code>	associated covariance matrix
<code>predict()</code>	(different types of) predictions for new data
<code>fitted()</code>	fitted values for observed data
<code>residuals()</code>	extract (different types of) residuals
<hr/>	
<code>terms()</code>	extract terms
<code>model.matrix()</code>	extract model matrix (or matrices)
<code>nobs()</code>	extract number of observations
<code>df.residual()</code>	extract residual degrees of freedom
<code>logLik()</code>	extract fitted log-likelihood
<hr/>	

Excursion: Object orientation

Furthermore: “Smart” generics can rely on suitable methods such as `coef()`, `vcov()`, `logLik()`, `nobs()`, etc.

<code>confint()</code>	confidence intervals
<code>AIC()</code> , <code>BIC()</code>	information criteria (AIC, BIC, ...)
<code>coeftest()</code>	partial Wald tests of coefficients (lmtest)
<code>waldtest()</code>	Wald tests of nested models (lmtest)
<code>linearHypothesis()</code>	Wald tests of linear hypotheses (car)
<code>lrtest()</code>	likelihood ratio tests of nested models (lmtest)
<code>sandwich()</code> , ...	sandwich/HC/HAC estimators of covariance matrices (sandwich)

Excursion: Object orientation

```
R> confint(j_lm)
```

```
                2.5 %  97.5 %  
(Intercept)      4.6559  4.8765  
log(price/citations) -0.6033 -0.4628
```

```
R> linearHypothesis(j_lm, "log(price/citations) = -0.5")
```

Linear hypothesis test

Hypothesis:

$\log(\text{price}/\text{citations}) = -0.5$

Model 1: restricted model

Model 2: $\log(\text{subs}) \sim \log(\text{price}/\text{citations})$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	179	100				
2	178	100	1	0.484	0.86	0.35

Excursion: Object orientation

```
R> coeftest(j_lm)
```

```
t test of coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.7662	0.0559	85.2	<2e-16
log(price/citations)	-0.5331	0.0356	-15.0	<2e-16

```
R> coeftest(j_lm, vcov = sandwich)
```

```
t test of coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.7662	0.0550	86.7	<2e-16
log(price/citations)	-0.5331	0.0338	-15.8	<2e-16

Illustration: Mobility in educational attainment

Data: From Winkelmann & Boes (2009). Cross-section of 675 14-year old children taken from the German Socio-Economic Panel (GSOEP), 1994–2002.

Model: Secondary school choice (Hauptschule, Realschule, Gymnasium) explained by mother's education (in years), correcting for mother's employment level, household income and size (in logs).

Comparison: Multinomial logit (MNL) and ordered logit model (OLM).

In R:

- MNL: `multinom()` from package **nnet** (because neural networks have same fitting algorithm).
- OLM: `polr()` from package **MASS** (because model is also known as proportional odds logistic regression in the statistics literature).
- Couple with **effects** package for visualizing predicted probabilities.

Illustration: Mobility in educational attainment

Exploratory display:

```
R> data("GSOEP9402", package = "AER")  
R> plot(school ~ medication, data = GSOEP9402,  
+       breaks = c(7, 9, 10.5, 11.5, 12.5, 15, 18))
```

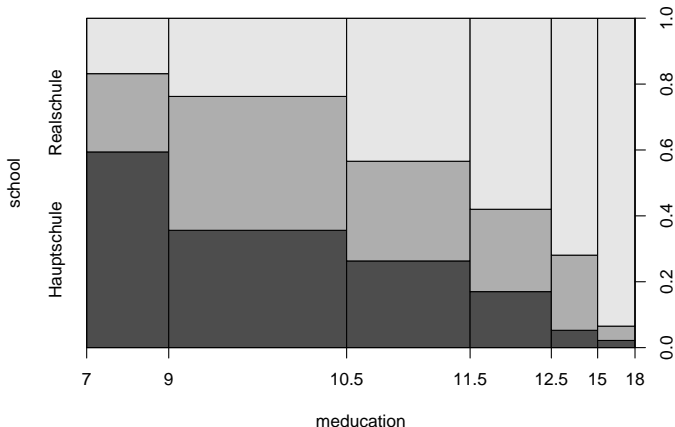


Illustration: Mobility in educational attainment

Model formula:

```
R> f <- school ~ meducation + memployment + log(income) + log(size)
```

