

Replizierbarkeit in den Wirtschaftswissenschaften – Eine Schadensbesichtigung

Christian Kleiber

Einleitung

McCullough, McGeary und Harrison (*Journal of Money, Credit and Banking*, 2006):

“In March 1989 at the University of Utah, Stanley Pons and Martin Fleischman claimed to have produced “cold fusion” which, if true, would have led to an almost limitless supply of cheap energy. ...

If Pons and Fleischman had published their cold fusion results in an economics journal, the world would still be awaiting lower utility bills. Distinct from most sciences, economics has not fully embraced the scientific method; in particular, there is no tradition of replication in economics.”

Gliederung

- Einleitung
- Problemfelder
 - Daten
 - Benchmarks
 - Software
- Beispiele
 - eine Nichtstandard-Verteilung in der Zeitreihenökometrie
 - vollständige Separation bei binären abhängigen Variablen
 - Regressionsmodelle für Zähldaten
- Ausblick

Einleitung

Andererseits ...

11.6% der statistischen Resultate in *Nature* 2001 sind “inkongruent”.

Garcia-Berthou und Alcaraz (*BMC Medical Research Methodology* 2004)

Einleitung

- 1982 *Journal of Money, Credit and Banking* (JMCB) Data Storage and Evaluation Project:
Dewald, Thursby and Anderson (*AER*, 1986) finden: 2 aus 54 Arbeiten replizierbar
- 1986 Replikationspolitik *American Economic Review*: Daten und Code
- JMCB Wiederholungsstudie (McCullough, McGeary and Harrison, JMCB 2006):
14 aus 62 replizierbar
- McCullough und Vinod (*AER*, 2003) versuchen Replikation einer gesamten Ausgabe der *AER*

Einleitung

Hier “nur” **Replizierbarkeit im engen Sinn**: sind die konkreten Berechnungen, Graphiken, etc. von Dritten mit vertretbarem Aufwand wiederholbar?

Keine Diskussion von Methodik, Fragestellung, etc.!

Warum sind Arbeiten nicht replizierbar?

- Daten nicht verfügbar
- Daten verfügbar, aber Programme nicht
- Daten und Programme verfügbar, aber
 - Änderungen in Software
 - Fehler in Software
 - numerische Probleme

Einleitung

- 2003 *Journal of Applied Econometrics* Replication Section
- seit 2004: Archiv“pflicht” für Daten und Code bei
American Economic Review, *Econometrica*, *Review of Economic Studies*, *Journal of Political Economy*
- seitdem weitere Zeitschriften mit Replication Section:
Empirical Economics, *Journal of Economic and Social Measurement*

Einleitung

Altman, M., Gill, J. und McDonald, M.P. (2004). *Numerical Issues of Statistical Computing for the Social Scientist*. Hoboken, NJ: John Wiley.

Lovell, M.C., und Selover, D.D. (1994). Econometric software accidents. *Economic Journal*, 104, 713–725.

McCullough, B.D. und H.D. Vinod (1999). The numerical reliability of econometric software. *Journal of Economic Literature*, 37, 633-665.

McCullough, B.D. und H.D. Vinod (2003). Verifying the solution from a nonlinear solver: A case study. *American Economic Review*, 93, 873-892.

Newbold, P., Agiakloglou, C., und Miller, J. (1994). Adventures with ARIMA software. *International Journal of Forecasting*, 10, 573–581.

Einleitung

Aus dem Vorwort von Altman, Gill und McDonald (2004):

“The overall purpose of this work is to address what we see as a serious deficiency in statistical work in the social and behavioral sciences, broadly defined. Quantitative researchers in these fields rely on statistical and mathematical computation as much as any of their colleagues in the natural sciences, yet there is less appreciation for the problems and issues in numerical computation.”

Problemfelder: Was sind die Grunfeld-Daten?

Beliebter Datensatz zu Panels und SUR: Grunfeld's Investitionsdaten aus

Grunfeld, Y. (1958). *The Determinants of Corporate Investment*. Unpublished Ph.D. Dissertation, University of Chicago.

