Universität Innsbruck

Applied Econometrics with R

Christian Kleiber, 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.
- Excursion: Object orientation.
- Further information.

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.,
  - **gmm**: Generalized method of moments.
  - **np**: Nonparametric kernel methods.
  - **plm**: Linear models for panel data.
  - **pscl**: Zero-inflated and hurdle models for count data.
  - **vars**: Vector autoregression and error correction models.

AER: Book and package

Book: *Applied Econometrics with R*, Springer-Verlag, New York.

Aims:

- Introduction to econometric computing with R.
- Not an econometrics book, rather "second book" for a course in econometrics.
- Emphasize applications/practical issues with challenging data sets.
- Bridge differences in jargon, explain some statistical concepts.
- Provide overview of relevant/useful R packages.

AER: Book and package

Basics: Programming, data management, object orientation, graphics.

Linear regression: OLS, systems of equations, panel models.

Diagnostics and alternative methods of regression: Quantile and resistant regression, sandwich covariances, diagnostic tests.

Models of microeconometrics: Logit, probit, Poisson regression (as GLMs), tobit, further models for count data (ZIP, hurdle), censored, or limited responses.

Time series: (S)ARIMA(X), unit roots and cointegration, structural change, ARCH models, structural time series models.

Programming your own analysis: Simulations, bootstrapping, likelihood optimization, reproducibility.

AER: Book and package

Package: `http://CRAN.R-project.org/package=AER`.

Contents:

- 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
  - Baltagi: *Econometrics*.
  - Cameron & Trivedi: *Regression Analysis of Count Data*.
  - Greene: *Econometric Analysis*.
  - Stock & Watson: *Introduction to Econometrics*.
  - Winkelmann & Boes: *Analysis of Microdata*.
  - and many others...
- New R functions: Extending/complementing methods previously available in R, e.g., `tobit()` convenience interface to `survreg()`.