Multinomial logit:

```
R> library("nnet")
```

```
R> gsoep_mnl <- multinom(f, data = GSOEP9402)
```

Ordered logit:

```
R> library("MASS")
```

```
R> gsoep_olm <- polr(f, data = GSOEP9402, Hess = TRUE)
```

Comparison:

```
R> AIC(gsoep_mnl, gsoep_olm)
```

	df	AIC
gsoep_mnl	12	1279
gsoep_olm	7	1277

Illustration: Mobility in educational attainment

Selected model:

```
R> coeftest(gsoep_olm)
```

z test of coefficients:

	Estimate	Std. Error	z value	Pr(> z)
meducation	0.4766	0.0513	9.28	< 2e-16
memploymentparttime	0.6932	0.2452	2.83	0.0047
memploymentnone	0.8124	0.2507	3.24	0.0012
log(income)	1.0392	0.1868	5.56	2.6e-08
log(size)	-1.2550	0.3230	-3.89	0.0001
Hauptschule Realschule	14.6720	1.9332	7.59	3.2e-14
Realschule Gymnasium	16.2233	1.9532	8.31	< 2e-16

Visualization:

```
R> library("effects")
```

```
R> plot(effect("meducation", gsoep_mnl), confint = FALSE)
```

```
R> plot(effect("meducation", gsoep_olm), confint = FALSE)
```

Illustration: Mobility in educational attainment

meducation effect plot

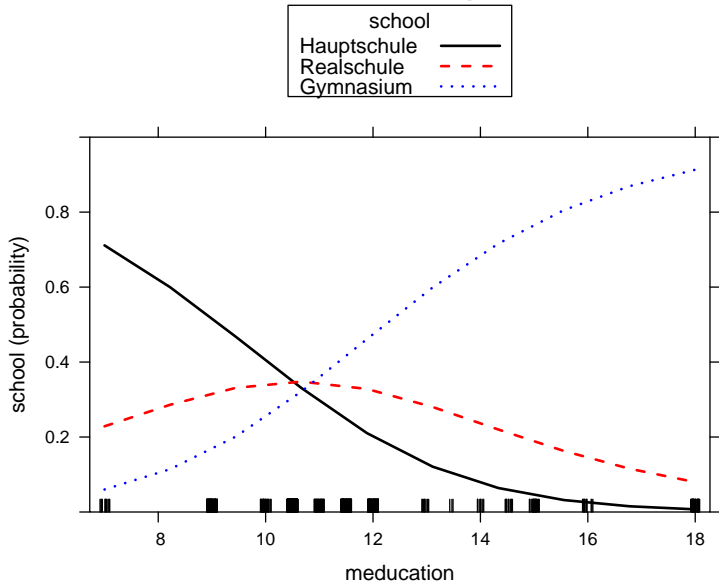
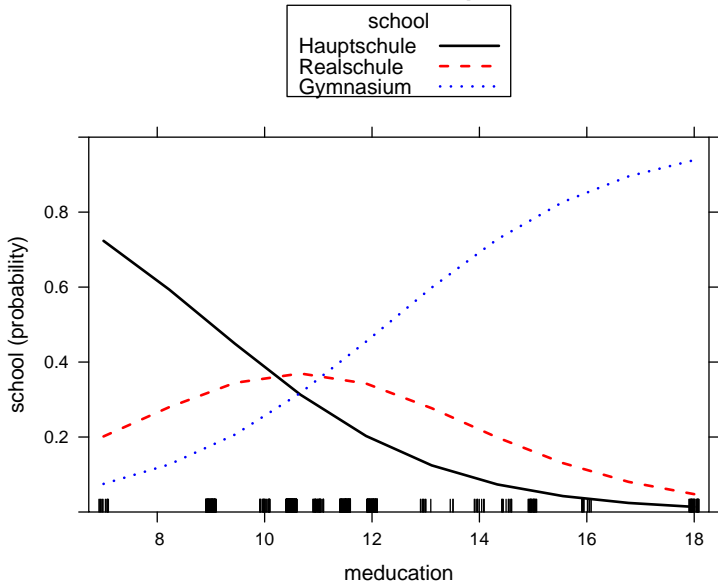


Illustration: Mobility in educational attainment

meducation effect plot



Excursion: Reproducible research

Idea: Facilitate reproducibility by keeping text and code in sync within the same document.

