J Labour Market Res. (2014) 47:43–62
DOI 10.1007/s12651-013-0135-0

ARTICLE

Estimation of standard errors and treatment effects in empirical economics—methods and applications

Olaf Hübler

Published online: 10 July 2013

Abstract This paper discusses methodological problems of standard errors and treatment effects. First, heteroskedasticity- and cluster-robust estimates are considered as well as problems with Bernoulli distributed regressors, outliers and partially identified parameters. Second, procedures to determine treatment effects are analyzed. Four principles are in the focus: difference-in-differences estimators, matching procedures, treatment effects in quantile regression analysis and regression discontinuity approaches. These methods are applied to Cobb-Douglas functions using IAB establishment panel data.

Different heteroskedasticity-consistent procedures lead to similar results of standard errors. Cluster-robust estimates show evident deviates. Dummies with a mean near 0.5 have a smaller variance of the coefficient estimates than others. Not all outliers have a strong influence on significance. New methods to handle the problem of partially identified parameters lead to more efficient estimates.

The four discussed treatment procedures are applied to the question whether company-level pacts affect the output. In contrast to unconditional difference-in-differences and to estimates without matching the company-level effect is positive but insignificant if conditional difference-in-differences, nearest-neighbor or Mahalanobis metric matching is applied. The latter result has to be specified under quantile treatment effects analysis. The higher the quantile the higher is the positive company-level pact effect and there is a tendency from insignificant to significant effects. A sharp regression discontinuity analysis shows a structural break at a probability of 0.5 that a company-level pact exists. No specific effect of the Great Recession can be detected. Fuzzy regression discontinuity estimates reveal that the company-level pact effect is significantly lower in East than in West Germany.

Keywords Standard errors · Outliers · Partially identified parameters · DiD estimators · Matching · Quantile regressions · Regression discontinuity

JEL Classification C21 · C26 · D22 · J53

Schätzung von Standardfehlern und Kausaleffekten in der empirischen Wirtschaftsforschung – Methoden und Anwendungen

Zusammenfassung Dieser Beitrag diskutiert Möglichkeiten zur Schätzung von Standardfehlern und Kausaleffekten. Zunächst werden heteroskedastie- und gruppenrobuste Schätzungen für Standardfehler betrachtet sowie Auffälligkeitenund Probleme bei Dummy-Variablen als Regressoren, Ausreißern und nur partiell identifizierten Parametern erörtert. Danach geht es um Verfahren zur Bestimmung von Treatmenteffekten. Vier Prinzipien werden hierzuvorgestellt: Differenz-von-Differenzen-Schätzer, Matchingverfahren, Kausaleffekte in der Quantilsregressionsanalyse und Ansätze zur Bestimmung von Diskontinuitäten bei Regressionsschätzungen. Anwendungen erfolgen im zweiten Teil der Arbeit auf Cobb-Douglas-Produktionsfunktionen unter Verwendung von IAB-Betriebspaneldaten.

Verschiedene heteroskedastiekonsistente Verfahren führen zu recht ähnlichen Ergebnissen bei den Standardfehlern. Clusterrobuste Schätzungen zeigen dagegen deutliche Abweichungen. Dummies als Regressoren mit einem Mittelwert in der Nähe von 0.5 weisen kleinere Varianzen der

O. Hübler (✉)
Institut für Empirische Wirtschaftsforschung, Leibniz Universität Hannover, Königsworther Platz 1, 30167 Hannover, Germany
e-mail: huebler@ewifo.uni-hannover.de

Koeffizienterschätzer auf als andere. Nicht alle Ausreißer haben einen nennenswerten Einfluss auf die Signifikanz. Neuere Methoden zur Behandlung des Problems von nur partiell identifizierten Parametern führen zu effizienteren Schätzungen.

Die vier diskutierten Verfahren zur Bestimmung der Wirkungen von Maßnahmen werden auf das Problem, ob betriebliche Bündnisse einen signifikanten Einfluss auf den Produktionsoutput haben, angewandt. Im Gegensatz zu nicht konditionalen Differenz-von-Differenzen-Schätzern und Schätzern ohne Matching sind die Effekte betrieblicher Bündnisse bei bedingten Differenz-von-Differenzen-Schätzern und Matching-Verfahren zwar positiv, aber insignifikant. Diese Aussage ist auf Basis der Treatment-Quantilsanalysezu präzisieren. Je höher die Quantile sind, umso größer ist die Wirkung betrieblicher Bündnisse mit einer Tendenz von insignifikanten zu signifikanten Effekten. Die deterministische Regressionsanalyse mit Diskontinuitäten zeigt einen Strukturbruch bei Wahrscheinlichkeit 0.5, dass ein betriebliches Bündnis existiert. Es lassen sich keine spezifischen Effekte während der Rezession 2009 ausmachen. Schätzungen im Rahmen stochastischer Diskontinuitätsansätze offenbaren, dass die Wirkungen betrieblicher Bündnisse in Ostdeutschland signifikant niedriger ausfallen als in Westdeutschland.

1 Introduction

Contents, questions and methods have changed in empirical economics in the last 20 years. Many methods were developed in the past but the application in empirical economics follows with a lag. Some methods are well-known but have experienced only little attention. New approaches focus on characteristics of the data, on modified estimators, on correct specifications, on unobserved heterogeneity, on endogeneity and on causal effects. Real data sets are not compatible with the assumptions of classical models. Therefore, modified methods were suggested for the estimation and inference.

The road map of the following considerations are four hypotheses where the first two and the second two belong together:

(1) Significance is an important indicator in empirical economics but the results are sometimes misleading.
(2) Assumptions' violation, clustering of the data, outliers and only partially identified parameters are often the reason of wrong standard errors using classical methods.
(3) The estimation of average effects is useful but subgroup analysis and quantile regressions are important supplements.
(4) Causal effects are of great interest but the determination is based on disparate approaches with varying results.

In the following some econometric methods are developed, presented and applied to Cobb-Douglas production functions.

2 Econometric methods

2.1 Significance and standard errors in regression models

The working horse in empirical economics is the classical linear model

$$y_i = x_i'\beta + u_i, \quad i = 1, \ldots, n.$$

The coefficient vector β is estimated by ordinary least squares (OLS)

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

and the covariance matrix by

$$\hat{V}(\hat{\beta}) = \hat{\sigma}^2 \left(X'X\right)^{-1},$$

where X is the design matrix and $\hat{\sigma}^2$ the estimated variance of the disturbances. The influence of a regressor, e.g. x_k, on the regressand y is called significant at a 5 percent level if $|t| = |\hat{\beta}_k / \sqrt{\hat{V}(\hat{\beta}_k)}| > t_{0.975}$. In empirical papers this result is often documented by an asterisk and implicitly interpreted as a good one, while insignificance is a negative signal. Ziliak and McCloskey (2008) and Krämer (2011) have criticized this procedure although the analysis is extended by robustness tests in many investigations. Three types of mistakes can lead to a misleading interpretation:

(1) There does not exist any effect but due to technical inefficiencies a significant effect is reported.
(2) The effect is small but due to the precision of the estimates a significant effect is determined.
(3) There exists a strong effect but due to the variability of the estimates the statistical effect cannot be detected.