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

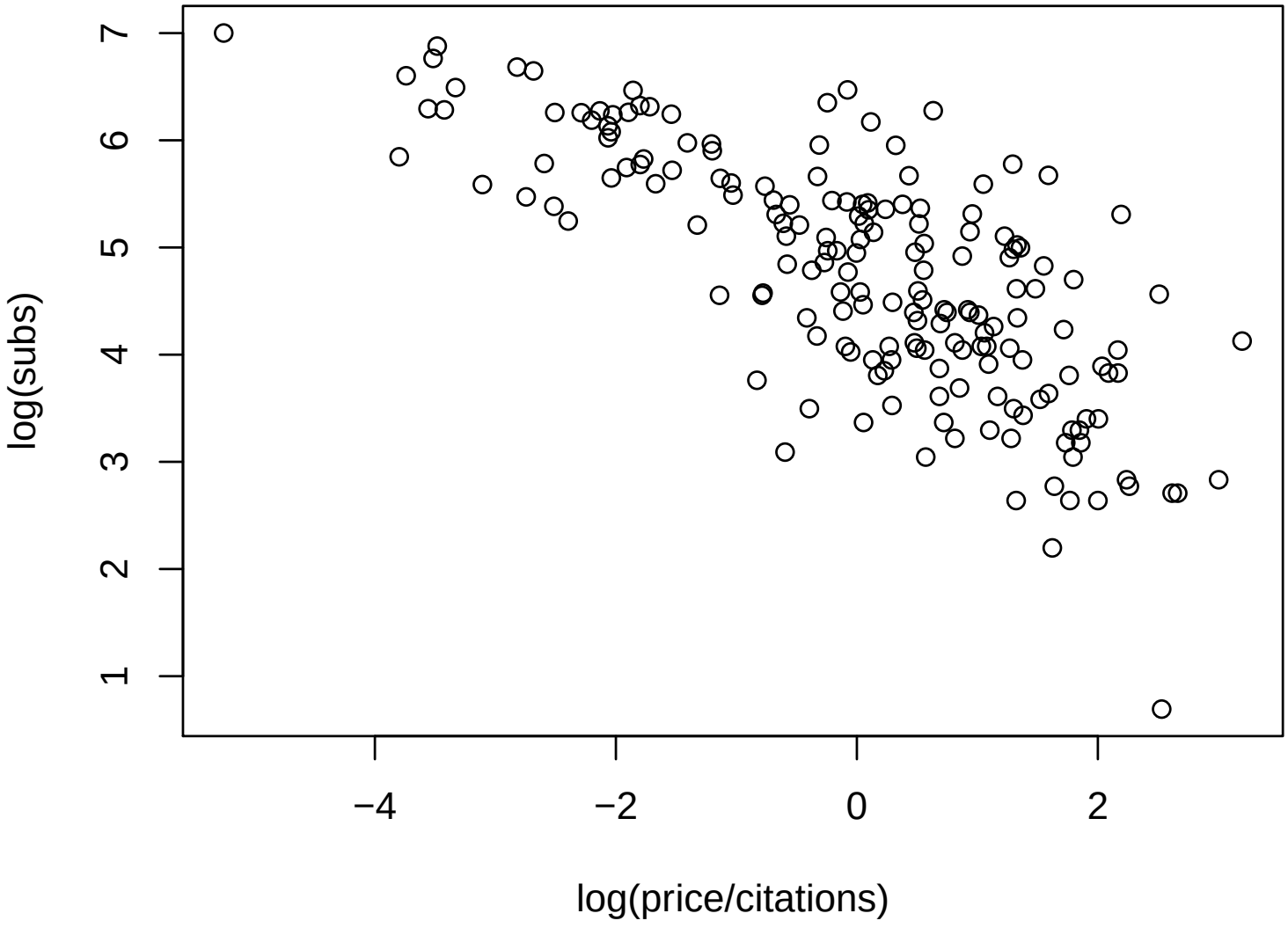


Illustration: Demand for economics journals

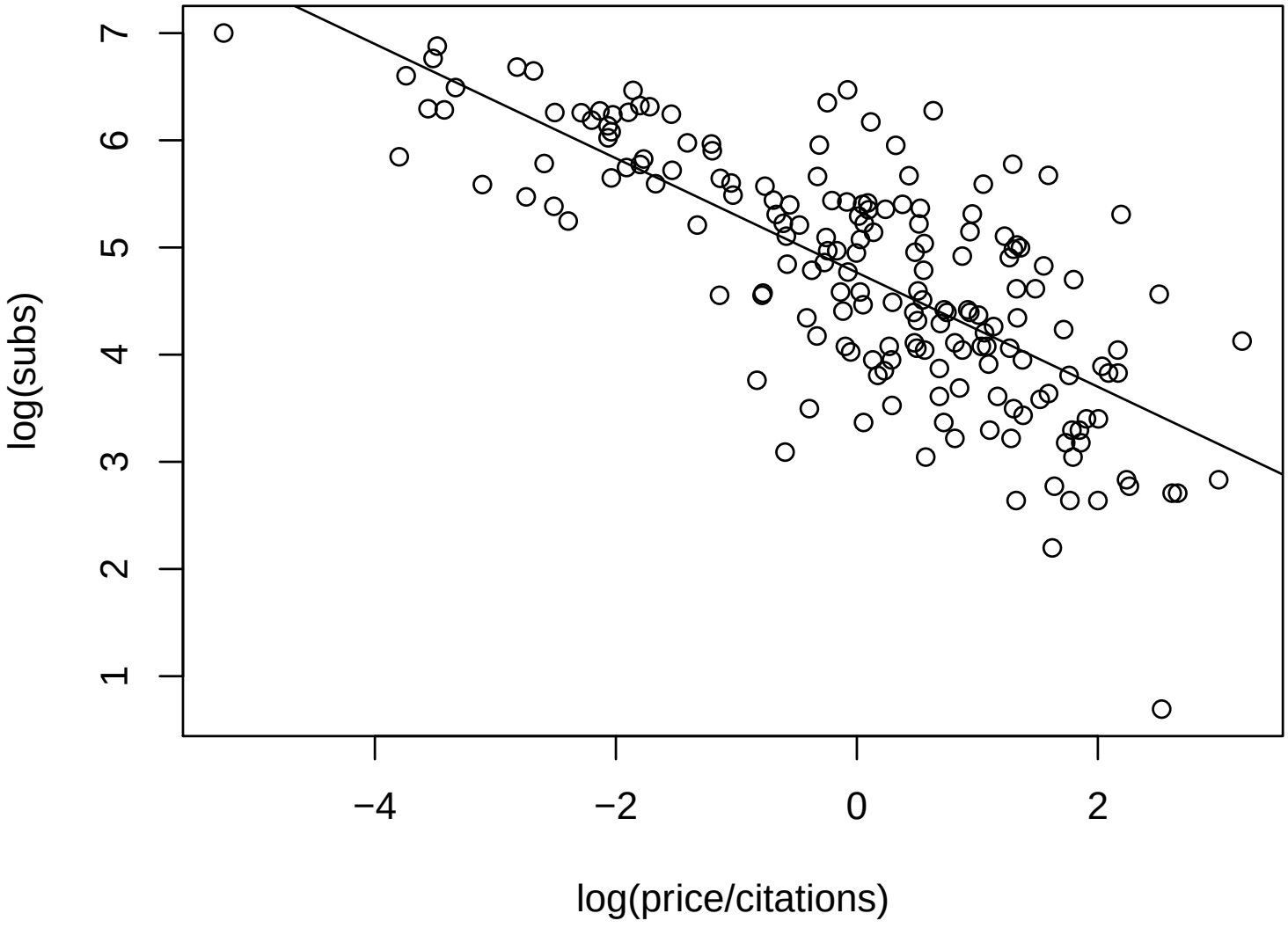


Illustration: Demand for economics journals

```
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 objecs, 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

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

Excursion: Object orientation

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

`confint()`	confidence intervals
`AIC()`, `BIC()`	information criteria (AIC, BIC, ...)
`coeftest()`	partial Wald tests of coefficients (**lmtest**)
`waldtest()`	Wald tests of nested models (**lmtest**)
`linearHypothesis()`	Wald tests of linear hypotheses (**car**)
`lrtest()`	likelihood ratio tests of nested models (**lmtest**)
`sandwich()`, ...	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(price/citations) = -0.5

Model 1: restricted model
Model 2: log(subs) ~ log(price/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: 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 ~ meducation, data = GSOEP9402,
+    breaks = c(7, 9, 10.5, 11.5, 12.5, 15, 18))
```