In R: `Sweave()` combines R code with \LaTeX text (or HTML, Markdown, ODF, ...).

- Single `.Rnw` file contains both text and code.
- *Tangling*: Extract code.
- *Weaving*: Execute code to produce all numbers, tables, figures, ...
- Optionally, R input and output can be shown or hidden.
- Results in “dynamic” or “revivable” documents.

Here: These slides are actually produced using `Sweave()`.

Illustration: Forensic econometrics of growth

Investigation: Cross-country growth behavior based on extended Solow model.

- Durlauf and Johnson (1995, *Journal of Applied Econometrics*) extend analysis by Mankiw, Romer, Weil (1992, *The Quarterly Journal of Economics*).
- Of interest: Output (GDP per capita) growth from 1960 to 1985 for 98 non-oil-producing countries.
- Variables: Real GDP per capita; fraction of real GDP devoted to investment; population growth; fraction of population in secondary schools; and adult literacy rate.
- Data taken from MRW. DJ added literacy rate. Available as `data.dj` in JAE data archive.

Models: OLS regressions for full sample and breaks based on initial output and literacy.

Illustration: Forensic econometrics of growth

Dependent variable: $\log(Y/L)_{i,1985} - \log(Y/L)_{i,1960}$.

	Full sample	$(Y/L)_{i,1960} < 1950$ $LR_{i,1960} < 54\%$	$(Y/L)_{i,1960} \geq 1950$ $LR_{i,1960} \geq 54\%$
Observations	98	42	42
Constant	3.040 (0.831)	1.400 (1.850)	0.450 (0.723)
$\log(Y/L)_{i,1960}$	-0.289 (0.062)	-0.444 (0.157)	-0.434 (0.085)
$\log(I/Y)_i$	0.524 (0.087)	0.310 (0.114)	0.689 (0.170)
$\log(n + 0.05)_i$	-0.505 (0.288)	-0.379 (0.468)	-0.545 (0.283)
$\log(SCHOOL)_i$	0.233 (0.060)	0.209 (0.094)	0.114 (0.164)

Illustration: Forensic econometrics of growth

Replication: Data is available from JAE archive, and OLS regression should be trivial ... right?

Data: Read, code missing values, and select non-oil countries.

```
R> dj <- read.table("data.dj", header = TRUE,  
+   na.strings = c("-999.0", "-999.00"))  
R> dj <- subset(dj, NONOIL == 1)
```

Model: R formula (converting percentages to fractions).

```
R> f1 <- I(log(GDP85) - log(GDP60)) ~ log(GDP60) +  
+   log(IONY/100) + log(POPGR0/100 + 0.05) + log(SCHOOL/100)
```

Regression: OLS fit for full sample and subsamples.

```
R> mrw <- lm(f1, data = dj)  
R> sub1 <- lm(f1, data = dj, subset = GDP60 < 1950 & LIT60 < 54)  
R> sub2 <- lm(f1, data = dj, subset = GDP60 >= 1950 & LIT60 >= 54)
```

Illustration: Forensic econometrics of growth

Full sample results: Success! Only minor deviations.

```
R> mrw <- lm(f1, data = dj)
R> coeftest(mrw)
```

	Durlauf & Johnson	Replication
Observations	98	98
Constant	3.040 (0.831)	3.022 (0.827)
$\log(Y/L)_{i,1960}$	-0.289 (0.062)	-0.288 (0.062)
$\log(I/Y)_i$	0.524 (0.087)	0.524 (0.087)
$\log(n + 0.05)_i$	-0.505 (0.288)	-0.506 (0.289)
$\log(SCHOOL)_i$	0.233 (0.060)	0.231 (0.059)