The consequence cannot be to neglect the instrument of significance. But what can we do? The following proposals may help to clarify why some standard errors are high and others low, why some influences are significant and others not, whether alternative procedures can reduce the danger of one of the three mistakes:

- Compute robust standard errors.
- Analyze whether variation within clusters is only small in comparison with variation between the clusters.
- Check whether dummies as regressors with high or low probability are responsible for insignificance.
- Test whether outliers induce large standard errors.
- Consider the problem of partially identified parameters.

- Detect whether collinearity is effective.
- Investigate alternative specifications.
- Use sub-samples and compare the results.
- Execute sensitivity analyses (Leamer 1985).
- Employ the sniff test (Hamermesh 2000) in order to detect whether econometric results are in accord with economic plausibility.

2.1.1 Heteroskedasticity-robust standard errors

OLS estimates are inefficient or biased and inconsistent if assumptions of the classical linear model are violated. We need alternatives which are robust to the violation of specific assumptions. In empirical papers we find often the hint that robust standard errors are displayed. This is imprecise. In most cases this means only heteroskedasticity-robust. This should be mentioned and also that the estimation is based on White's approach. If we know the type of heteroskedasticity, a transformation of the regression model should be preferred, namely

$$\frac{y_i}{\sigma_i} = \frac{\beta_0}{\sigma_i} + \beta_1 \frac{x_{1i}}{\sigma_i} + \cdots + \beta_K \frac{x_{Ki}}{\sigma_i} + \frac{u_i}{\sigma_i},$$

where $i = 1, \ldots, n$. Typically, the individual variances of the error term are unknown. In the case of unknown and unspecific heteroscedasticity White (1980) recommends the following estimation of the covariance matrix

$$\hat{V}_{white}(\hat{\beta}) = \left(X'X\right)^{-1}\left(\sum \hat{u}_i^2 x_i x_i'\right)\left(X'X\right)^{-1}.$$

Such estimates are asymptotically heteroscedasticity-robust. In many empirical investigations this robust estimator is routinely applied without testing whether heteroskedasticity exists. We should stress that those estimated standard errors are more biased than conventional estimators if residuals are homoskedastic. As long as there is not too much heteroskedasticity, robust standard errors are also biased downward. In the literature we find some suggestions to modify this estimator, namely to weight the squared residuals $\hat{u}_i^2$:

$$hc_1 = \frac{n}{n-K}\hat{u}_i^2$$

$$hc_j = \frac{1}{(1-c_{ii})^{\delta_j}}\hat{u}_i^2,$$

where $j = 2, 3, 4$, c_{ii} is the main diagonal element of $X'(X'X)^{-1}X$ and $\delta_j = 1; 2; \min[\gamma_1, (nc_{ii})/K] + \min[\gamma_2, (nc_{ii})/K]$, γ_1 and γ_2 are real positive constants.

The intention is to obtain more efficient estimates. It can be shown for hc_2 that under homoskedasticity the mean of $\hat{u}_i^2$ is the same as $\sigma^2(1-c_{ii})$. Therefore, we should expect that the hc_2 option leads under homoskedasticity to better estimates in small samples than the simple hc_1 option. Then $E(\hat{u}_i^2/(1-c_{ii}))$ is σ^2. The second correction is presented by MacKinnon and White (1985). This is an approximation of a more complicated estimator which is based on a jackknife estimator—see Sect. 2.1.2. Applications demonstrate that the standard error increases started with OLS via hc_1, hc_2 to the hc_3 option. Simulations, however, do not show a clear preference. As one cannot be sure which case is the correct one, a conservative choice is preferable (Angrist and Pischke 2009, p. 302). The estimator should be chosen that has the largest standard error. This means the null hypothesis (H_0: no influence on the regressand) keeps up longer than with other options.

Cribari-Neto and da Silva (2011) suggest $\gamma_1 = 1$ and $\gamma_2 = 1.5$ in hc_4. The intention is to weaken the effect of influential observations compared with hc_2 and hc_3 or in other words to enlarge the standard errors. In an earlier version (Cribari-Neto et al. 2007) a slight modification is presented: $hc_4^* = 1/(1-c_{ii})^{\delta_{4*}}$, where $\delta_{4*} = \min(4, nc_{ii}/K)$. It is argued that the presence of high leverage observations is more decisive for the finite-sample behavior of the consistent estimators of $V(\hat{\beta})$ than the intensity of heteroskedasticity, hc_4 and hc_{4*} aim at discounting for leverage points—see Sect. 2.1.5—more heavily than hc_2 and hc_3. The same authors formulate a further estimator

$$hc_5 = \frac{1}{(1-c_{ii})^{\delta_5}}\hat{u}_i^2,$$

where $\delta_5 = \min(\frac{nc_{ii}}{K}, \max(4, \frac{nkc_{ii,\max}}{K}))$, k is a predefined constant, where $k = 0.7$ is suggested. In this case squared residuals are affected by the maximal leverage.

2.1.2 Re-sampling procedures

Other possibilities to determine the standard error are the jackknife and the bootstrap estimator. These are re-sampling procedures, which construct sub-samples with $n-1$ observations in the jackknife case. Sequentially, one observation is eliminated. The former methods compare the estimated coefficients of the total sample size $\hat{\beta}$ with those after eliminating one observation $\hat{\beta}_{-i}$. The *jackknife* estimator of the covariance matrix is

$$\hat{V}_{\text{jack}} = \frac{n-K}{n}\sum_{i=1}^{n}(\hat{\beta}_{-i} - \hat{\beta})(\hat{\beta}_{-i} - \hat{\beta})'.$$

There exist many ways to *bootstrap* regression estimates. The basic idea is assume that the sample with n elements is the population and B times m elements (sampling with replacement) are drawn, where $m \leq n$ and $m > n$ is feasible. If $\hat{\beta}'_{\text{boot}} = (\hat{\beta}(1)'_m, \ldots, \hat{\beta}(B)'_m)$ are the bootstrap estimators

of the coefficients the asymptotic covariance matrix is

$$\hat{V}_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B}\big(\hat{\beta}(b)_m - \hat{\beta}\big)\big(\hat{\beta}(b)_m - \hat{\beta}\big)',$$

where $\hat{\beta}$ is the estimator with the original sample size n. Alternatively, $\hat{\beta}$ can be substituted by $\bar{\beta} = 1/B\sum\hat{\beta}(b)_m$. Bootstrap estimates of the standard error are especially helpful when it is difficult to compute standard errors by conventional methods, e.g. 2SLS estimators under heteroskedasticity or cluster-robust standard errors when many small clusters or only short panels exist. The jackknife can be viewed as a linear approximation of the bootstrap estimator. A further popular way to estimate the standard errors is the *delta method*. This approach is especially used for nonlinear functions of parameter estimates $\hat{\gamma} = g(\hat{\beta})$. An asymptotic approximation of the covariance matrix of a vector of such functions is determined. It can be shown that

$$n^{1/2}(\hat{\gamma} - \gamma_0) \sim N\big(0, G_0 V^{\infty}(\hat{\beta})G_0'\big),$$

where γ_0 is the vector of the true values of γ, G_0 is an $l \times K$ matrix with typical element $\partial g_i(\beta)/\partial\beta_j$, evaluated at β_0, and V^{∞} is the asymptotic covariance matrix of $n^{1/2}(\hat{\beta} - \beta_0)$.