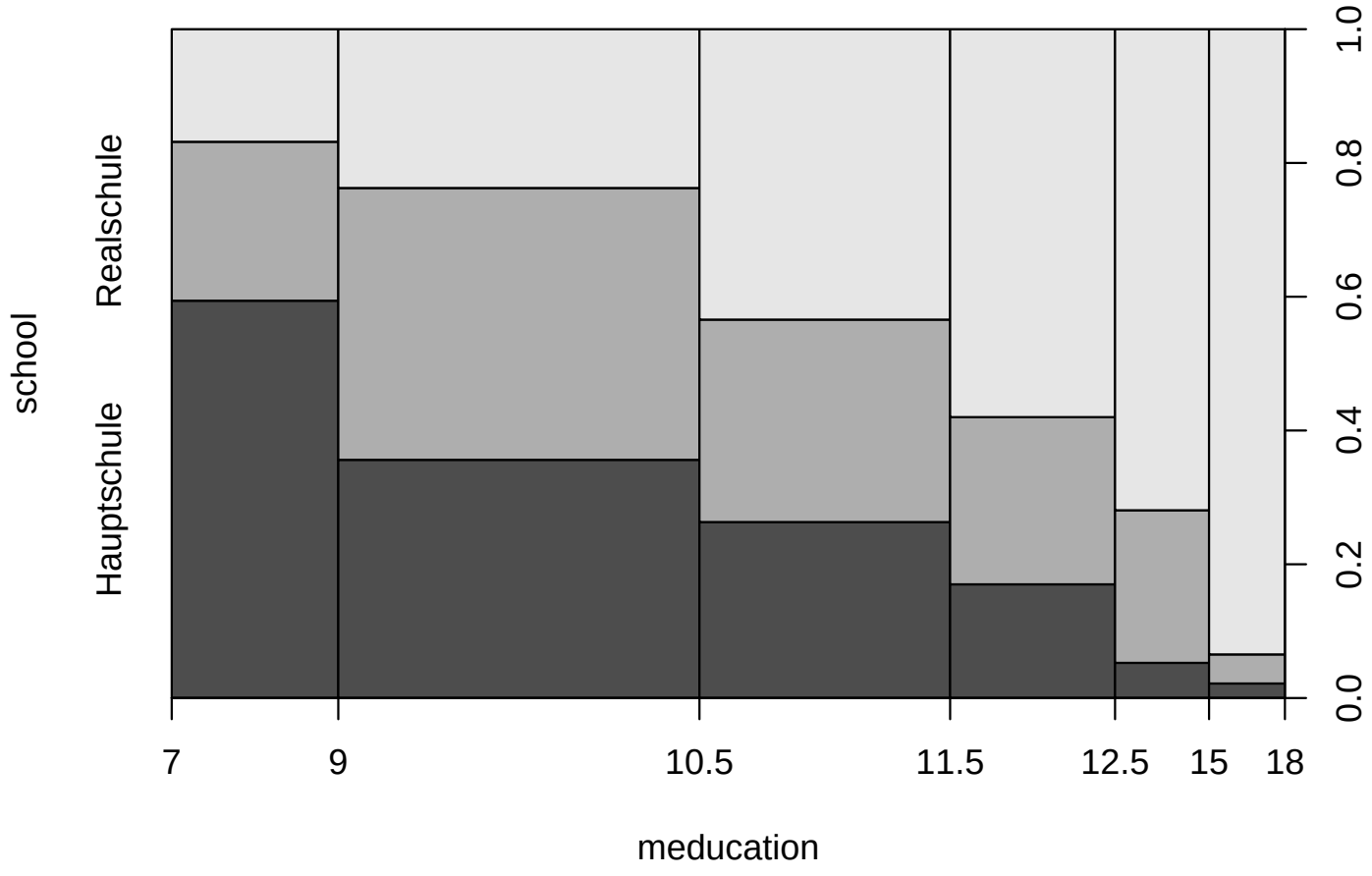


Illustration: Mobility in educational attainment

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

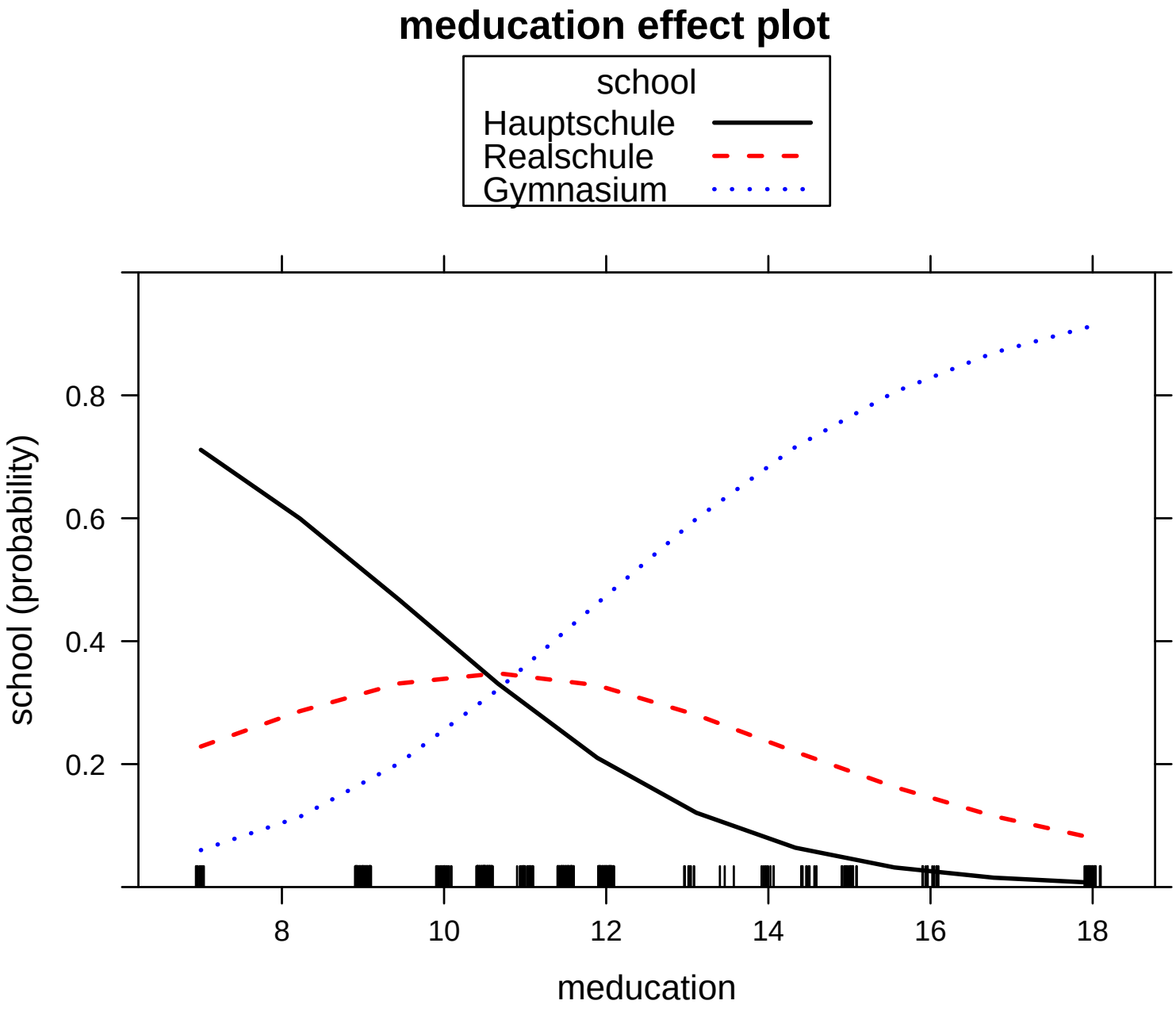
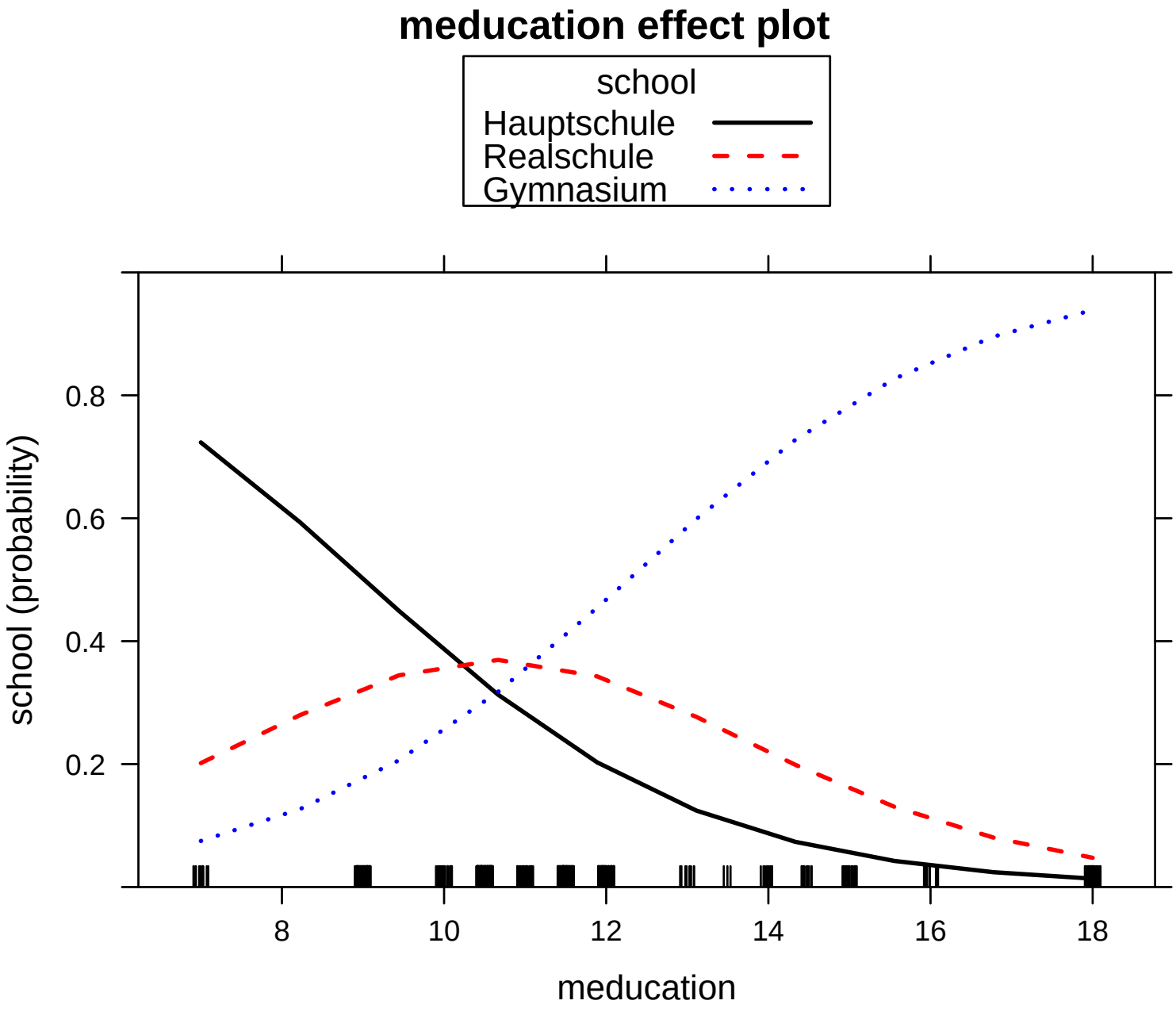


Illustration: Mobility in educational attainment



Further information

CRAN task view:
`http://CRAN.R-project.org/view=Econometrics`
and related views `Finance`, `SocialSciences`, `TimeSeries`.

Mailing lists: R-help, R-SIG-Finance, ...
`https://stat.ethz.ch/mailman/listinfo/R-SIG-Finance`

Journals:

- The R Journal.
  `http://journal.R-project.org/`
- Journal of Statistical Software.
  `http://www.jstatsoft.org/`
  See Volume 27 (Zeileis & Koenker 2008) for overview.

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`

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/`

Koenker R, Zeileis A (2009). "On Reproducible Econometric Research." *Journal of Applied Econometrics*, **24**(5), 833–847.
`doi:10.1002/jae.1083`