10 Firmen über je 20 Jahre (1935-1954):

[1] "General Motors"	"US Steel"	"General Electric"
[4] "Chrysler"	"Atlantic Refining"	"IBM"
[7] "Union Oil"	"Westinghouse"	"Goodyear"
[10] "Diamond Match"	"American Steel"	

Verwendet in diversen Lehrbüchern, u.a. bei

- Maddala (1977): *Econometrics* (10 Firmen)
- Greene (2003): *Econometric Analysis*, 5e (5 Firmen)
- Baltagi (2005): *Econometric Analysis of Panel Data*, 3e (10 Firmen)

Einleitung

- kulturelle Unterschiede
Wirtschafts- und Sozialwissenschaften \longleftrightarrow Naturwissenschaften
- Verbreitung von PCs und (naive) Nutzung von Voreinstellungen
- (falsche) Anreize für Softwareentwickler:
neueste Verfahren und Schnelligkeit \longleftrightarrow numerische Genauigkeit
- Mangel an numerischen Benchmarks: nur erhältlich für klassische Methoden

Problemfelder: Benchmarks

- National Institute of Standards and Technology (NIST) archiviert Statistical Reference Datasets (StRD)

“The purpose of this project is to improve the accuracy of statistical software by providing reference datasets with certified computational results that enable the objective evaluation of statistical software.”

- Mangel an numerischen Benchmarks: nur erhältlich für
 - elementare Statistik
 - lineare Regression und Varianzanalyse
 - Matrixalgebra
 - Zufallszahlengeneratoren und MCMC
 - ausgewählte statistische Verteilungen
 - nichtlineare Regressionsprobleme
- Ökonometrie: diverse Testprobleme bei C. Cummins (TSP)
Neu: GARCH-Benchmark aus Bollerslev und Ghysels, JBES 1996

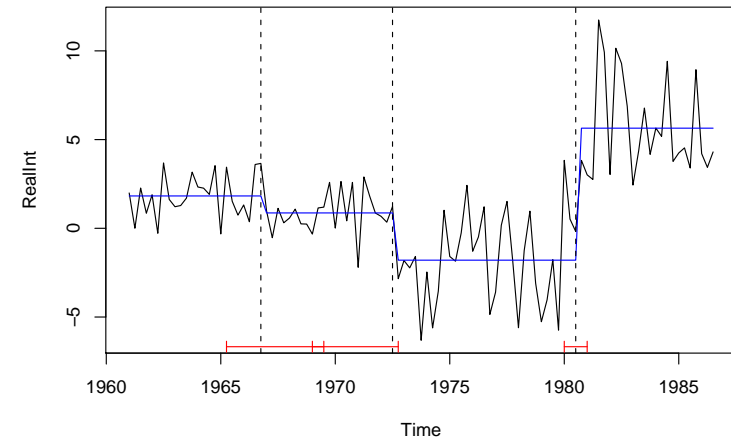
Nichtstandard-Reihenfolge von Operatoren:

Was ist -3^2 ? ... in MS Excel: 9

R.L. Berger (2007): Nonstandard operator precedence in Excel. *Computational Statistics & Data Analysis*, 51, 2788–2791.

Beispiele: Eine Nichtstandard-Verteilung

- Bai und Perron (*JAE*, 2003):
Regression auf Konstante, Standardfehler über HAC mit automatischer Bandbreitenwahl
- Punktschätzer der Brüche (Zeitpunkte) gut replizierbar, aber nur 2 von 3 Konfidenzintervallen
- Ursachen:
 - Konfidenzintervall braucht Quantile einer Nichtstandardverteilung
 - Programmierfehler
 - Softwarefehler
- mehr dazu: Zeileis und Kleiber (*JAE*, 2005)



Beispiele: Eine Nichtstandard-Verteilung

Modell: multiples lineares Regressionsmodell mit m Brüchen (äquivalent: $m + 1$ Regimes)

$$y_t = x_t^\top \beta_j + u_t, \quad t = T_{j-1} + 1, \dots, T_j, \quad j = 1, \dots, m + 1.$$

Zielfunktion:

$$RSS(T_1, \dots, T_m) = \sum_{j=1}^{m+1} r_{ss}(T_{j-1} + 1, T_j)$$

Bestimme Bruchpunkte $\{\hat{T}_1, \dots, \hat{T}_m\} =: \mathcal{I}_{m,n}$ als

$$(\hat{T}_1, \dots, \hat{T}_m) = \operatorname{argmin}_{(T_1, \dots, T_m)} RSS(T_1, \dots, T_m)$$

über Methoden der dynamischen Optimierung (\rightarrow **Bellman-Prinzip**)

$$RSS(\mathcal{I}_{m,n}) = \min_{mn_h \leq t \leq n - n_h} [RSS(\mathcal{I}_{m-1,t}) + r_{ss}(T_j + 1, n)]$$