2.1.3 The Moulton problem

The variance of a regressor is low if this variable strongly varies between groups but only little within groups (Moulton 1986, 1987, 1990). This is especially the case if industry, regional and macroeconomic variables are introduced in a microeconomic model or panel data are considered. In a more general context this is called the problem of cluster sampling. Individuals or establishments are sampled in groups or clusters. Consequence may be a weighted estimation that adjust for differences in sampling rates. However, weighting is not always necessary and estimates may understate the true standard errors. Some empirical investigations note that cluster-robust standard errors are displayed but do not mention the cluster variable. If panel data are used then this is usually the identification variable of the individuals or firms. In many specifications more than one cluster variable, e.g. a regional and an industry variable, is incorporated. Then it is misleading if the cluster variable is not mentioned. Furthermore, then a sequential determination of a cluster-robust correction is not qualified if there is a dependency between the cluster variables. If we can assume that there is a hierarchy of the cluster variables then a multilevel approach can be applied (Raudenbush and Bryk 2002; Goldstein 2003). Cameron and Miller (2010) suggest a two-way clustering procedure. The covariance matrix can be determined by

$$\hat{V}_{\text{two-way}}(\hat{\beta}) = \hat{V}_1(\hat{\beta}) + \hat{V}_1(\hat{\beta}) - \hat{V}_{1\cap 2}(\hat{\beta})$$

when the three components are computed by

$$\hat{V}(\hat{\beta}) = \big(X'X\big)^{-1}\hat{B}\big(X'X\big)^{-1}$$

$$\hat{B} = \left(\sum_{g=1}^{G} X_g'\hat{u}_g\hat{u}_g'X_g\right).$$

Different ways of clustering can be used. Cluster-robust inference asymptotics are based on $G \to \infty$. In many applications there are only a few clusters. In this case $\hat{u}_g$ has to be modified. One way is the following transformation

$$\tilde{u}_g = \sqrt{\frac{G}{G-1}}\hat{u}_g.$$

Further methods and suggestions in the literature are presented by Cameron and Miller (2010) and Wooldridge (2003).

A simple and extreme example shall demonstrate the cluster problem.

Example Assume a data set with 5 observations ($n = 5$) and 4 variables ($V1$–$V4$).

i	$V1$	$V2$	$V3$	$V4$
1	24	123	−234	−8
2	875	87	54	3
3	−12	1234	−876	345
4	231	−87	−65	9808
5	43	34	9	−765

The linear model

$$V1 = \beta_1 + \beta_2 V2 + \beta_3 V3 + \beta_4 V4 + u$$

is estimated by OLS using the original data set (1M). Then the data set is doubled (2M), quadrupled (4M) and octuplicated (8M). The following OLS estimates result.

		1M	2M	4M	8M
	$\hat{\beta}$	$\hat{\sigma}_{\hat{\beta}}$	$\hat{\sigma}_{\hat{\beta}}$	$\hat{\sigma}_{\hat{\beta}}$	$\hat{\sigma}_{\hat{\beta}}$
$V2$	1.7239	1.7532	0.7158	0.4383	0.2922
$V3$	2.7941	2.3874	0.9747	0.5969	0.3979
$V4$	0.0270	0.0618	0.0252	0.0154	0.0103
const	323.2734	270.5781	110.463	67.64452	45.0963

The coefficients of 1M to 8M are the same, however, the standard errors decrease if the same data set is multiplied. Namely, the variance is only 1/6, 1/16 and 1/36 of the original variance. The general relationship can be shown as follows. For the original data set (X_1) the covariance matrix is

$$\hat{V}_1(\hat{\beta}) = \hat{\sigma}_1^2\big(X_1'X_1\big)^{-1}.$$

Using $X_1 = \cdots = X_F$ the F times enlarged data set with the design matrix $X' =: (X'_1 \cdots X'_F)$ leads to

$$\hat{\sigma}_F^2 = \frac{1}{F \cdot n - K} \sum_{i=1}^{F \cdot n} \hat{u}_i^2 = \frac{F(n-K)}{F \cdot n - K} \hat{\sigma}_1^2$$

and

$$\hat{V}_F(\hat{\beta}) = \hat{\sigma}_F^2 \left(X'X\right)^{-1} = \hat{\sigma}_F^2 \frac{1}{F} \cdot \left(X'_1 X_1\right)^{-1}$$
$$= \frac{n-K}{F \cdot n - K} \hat{V}_1(\hat{\beta}).$$

K is the number of regressors including the constant term, n is the number of observations in the original data set (number of clusters), F is the number of observations within a cluster. In the numerical example with $F = 8$, $K = 4$, $n = 5$ the Moulton factor MF that indicates the deflation factor of the variance is

$$MF = \frac{n-K}{F \cdot n - K} = \frac{1}{36}.$$

This is exactly the same as it was demonstrated in the numerical example. Analogously the estimated values $1/6$ and $1/16$ can be determined. As the multiplying of the data set does not add any further information to the simple original data set not only the coefficients but also the standard errors should be the same. Therefore, it is necessary to correct the covariance matrix. Statistical packages, e.g. Stata, supply cluster-robust estimates

$$\hat{V}(\hat{\beta})_C = \left(\sum_{c=1}^{C} X'_c X_c\right)^{-1} \sum_{c=1}^{C} X'_c \hat{u}_c \hat{u}_c X_c \left(\sum_{c=1}^{C} X'_c X_c\right)^{-1},$$

where C is the number of clusters. In our specific case this is the number of observations n. This approach implicitly assumes that F is small and $n \to \infty$. If this assumption does not hold a degrees-of-freedom correction

$$df_C = \frac{F \cdot n - 1}{F \cdot n - K} \cdot \frac{n}{n-1}$$

is helpful. $df_C \cdot \hat{V}(\hat{\beta})_C$ is the default option in Stata and corrects for the number of clusters in practice being finite. Nevertheless, this correction eliminates only partially the underestimated standard errors. In other words, the corrected t-statistic of the regressor x_k is larger than that of $\hat{\beta}_k / \sqrt{\hat{V}_{1k}}$.