Illustration: Forensic econometrics of growth

Subsample results: Failure! Not even sample size is correct.

```
R> sub2 <- lm(f1, data = dj, subset = GDP60 >= 1950 & LIT60 >= 54)
R> coeftest(sub2)
```

	Durlauf & Johnson	Replication
Observations	42	39
Constant	0.450 (0.723)	3.952 (1.337)
$\log(Y/L)_{i,1960}$	-0.434 (0.085)	-0.425 (0.104)
$\log(I/Y)_i$	0.689 (0.170)	0.653 (0.187)
$\log(n + 0.05)_i$	-0.545 (0.283)	-0.587 (0.361)
$\log(SCHOOL)_i$	0.114 (0.164)	0.137 (0.180)

Illustration: Forensic econometrics of growth

Problem 1: Grid search plus educated guessing leads to different breaks.

```
R> sub2b <- lm(f1, data = dj, subset = GDP60 >= 1800 & LIT60 >= 50)
R> coeftest(sub2b)
```

	Durlauf & Johnson	Replication
Observations	42	42
Constant	0.450 (0.723)	4.147 (1.230)
$\log(Y/L)_{i,1960}$	-0.434 (0.085)	-0.435 (0.096)
$\log(I/Y)_i$	0.689 (0.170)	0.689 (0.178)
$\log(n + 0.05)_i$	-0.545 (0.283)	-0.545 (0.345)
$\log(SCHOOL)_i$	0.114 (0.164)	0.114 (0.171)

Illustration: Forensic econometrics of growth

Problem 2: Population growth and schooling not fractions but percent.

```
R> sub2c <- update(sub2b, . ~ log(GDP60) +  
+   log(IONY) + log(POPGRO/100 + 0.05) + log(SCHOOL))
```

	Durlauf & Johnson	Replication
Observations	42	42
Constant	0.450 (0.723)	0.450 (0.899)
$\log(Y/L)_{i,1960}$	-0.434 (0.085)	-0.435 (0.096)
$\log(I/Y)_i$	0.689 (0.170)	0.689 (0.178)
$\log(n + 0.05)_i$	-0.545 (0.283)	-0.545 (0.345)
$\log(SCHOOL)_i$	0.114 (0.164)	0.114 (0.171)

Illustration: Forensic econometrics of growth

Problem 3: Robust sandwich standard errors.

```
R> coeftest(sub2c, vcov = sandwich)
```

	Durlauf & Johnson	Replication
Observations	42	42
Constant	0.450 (0.723)	0.450 (0.723)
$\log(Y/L)_{i,1960}$	-0.434 (0.085)	-0.435 (0.085)
$\log(I/Y)_i$	0.689 (0.170)	0.689 (0.170)
$\log(n + 0.05)_i$	-0.545 (0.283)	-0.545 (0.283)
$\log(SCHOOL)_i$	0.114 (0.164)	0.114 (0.164)

Illustration: Forensic econometrics of growth

Summary:

- Cutoffs actually used did not match those indicated.
- Usage of standard errors inconsistent.
- Scaling of variables (and hence intercepts) inconsistent.
- Other models in DJ paper: Similar problems, and some inference not reproducible at all.

Implications:

- Casts doubt results. (Even though – in this case, so far – qualitative results remain unchanged.)
- Very hard to track down without original code.
- Might have been impossible for less standard models.
- *Hence*: Keep analysis and documentation/manuscript in sync. Provide replication code even for simple things and details.

Summary

- R system is a free open-source environment with tools for reproducible research.
- Wide variety of econometric methods already available.
- Workflow, development process, and terminology may sometimes be unfamiliar to econometricians.
- Many resources available to bridge differences: Examples, demos, textbooks, software papers, . . .

References

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

URL <http://CRAN.R-project.org/package=AER>

Koenker R, Zeileis A (2009). “On Reproducible Econometric Research.” *Journal of Applied Econometrics*, **24**(5), 833–847.

doi:10.1002/jae.1083

Zeileis A, Koenker R (2008). “Econometrics in R: Past, Present, and Future.” *Journal of Statistical Software*, **27**(1), 1–5.

URL <http://www.jstatsoft.org/v27/i01/>