Beispiele: Eine Nichtstandard-Verteilung

Asymptotik der Bruchpunkte: Grenzverteilung ist Verteilung von

$$\operatorname{argmax}_s V(s)$$

mit

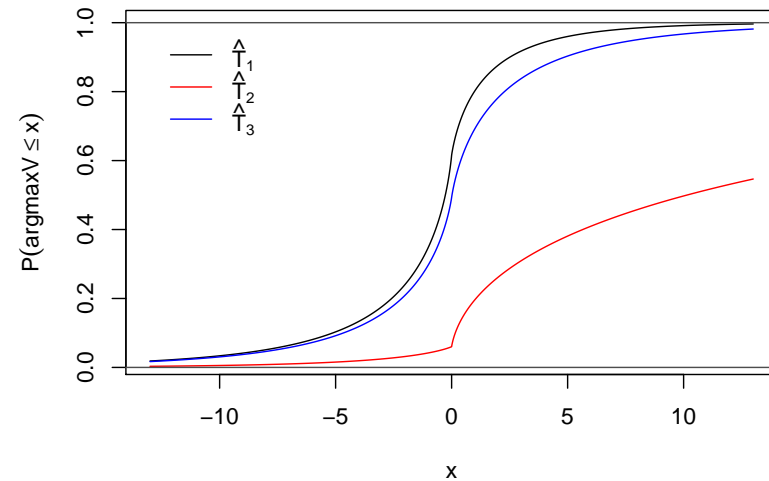
$$V(s) = \begin{cases} W_1(-s) - |s|/2 & \text{für } s \leq 0, \\ \sqrt{\xi}(\phi_2/\phi_1)W_2(s) - \xi s/2 & \text{für } s > 0. \end{cases}$$

→ zweiseitige Brownsche Bewegung mit unterschiedlichen Skalen und linearer Drift

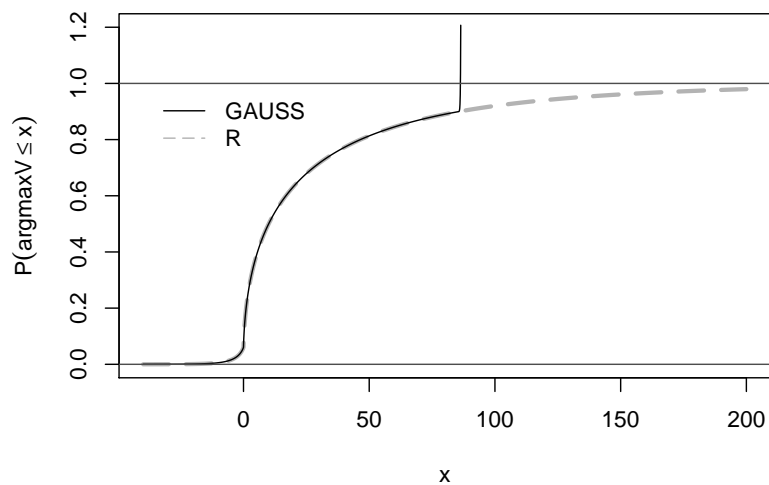
Rechter Arm der Grenzverteilung:

$$G(x) = 1 + \frac{\xi}{\sqrt{\phi} \cdot 2\pi} \sqrt{x} \exp\left\{-\frac{\xi^2}{8\phi} x\right\} + c \exp(ax)\Phi(-b\sqrt{x}) + \left(-d + 2 - \frac{\xi}{2\phi} x\right) \Phi\left(\frac{\xi}{2\sqrt{\phi}}\sqrt{x}\right) \quad (x > 0).$$

Beispiele: Eine Nichtstandard-Verteilung



Beispiele: Eine Nichtstandard-Verteilung



Beispiele: Vollständige Separation

Empirisches Beispiel: Daten aus

Maddala, G.S. (2001). *Introduction to Econometrics*, 3rd ed. New York: John Wiley.

Daten für 44 US-Bundesstaaten für 1950. Variablen sind

rate	Mordrate per 100'000 Einwohner (für 1950).
convictions	Anteil Verurteilungen 1950.
executions	Anteil Exekutionen an Verurteilungen.
time	Haftdauer (Median, in Monaten) von 1951 entlassenen Mördern.
income	Familieneinkommen (Median) für 1949 (in 1,000 USD).
lfp	Arbeitsmarktpartizipationsrate für 1950 (in Prozent).
noncauc	Bevölkerungsanteil nichtweiss
southern	Region (Faktor).

Stokes, H. (2004). On the advantage of using two or more econometric software systems to solve the same problem. *J. Economic and Social Measurement*, 29, 307–320.