2.1.4 *Large standard errors of dichotomous regressors with small or large mean*

Another problem with estimated standard errors can be induced by Bernoulli distributed regressors. Assume a simple two-variable classical regression model

$$y = a + b \cdot D + u.$$

D is a dummy variable and the variance of $\hat{b}$ is

$$V(\hat{b}) = \frac{\sigma^2}{n} \cdot \frac{1}{s_D^2},$$

where

$$s_D^2 = \hat{P}(D=1) \cdot \hat{P}(D=0) =: \hat{p}(1-\hat{p})$$
$$= \frac{(n|D=1)}{n} \cdot \left(1 - \frac{(n|D=1)}{n}\right).$$

If s_D^2 is determined by $\bar{D} = (n|D=1)/n$ we find that $\bar{D}$ is at most 0.5. $V(\hat{b})$ is minimal at given n and σ^2 when the sample variance of D reaches the maximum, if $\bar{D} = 0.5$. This result holds only for inhomogeneous models.

Example An income variable ($Y = Y_0/10^7$) with 53,664 observations is regressed on a Bernoulli distributed random variable RV. The coefficient β_1 of the linear model $Y = \beta_0 + \beta_1 RV + u$ is estimated by OLS, where alternative values of the mean of RV ($\overline{RV}$) are assumed $(0.1, 0.2, \ldots, 0.9)$

Y	$\hat{\beta}_1$	std.err.
$\overline{RV} = 0.1$	−0.3727	0.6819
$\overline{RV} = 0.2$	−0.5970	0.5100
$\overline{RV} = 0.3$	−0.4768	0.4455
$\overline{RV} = 0.4$	0.3068	0.4170
$\overline{RV} = \mathbf{0.5}$	0.1338	**0.4094**
$\overline{RV} = 0.6$	0.0947	0.4187
$\overline{RV} = 0.7$	−0.0581	0.4479
$\overline{RV} = 0.8$	−0.1860	0.5140
$\overline{RV} = 0.9$	−0.1010	0.6827

This example confirms the theoretical result. The standard error is smallest if $\overline{RV} = 0.5$ and increases systematically if the mean of RV decreases or increases. An extension to multiple regression models seems possible—see applications in the Appendix, Tables 11, 12, 13, 14. The more $\bar{D}$ deviates from 0.5, the larger or smaller is the mean of D, the higher is the tendency to insignificant effects. A caveat is necessary. The conclusion that the t-value of a dichotomous regressor D_1 is always smaller than that of D_2, when $V(D_1) > V(D_2)$, is not unavoidable. The basic effect of D_1 on y may be larger than that of D_2 on y. The theoretical result aims on specific variables and not on the comparison between regressors. In practice, significance is determined by $t = \hat{b} / \sqrt{\hat{V}(\hat{b})}$. However, we do not find a systematic influence of $\hat{b}$ on t if $\bar{D}$ varies. Nevertheless, the random differences in the influence of D on y can dominate the $\bar{D}$ effect via s_D^2. The comparison of Table 13 with Table 14 shows that the influence of a works council (WOCO) is stronger

than that of a company-level pact (CLP). The coefficients of the former regressor are larger and the standard errors are lower than that of the latter regressor so that the t-values are larger. In both cases the standard errors increase if the mean of the regressor is reduced. The comparison of line 1 in Table 13 with line 9 in Table 14, where the mean of CLP and WOCO is nearly the same, makes clear that the stronger basic effect of WOCO on $\ln Y$ dominates the mean reduction effect of WOCO. The t-value in line 9 of Table 14 is smaller than that in line 1 of Table 14 but still larger than that in line 1 of Table 13. Not all deviations of the mean of a dummy D as regressor from 0.5 induce the described standard error effects. A random variation of $\bar{D}$ is necessary. An example, where this is not the case, is matching—see Sect. 2.2 and the application in Sect. 3. $\bar{D}$ increases due to the systematic elimination of those observations with $D = 0$ that are dissimilar to those of $D = 1$ in other characteristics.

2.1.5 *Outliers and influential observations*

Outliers may have strong effects on the estimates of the coefficients, of the dependent variable and on standard errors and therefore on significance. In the literature we find some suggestions to measure outliers that are due to large or small values of the dependent variable or on the independent variables. Belsley et al. (1980) use the main diagonal elements c_{ii} of the hat matrix $C = X(X'X)^{-1}X'$ to determine the effects of a single observation on the coefficient estimator $\hat{\beta}$, on the estimated endogenous variable $\hat{y}_i$ and on the variance $\hat{V}(\hat{y})$. The higher c_{ii}, the higher is the difference between the estimated dependent variable with and without the ith observation. A rule of thumb orients on the relation

$$c_{ii} > \frac{2K}{n}.$$

An observation i is called an influential observation with a strong leverage if this inequality is fulfilled. The effects of the ith observation on $\hat{\beta}$, $\hat{y}$ and $\hat{V}(\hat{\beta})$ and the rules of thumb can be expressed by

$$\left|\hat{\beta}_k - \hat{\beta}_k(i)\right| > \frac{2}{\sqrt{n}}$$

$$\left|\frac{\hat{y}_i - \hat{y}_{i(i)}}{s(i)\sqrt{c_{ii}}}\right| > 2\sqrt{\frac{K}{n}}$$

$$\left|\frac{\det(s^2(i)(X'(i)X(i))^{-1}}{\det(s^2(X'X)^{-1})}\right| > \frac{3K}{n}.$$

If the inequalities are fulfilled, this indicates a strong influence of observation i where (i) means that observation i is not considered in the estimates. The determination of an outlier is based on externally studentized residuals

$$\hat{u}_i^* = \frac{\hat{u}_i}{s(i)\sqrt{1 - c_{ii}}} \sim t_{n-K-1}.$$

Observations which fulfill the inequality $|\hat{u}_i^*| > t_{1-\alpha/2;n-K-1}$ are called outliers. Alternatively, a mean shift outlier model can be formulated

$$y = X\beta + A_j\delta + \epsilon,$$

where

$$A_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

Observation j has a statistical effect on y if δ is significantly different from zero. The estimated t-value is the same as $\hat{u}_j^*$. This procedure does not separate whether the outlier j is due to unusual y- or unusual x-values.

Hadi (1992) proposes an outlier detection with respect to all regressors. The decision whether the design matrix X contains outliers is based on an elliptical distance

$$d_i(c, W) = \sqrt{(x_i - c)'W(x_i - c)},$$

where intuitively the classical choices of c and W are the arithmetic mean $(\bar{x})$ and the inverse of the sample covariance matrix (S^{-1}) of the estimation function of β, respectively, so that the Mahalanobis distance follows. If

$$d_i\left(\bar{x}, S^{-1}\right)^2 > \chi_K^2,$$

observation i is identified as an outlier. As $\bar{x}$ and S react sensitive to outliers it is necessary to estimate an outlier-free mean and sample covariance matrix. For this purpose, only outlier-free observations are considered to determine $\bar{x}$ and S. Another way to avoid the sensitivity problem is to use more robust estimators of the location and covariance matrix, e.g. the median but not the mean is robust to outliers. Finally, an outlier vector MOD (multiple outlier dummy) instead of A is incorporated in the model in order to test whether the identified outlier observations have a significant effect on the dependent variable. A second problem is whether we should eliminate all outliers or only some of them or no outlier. The situation is obvious if an outlier is induced by measurement errors. Then we should eliminate this observation if we have no information to correct the error. Typically, however, we cannot be sure that an anomalous value is due to measurement errors. Insofar, the correct estimation is based between the two extremes: all outliers are considered or all outliers are eliminated. A solution is presented in the next subsection.

2.1.6 *Partially identified parameters*

Assume that some observations are unknown or not exactly measured. Consequence is that a parameter cannot exactly be determined but only within a range. The outlier situation leads to such a partial identification problem. There exist many other similar constellations.

Example The share of unemployed persons is 8 % but 5 % have not answered to the question of the employment status. Therefore, the unemployment rate can only be calculated within certain limits, namely between the two extremes:

- all persons who have not answered are employed
- all persons who have not answered are unemployed.

In the first case the unemployment rate is 7.6 % and in the second case 12.6 %.

The main methodological focus of partially identified parameters is the search for the best statistical inference. Chernozhukov et al. (2007), Imbens and Manski (2004), Romano and Shaikh (2010), Stoye (2009) and Woutersen (2009) have discussed solutions.

If $\Theta_0 = [\theta_l, \theta_u]$ describes the lower and the upper bound based on the two extreme situations Stoye (2009) develops the following confidence interval

$$CI_\alpha = \left[\hat{\theta}_l - \frac{c_\alpha \hat{\sigma}_l}{\sqrt{n}}, \hat{\theta}_u - \frac{c_\alpha \hat{\sigma}_l}{\sqrt{n}}\right],$$

where $\hat{\sigma}_l$ is the standard error of the estimation function $\hat{\theta}_l$. c_α is chosen by

$$\Phi\left(c_\alpha + \frac{\sqrt{n}\hat{\Delta}}{\hat{\sigma}_l}\right) - \Phi(-c_\alpha) = 1 - \alpha,$$

where $\Delta = \theta_u - \theta_l$. As Δ is unknown, the interval has to be estimated ($\hat{\Delta}$).

2.2 Treatment evaluation

The objective of treatment evaluation is the determination of causal effects of economic measures. The simplest form to measure the effect is to estimate α in the linear model

$$y = X\beta + \alpha D + u,$$

where D is the intervention variable and measured by a dummy: 1 if an individual or an establishment is assigned to treatment; 0 otherwise. Typically, this is not the causal effect. An important reason for this failure are unobserved variables that influence y and D, when D and u correlate.

In the last 20 years a wide range of methods was developed to determine the "correct" causal effect. Which approach should be preferred depends on the data, the behavior of the economic agents and the assumptions of the model. The major difficulty is that we have to compare an observed situation with an unobserved situation. Depending on the available information the latter is estimated. We have to ask what would occur if not $D = 1$ but $D = 0$ (treatment on the treated) would take place. This counterfactual is unknown and has to be estimated. Inversely, if $D = 0$ is observable we can search for the potential result under $D = 1$ (treatment on the untreated). A further problem is the fixing of the control group. What is the meaning of "otherwise" in the definition of D? Or in other words: What is the causal effect of an unobserved situation? Should we determine the average causal effect or only that of a subgroup?

Neither a before-after comparison $(\bar{y}_1|D = 1) - (\bar{y}_0|D = 1)$ nor a comparison of $(\bar{y}_t|D = 1)$ and $(\bar{y}_t|D = 0)$ in cross-section is usually appropriate. *Difference-in-differences estimators* (DiD), a combination of these two methods, are very popular in applications

$$\bar{\Delta}_1 - \bar{\Delta}_0 = \left[(\bar{y}_1|D = 1) - (\bar{y}_1|D = 0)\right] - \left[(\bar{y}_0|D = 1) - (\bar{y}_0|D = 0)\right].$$

The effect can be determined in the following unconditional model

$$y = a_1 + b_1 T + b_2 D + b_3 T D + u,$$

where $T = 1$ means a period that follows the period of the measure ($D = 1$). $T = 0$ is a period before the measure takes place. In this approach $\hat{b}_3 = \bar{\Delta}_1 - \bar{\Delta}_0$ is the causal effect. The equation can be extended by further regressors X. This is called a conditional DiD estimator. Nearly all DiD investigations neglect a potential bias in standard error estimates induced by serial correlation. A further problem results under endogenous intervention variables. Then an instrumental variables estimator should be employed avoiding the endogeneity bias. This procedure will be considered in the quantile regression analysis. If the dependent variable is a dummy a nonlinear estimator has to be applied. Suggestions are presented by Ai and Norton (2003) and Puhani (2012).

Matching procedures were developed with the objective to find a control group that is very similar to the treatment group. Parametric and non-parametric procedures can be employed to determine the control group. Kernel, inverse probability, radius matching, local linear regression, spline smoothing or trimming estimators are possible. Mahalanobis metric matching with or without propensity scores and nearest neighbor matching with or without caliper are typical procedures—see e.g. Guo and Fraser (2010). The Mahalanobis distance is defined by

$$(u - v)' S^{-1} (u - v),$$

where u (v) is a vector that incorporates the values of matching variables of participants (non-participants) and S is the empirical covariance matrix from the full set of non-treated participants.

An observed or artificial statistical twin can be determined to each participant. The probability of all non-participants to participate on the measure is calculated based on probit estimates (*propensity score*). The statistical twin j of a participant i is that who has a propensity score (ps_j) nearest to that of the participant. The absolute distance between i and j may not exceed a given value ϵ

$$|ps_i - ps_j| < \epsilon,$$

where ϵ is a predetermined tolerance (caliper). A quarter of a standard deviation of the sample estimated propensity scores is suggested as the caliper size (Rosenbaum and Rubin 1985). If the control group is identified the causal effect can be estimated using the reduced sample (treatment observations and matched observations). In applications α from the model $y = X\beta + \alpha D + u$ or b_3 from the DiD approach is determined as causal effect. Both estimators implicitly assume that the causal effect is the same for all subgroups of individuals or firms and that no unobserved variables exist that are correlated with observed variables. Insofar matching procedures suffer from the same problem as OLS estimators.

If the interest is to detect whether and in which amount the effects of intervention variables differ between the percentiles of the distribution of the objective variable y a quantile regression analysis is an appropriate instrument. The objective is to determine *quantile treatment effects* (QTE). The distribution effect of a measure can be estimated by the difference Δ of the dependent variable with (y_1) and without (y_0) treatment ($D = 1;\ D = 0$) separate for specific quantiles Q^τ where $0 < \tau < 1$

$$\Delta^\tau = Q^\tau_{y1} - Q^\tau_{y0}.$$

The empirical distribution function of an observed situation and that of the counterfactual is identified. From the view of modeling four major cases are developed in the literature that differ in the assumptions. The measure is assumed exogenous or endogenous and the effect on y is unconditional or conditional analogously to DiD.