Beispiele: Vollständige Separation

Schätzung mit Voreinstellungen liefert:

z test of coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	10.9933	20.7734	0.53	0.597
time	0.0194	0.0104	1.87	0.062
income	10.6101	5.6541	1.88	0.061
lfp	-0.6676	0.4767	-1.40	0.161
noncauc	70.9879	36.4118	1.95	0.051
southernyes	17.3313	2872.1707	0.01	0.995

Beispiele: Vollständige Separation

Ändern des Abbruchkriteriums liefert aber:

z test of coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.10e+01	2.08e+01	0.53	0.597
time	1.94e-02	1.04e-02	1.87	0.062
income	1.06e+01	5.65e+00	1.88	0.061
lfp	-6.68e-01	4.77e-01	-1.40	0.161
noncauc	7.10e+01	3.64e+01	1.95	0.051
southernyes	3.13e+01	1.73e+07	1.8e-06	1.000

Ursache ist "quasi-vollständige Separation"

	no	yes
FALSE	9	0
TRUE	20	15

Beispiele: Regressionsmodelle für Zähldaten

Beispiel: Freizeitverhalten/Tourismus ($n = 659$, USA 1980)
 y_i Anzahl Bootsausflüge auf einem See in Texas

Kovariablen:

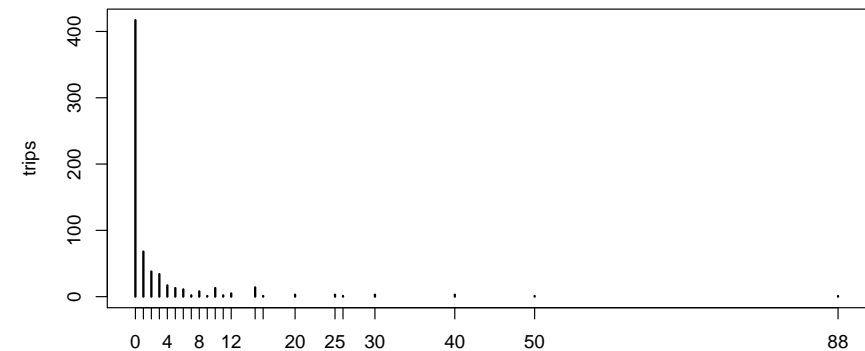
Einkommen, Qualitätsindikator, Benutzungsgebühr (Faktor), Wasserski (Faktor), Opportunitätskosten.

Datensatz wird in diversen Publikationen verwendet, u.a. in

- Sellar, Stoll und Chavas, *Land Economics* 1985
- Ozuna und Gomez, *Empirical Economics* 1995
- Gurmu und Trivedi, *Journal of Business and Economic Statistics* 1996
- Cameron und Trivedi, *Regression Models for Count Data*, CUP 1998

Beispiele: Regressionsmodelle für Zähldaten

Zählvariable trips



Beispiele: Regressionsmodelle für Zähldaten

Ozuna und Gomez: Specification and Testing of Count Data Recreation Demand Functions. *Empirical Economics* 1995.

schätzen Poisson- und Negativ-Binomial-Modell.

Fussnote:

“It should be noted that one of the anonymous referees re-estimated the models used in this study using the same data set and he obtained different parameter estimates. The referee and the authors of this article agreed that the problem was in the software used to estimate the models. The referee used LIMDEP 6.0 for the Poisson and MICROFIT 3.0 for the NLS models whereas the authors used GAUSS 3.0. This is an important observation since the parameter estimates affect consumer surplus. Researchers should thus be cautious of the software they use to estimate the models.”

Beispiele: Regressionsmodelle für Zähldaten

Beispiel: Wahl von Logit + Poisson ergibt

$$P(y_i = 0) = (1 - \xi_i)F(z_i^\top \gamma) + \xi_i\{1 - F(z_i^\top \gamma)\}, \quad \xi_i \in \{0, 1\}$$
$$P(y_i = k_i | y_i > 0) = \frac{e^{-\lambda_i k_i} \lambda_i^{k_i}}{k_i!(1 - e^{-\lambda_i})}, \quad k_i = 1, 2, 3, \dots$$

Führt auf (Log-)Likelihood mit zwei Komponenten:

$$\ell_1(\gamma) = \sum_{y_i=0} \log F(z_i^\top \gamma) + \sum_{y_i>0} \log\{1 - F(z_i^\top \gamma)\}$$
$$\ell_2(\beta) = \sum_{y_i>0} \left\{ y_i(x_i^\top \beta) - e^{x_i^\top \beta} - \log\{1 - \exp(-e^{x_i^\top \beta})\} \right\} - \sum_{y_i>0} \log y_i!$$

Log-Likelihood ist dann $\ell(\beta, \gamma) = \ell_1(\gamma) + \ell_2(\beta)$

Dabei können ℓ_1 und ℓ_2 prinzipiell separat maximiert werden.