	Unconditional	Conditional
Exogenous	(1) Firpo (2007)	(2) Koenker and Bassett (1978)
Endogenous	(3) Frölich and Melly (2012)	(4) Abadie et al. (2002)

In case (1) the quantile treatment effect $Q^\tau_{y1} - Q^\tau_{y0}$ is estimated by

$$Q^\tau_{yj} = \arg\min_{\alpha_0;\alpha_1} E\big[\rho_\tau(y - q_j)(W|D = j)\big],$$

where $j = 0; 1$, $q_j = \alpha_0 + \alpha_1(D|D = j)$, $\rho_\tau = a(\tau - \mathbf{1}(a \leq 0))$ is a check function; a is a real number. The weights are

$$W = \frac{D}{p(X)} + \frac{1 - D}{1 - p(X)}.$$

The estimation is characterized by two stages. First, the propensity score is determined by a large number of regressors X via a nonparametric method—$\hat{p}(X)$. Second, in Q^τ_{yj} the probability $p(X)$ is substituted by $\hat{p}(X)$.

Case (2) follows Koenker and Bassett (1978).

$$\sum_{(i|y_i \geq x_i'\beta)=1}^{n_1} \tau \cdot \left|y_i - \alpha(D_i|D_i = j) - x_i'\beta\right|$$

$$+ \sum_{(i|y_i < x_i\beta)=n_1+1}^{n} (1 - \tau) \cdot \left|y_i - \alpha(D_i|D_i = j) - x_i'\beta\right|$$

has to be minimized with respect to α and β, where τ is given. In other words,

$$Q^\tau_{yj} = \arg\min_{\alpha;\beta} E\big[\rho_\tau(y - q_j)(W|D = j)\big],$$

where $j = 0; 1$, $q_j = \alpha(D|D = j) + x'\beta$.

The method of case (3) is developed by Frölich and Melly (2012). Due to the endogeneity of the intervention variable D, an instrumental variables estimator is used with only one instrument Z and this is a dummy. The quantiles follow from

$$Q^\tau_{y^j|c} = \arg\min_{\alpha_0;\alpha_1} E\big[\rho_\tau(y - q_j) \cdot (W|D = j)\big],$$

where $j = 0; 1$, $q_j = \alpha_0 + \alpha_1(D|D = j)$, c means complier. The weights are

$$W = \frac{Z - p(X)}{p(X)(1 - p(X))}(2D - 1).$$

Abadie et al. (2002) investigate case (4) and suggest a weighted linear quantile regression. The estimator is

$$Q^\tau_{yj} = \arg\min_{\alpha,\beta} E\big[\rho_\tau\big(y - \alpha D - x'\beta\big)(W|D = j)\big],$$

where the weights are

$$W = 1 - \frac{D(1 - Z)}{1 - p(Z = 1|X)} - \frac{(1 - D)Z}{p(Z = 1|X)}.$$

Regression discontinuity (RD) design allows to determine treatment effects in a special situation. This approach

uses information on institutional and legal regulations that are responsible that changes occur in the effects of economic measures. Thresholds are estimated indicating discontinuity of the effects. Two forms are distinguished: sharp and fuzzy RD. Either the change of the status is exactly effective at a fixed point or it is assumed that the probability of a treatment change or the mean of a treatment change is discontinuous.

In the case of *sharp RD* individuals or establishments ($i = 1, \ldots, n$) are assigned to the treatment or the control group on the base of the observed variable S. The latter is a continuous or an ordered categorial variable with many parameter values. If variable S_i is not smaller than a fixed bound $\bar{S}$ then i belongs to the treatment group ($D = 1$)

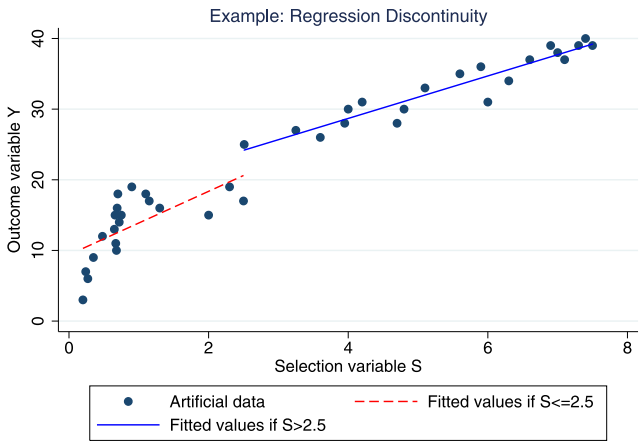
$$D_i = 1[S_i \geq \bar{S}].$$

The following graph based on artificial data with $n = 40$ demonstrates the design. Assuming we know that an institutional rule changes the conditions if $S > \bar{S} = 2.5$ and we want to determine the causal effect induced by the adoption of the new rule. This can be measured by the difference of the two estimated regressions at $\bar{S}$.

In a simple regression model $y = \beta_0 + \beta_1 D + u$ the OLS estimator of β_1 would be inconsistent when D and u correlate. If, however, the conditional mean $E(u|S, D) = E(u|S) = f(S)$ is additionally incorporated in the outcome equation ($y = \beta_0 + \beta_1 D + f(S) + \epsilon$, where $\epsilon = y - E(y|S, D)$), the OLS estimator of β_1 is consistent. Assume $f(S) = \beta_2 S$, the estimator of β_1 corresponds to the difference of the two estimated intercepts of the parallel regressions

$$\hat{y}_0 = \hat{E}(y|D = 0) = \hat{\beta}_0 + \hat{\beta}_2 S$$
$$\hat{y}_1 = \hat{E}(y|D = 1) = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 S.$$

The sharp RD approach identifies the causal effect by distinguishing between the nonlinear function due to the discontinuous character and the smoothed linear function. If, however, a nonlinear function of the general type $f(S)$ is given, modifications have to be regarded.



Assume, the true function $f(S)$ is a polynomial of pth order

$$y_i = \beta_0 + \beta_1 D_i + \beta_{21} S_i + \beta_{22} S_i^2 + \cdots + \beta_{2p} S_i^p + u_i$$

but two linear models are estimated, then the difference between the two intercepts, interpreted as the causal effect, is biased. What looks like a jump is in reality a neglected nonlinear effect.

Another strategy is to determine the treatment effect exactly at the fixed discontinuity point $\bar{S}$ assuming a local linear regression. Two linear regressions are considered

$$y_0 - E(y_0|S = \bar{S}) = \delta_0(S - \bar{S}) + u_0$$
$$y_1 - E(y_1|S = \bar{S}) = \delta_1(S - \bar{S}) + u_1,$$

where $y_j = E(y|D = j)$ and $j = 0; 1$. In combination with

$$y = (1 - D)y_0 + Dy_1$$

follows

$$y = (1 - D)\big(E(y_0|S = \bar{S}) + \delta_0(S - \bar{S}) + u_0\big) + D\big(E(y_1|S = \bar{S}) + \delta_1(S - \bar{S}) + u_1\big).$$

The linear regression

$$y = \gamma_0 + \gamma_1 D + \gamma_2(S - \bar{S}) + \gamma_3 D(S - \bar{S}) + \tilde{u}$$

can be estimated, where $\tilde{u} = u_0 + D(u_1 - u_0)$. This looks like the DiD estimator but now $\gamma_1 = E(y_1|S = \bar{S}) - E(y_0|S = \bar{S})$ and not γ_3 is of interest. The estimated coefficient $\hat{\gamma}_1$ is a global but not a localized average treatment effect.

The localized average follows if a small interval around $\bar{S}$ is modeled, i.e. $\bar{S} - \Delta S < S_i < \bar{S} + \Delta S$. The treatment effect corresponds to the difference of the two former determined intercepts, restricted to $\bar{S} < S_i < \bar{S} + \Delta S$ on the one hand and to $\bar{S} - \Delta S < S_i < \bar{S}$ on the other hand.

A combination of the latter linear RD model with the DiD approach leads to an extended interaction model. Again, two linear regressions are considered

$$y_0 = \gamma_{00} + \gamma_{10} D + \gamma_{20}(S - \bar{S}) + \gamma_{30} D(S - \bar{S}) + \tilde{u}_0$$
$$y_1 = \gamma_{01} + \gamma_{11} D + \gamma_{21}(S - \bar{S}) + \gamma_{31} D(S - \bar{S}) + \tilde{u}_1,$$

where the first index of γ_{jt} with $j = 0; 1$ refers to the treatment and the second index with $t = 0; 1$ refers to the period. In contrast to the pure RD model, where y_j and $j = 0; 1$ is considered, now the index of y is a time index, i.e. y_T and $T = 0; 1$. Using

$$y = (1 - T)y_0 + Ty_1$$

follows

$$
\begin{aligned}
y &= \gamma_{00} + \gamma_{10} D + \gamma_{20}(S - \bar{S}) + \gamma_{30} D(S - \bar{S}) \\
&\quad + (\gamma_{01} - \gamma_{00})T + (\gamma_{11} - \gamma_{10})DT \\
&\quad + (\gamma_{21} - \gamma_{20})(S - \bar{S})T \\
&\quad + (\gamma_{31} - \gamma_{30})D(S - \bar{S})T + \big(\tilde{u}_0 + (\tilde{u}_1 - \tilde{u}_0)T\big) \\
&=: \beta_0 + \beta_1 T + \beta_2 D + \beta_3(S - \bar{S}) + \beta_4 D(S - \bar{S}) \\
&\quad + \beta_5 DT + \beta_6(S - \bar{S})T + \beta_7 D(S - \bar{S})T + \tilde{u}.
\end{aligned}
$$

Now, it is possible to determine whether the treatment effect varies between $T = 1$ and $T = 0$. The difference follows by a DiD approach

$$
\begin{aligned}
&\big[(y_1|D = 1) - (y_1|D = 0)\big] - \big[(y_0|D = 1) - (y_0|D = 0)\big] \\
&\quad = (\gamma_{11} - \gamma_{10}) + (\gamma_{31} - \gamma_{30})(S - \bar{S}) \\
&\quad = \beta_5 + \beta_7(S - \bar{S})
\end{aligned}
$$

under the assumption that the disturbance term does not change between the periods. The hypothesis of a time-invariant break cannot be rejected if DT and $D(S - \bar{S})T$ have no statistical influence on y.

The *fuzzy RD* assumes that the propensity score function of treatment $P(D = 1|S)$ is discontinuous with a jump in $\bar{S}$

$$
P(D_i = 1|S_i) = \begin{cases} g_1(S_i) & \text{if } S_i \geq \bar{S} \\ g_0(S_i) & \text{if } S_i < \bar{S}, \end{cases}
$$

where it is assumed that $g_1(\bar{S}) > g_0(\bar{S})$. Therefore, treatment in $S_i \geq \bar{S}$ is more likely. In principle, the functions $g_1(S_i)$ and $g_0(S_i)$ are arbitrary, e.g. a polynomial of pth order can be assumed but the values have to be within the interval [0; 1] and different values in $\bar{S}$ are necessary.

The conditional mean of D that depends on S is

$$
\begin{aligned}
E(D_i|S_i) &= P(D_i = 1|S_i) \\
&= g_0(S_i) + \big(g_1(S_i) - g_0(S_i)\big)T_i,
\end{aligned}
$$

where $T_i = 1(S_i \geq \bar{S})$ is a dummy indicating the point where the mean is discontinuous. If a polynomial of pth order is assumed the interaction variables $S_i T_i$, $S_i^2 T_i \cdots S_i^p T_i$ and the dummy T_i are instruments of D_i. The simplest case is to use only T_i as an instrument if $g_1(S_i)$ and $g_0(S_i)$ are discriminable constants.

We can determine the treatment effect around $\bar{S}$

$$
\lim_{\Delta \to 0} \frac{E(y_i|\bar{S} < S_i < \bar{S} + \Delta) - E(y_i|\bar{S} - \Delta < S_i < \bar{S})}{E(D_i|\bar{S} < S_i < \bar{S} + \Delta) - E(D_i|\bar{S} - \Delta < S_i < \bar{S})}.
$$

The empirical analogon is the Wald (1940) estimator that was first developed for the case of measurement errors

$$
\frac{(\bar{y}|\bar{S} < S_i < \bar{S} + \Delta) - (\bar{y}|\bar{S} - \Delta < S_i < \bar{S})}{(\bar{D}|\bar{S} < S_i < \bar{S} + \Delta) - (\bar{D}|\bar{S} - \Delta < S_i < \bar{S})}.
$$

QTE and RD analysis allow the determination of variable causal effects with a different intention. A further possibility is a separate estimation for subgroups, e.g. for industries or regions.

3 Applications: Some New Estimates of Cobb-Douglas Production Functions

This section presents some estimates of production functions, where IAB establishment panel data are used. The empirical analysis is restricted to the period 2006–2010. The decision to start with 2006 is the following: in this year information on company level-pacts (CLPs) were collected in the IAB establishment panel for the first time and many of the following applications deal with CLPs. Methods of Sect. 2 are applied. The intention of Sect. 3 is to illustrate that the discussed methods work with implemented STATA programmes. It is not discussed whether the applied methods are best for the given data set and the substantial problems. From a didactical perspective the paper is always concerned with only one issue and different suggestions to solve the problem are compared. The results can be found in Tables 1–10.

Table 1 focus on alternative estimates of standard errors—see Sects. 2.1.1–2.1.3—of Cobb-Douglas production functions (CDF) in the logarithm representation with the input factors $\ln L$ and $\ln K$. The estimation of conventional standard errors can be found for comparing in Table 3, column 1. The small standard deviations and therefore the large t-values are remarkable. Though the cluster-robust standard errors in Table 1, column 5 are larger, they are still by far too low. This is due to unobserved heterogeneity. Fixed effects estimates can partially solve this problem as can be seen in the Appendix, Table 15.