Beispiele: Regressionsmodelle für Zähldaten

Hürdenmodell (Mullahy, *JoE* 1986) ist Multiple-Index-Modell mit 2 Komponenten:

- ist y_i gleich Null oder positiv (“Hürde wird überquert”)? → binäre Variable
- falls y_i positiv, wie gross ist y_i ? → Zählvariable, gestutzt in Null

$$f_{\text{hurdle}}(y_i; x_i, z_i, \beta, \gamma) = \begin{cases} f_{\text{zero}}(0; z_i, \gamma), & \text{falls } y_i = 0, \\ \{1 - f_{\text{zero}}(0; z_i, \gamma)\} \cdot f_{\text{count}}(y_i; x_i, \beta) / \{1 - f_{\text{count}}(0; x_i, \beta)\}, & \text{falls } y_i > 0. \end{cases}$$

Beispiele: Regressionsmodelle für Zähldaten

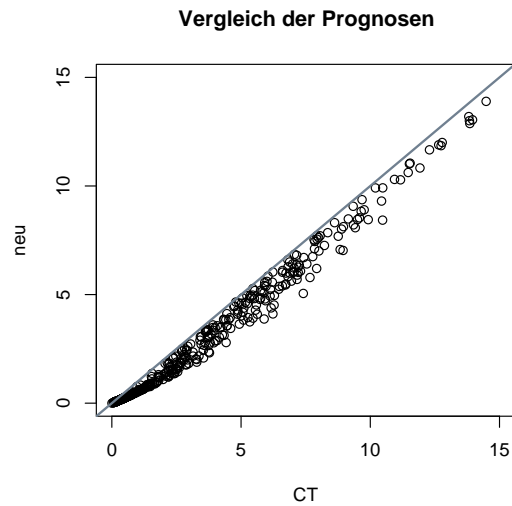
Zwei Schätzungen im Vergleich:

- Cameron und Trivedi (1998). *Regression Models for Count Data*, CUP
Poisson-Poisson-Hürde ergibt Likelihood -1209.582
- Neuberechnung mit R-Paket **pscl**
Poisson-Poisson-Hürde ergibt Likelihood -1181.612

Korrelation der Prognosen mit 0.999 aber hoch.

Unterschiede kommen aus Binärteil.

Beispiele: Regressionsmodelle für Zähldaten



C Kleiber

29

U Basel

Beispiele: Regressionsmodelle für Zähldaten

Ursache ist (wieder ...) **quasi-vollständige Separation**:

```
userfee
trips > 0 no yes
FALSE 417 0
TRUE 229 13
```

Implizites Konfidenzintervall bei Logit-Schätzung:

```
2.5 % 97.5 %
-98.85 NA
```

Unschön:

- mit Voreinstellungen von `glm.control`, insb. `epsilon = 1e-8` keine Warnung
- Warnung erst ab `epsilon = 1e-12`

C Kleiber

30

U Basel

Beispiele: Regressionsmodelle für Zähldaten

Fazit Datensatz RecreationDemand:

- eine Variable wird typischerweise falsch behandelt
- eine andere Variable ist bedingt nützlich
- noch 1995 schreiben Autoren selbst Code für eine Poisson-Regression
- Multiple Index-Modelle für Zähldaten bergen Überraschungen

C Kleiber

31

U Basel

Schlussfolgerungen

Massnahmen:

- Autoren:
 - Vorsicht mit Voreinstellungen
 - Info zu Software nebst Version und OS
 - Daten und Programme zur Verfügung stellen
- Zeitschriften:
 - Daten und Programme verpflichtend
 - mehr Archive
- Dozenten:
 - verwende Archive in der Lehre

C Kleiber

32

U Basel

Schlussfolgerungen

Wünschenswert: mehr als Daten und Code – komplett replizierbare Analysen

Eine Lösung: R-Funktion Sweave kombiniert

R und \LaTeX

Schlussfolgerungen

Schlussfolgerungen

... Confidence intervals for the breakpoints can be computed from the fitted `\texttt{bp.ri}` object for any number of breaks (smaller than the maximal number of breaks admissible) using the `\texttt{confint}` method from `\texttt{strucchange}`. A function for estimating the covariance matrix, here `\texttt{kernHAC}`, may again be supplied.

```
\begin{Schunk}
\begin{Sinput}
R> confint(bp.ri, breaks = 3, vcov = kernHAC)
\end{Sinput}
\end{Schunk}
```

This returns the breakpoints and corresponding confidence intervals (at the default 95\% level) coded by ...

... und die BWL?