The estimated coefficients in column 1–3 and 5 of Table 1 are identical. Estimates with $hc2$ and $hc4$—not presented in the tables—deviate only slightly from those with $hc1$. This could mean that it is not necessary to distinguish between $hc1$ to $hc4$. However, one could guess that stronger differences are observed if the sample is small. Empirical investigations, where only 10, 1 and 0.1 percent of the original sample size is used, do not support this presumption. The jackknife estimates of standard errors and t-values are also not so far away from the heteroskedasticity-consistent estimates with $hc1$ and $hc3$. The nearness to estimates with $hc3$ is plausible because the latter is only a slightly simplified version of what one gets by employing the jackknife technique. Furthermore, Table 1 demonstrates that bootstrap and cluster-robust estimates of the t-values differ strongest of the input factor labor ($\ln L$), measured by the number of employees in the firm. Capital ($\ln K$), approximated by the sum of investments of the last four years, has evidently

Table 1 Estimates of Cobb-Douglas production functions under alternative determination of standard errors using $hc1$, $hc3$, bootstrap, jackknife and cluster-robust estimates

	$hc1$	$hc3$	bootstrap	jackknife	cluster (idnum)
$\ln L$	0.9472	0.9472	0.9472	0.9582	0.9472
	(184.02)	(183.99)	(227.40)	(184.49)	(126.29)
$\ln K$	0.2225	0.2225	0.2225	0.2178	0.2225
	(60.80)	(60.79)	(60.40)	(59.58)	(43.04)
const	9.0810	9.0810	9.0810	9.0908	9.0810
	(307.86)	(307.81)	(271.82)	(308.83)	(215.20)

Note: $n = 34{,}308$; $R^2 = 0.843$; t-ratios in parentheses; idnum—identification number of the firm

Table 2 OLS estimates of an extended CDF with Bernoulli distributed regressors

	Mean	Coef.	Std.err.	t
$\ln L$		0.8808	0.0061	144.33
$\ln K$		0.2049	0.0041	49.55
CLP	0.0871	0.0307	0.0236	1.30
WOCO	0.3035	0.3915	0.0184	21.19
CB	0.3819	0.1385	0.0133	10.36
P1	0.0834	0.2462	0.0231	10.65
P2	0.3695	0.1032	0.0132	7.78
const		9.2905	0.0367	253.03

Note: $n = 20{,}332$; $R^2 = 0.846$. The regressors CLP (company-level pact), WOCO (works council), CB (collective industry-wide bargaining), P1 (profits last year: very good) and P2 (profits last year: good) are dummies

Table 3 OLS estimates of CDFs with and without outliers, t-values in parentheses; dependent variable: logarithm of sales—$\ln Y$

	With outliers	Without outliers	Without strong leverages	With Hadi-MOD
$\ln L$	0.9472	0.9415	1.0409	0.9412
	(222.12)	(240.28)	(169.10)	(240.10)
$\ln K$	0.2225	0.2242	0.1724	0.2243
	(70.11)	(77.04)	(36.33)	(77.08)
MOD				1.8810
				(2.33)
const	9.0811	9.0498	9.3445	9.0490
	(333.20)	(362.66)	(238.53)	(362.62)
n	34,308	33,851	27,262	34,308
R^2	0.866	0.866	0.805	0.843

larger cluster-robust estimates of standard errors than that from the other methods.

An extended version of the Cobb-Douglas function in Table 1 is presented in Table 2. The latter estimates show smaller coefficients and smaller t-values of the input factors labor and capital. The major intention of Table 2 is to demonstrate that also in this example there is—as maintained in Sect. 2.1.4—a clear relationship between $\bar{D}$, the mean of a dummy as independent variable, and the estimated standard errors. The nearer $\bar{D}$ to 0.5 the smaller is the standard error. The results in Table 2 cannot be generalized in contrast to that in Table 11 because the standard error of a dummy is not only determined by the mean. Each regressor has a specific influence on the dependent variable independent of the regressor's variance.

Outliers—see Sect. 2.1.5—may have strong effects on coefficient and standard error estimates. However, estimates do not react sensitively to all outliers. This can be demonstrated if the results with and without outliers are compared. Table 3 presents an example for simple Cobb-Douglas functions in column 1 and 2. An observation in column 2 is defined as an outlier if $|\hat{u}^*| > 3$. The coefficients in column 1 and 2 are very similar while the differences of the standard errors become more evident. The differences are enlarged

Table 4 Confidence intervals (CI) of output elasticities of labor and capital based on a Cobb-Douglas production function, estimated with and without outliers, Stoye's confidence interval at partially identified parameters; dependent variable: logarithm of sales—$\ln Y$

	CI with outliers	CI without outliers	Stoye CI
$\hat{\beta}_{\ln L;u}$	0.9555	0.9492	0.9511
$\hat{\beta}_{\ln L;l}$	0.9388	0.9339	0.9376
$\hat{\beta}_{\ln K;u}$	0.2287	0.2299	0.2282
$\hat{\beta}_{\ln K;l}$	0.2162	0.2185	0.2184
$\Delta\hat{\beta}_{\ln L}$	0.0167	0.0153	0.0135
$\Delta\hat{\beta}_{\ln K}$	0.0125	0.0114	0.0098

under a wider definition of an outlier, e.g. if 3 is substituted by 2. The picture becomes also clearer if observations with high leverage are eliminated—see column 3. Coefficients and standard errors in column 1 and 3 reveal a clear disparity for both input factors. This result is not unexpected but the consequence is ambiguous. Is column 1 or 3 preferable? If all observations with strong leverages are due to measurement errors the decision speaks in favor of the estimates in column 3. As no information is available to this question both estimates may be useful.

Column 4 extends the consideration to outliers following Hadi (1992).The squared difference between individual regressor values and the mean for all regressors—here $\ln L$ and $\ln K$—is determined for each observation weighted by the estimated covariance matrix—see Sect. 2.1.5. The decision whether establishment i is an outlier is now based on the Mahalanobis distance. MOD, the vector of multiple outlier dummies ($\text{MOD}_i = 1$ if i is an outlier; $=0$ otherwise), is incorporated as an additional regressor. The estimates show that outliers have a significant effect on the output variable $\ln Y$. The coefficients and the t-values in column 2 and 4 are very similar. This is a hint that the outliers defined via $\hat{u}^*$ are mainly determined by large deviations of the regressor values. From $\hat{u}^*$ it is unclear whether the values of the dependent variable or the independent variables are responsible for the fact that an observation is an outlier.

As it is not obvious whether the outliers are due to measurement errors that should be eliminated or whether these are unusual but systematically induced observations that should be accounted for, parameters can only partially be identified. Therefore, in Table 4 confidence intervals are not only presented for the two extreme cases (column 1: all outliers are induced by specific events; column 2: all outliers are due to random measurement errors). Additionally, in column 3 the confidence interval (CI) based on Stoye's method is displayed. The results show that the lower and upper coefficient estimates of $\ln L$ by Stoye lies within the estimated coefficients in column 1 and 2. The upper coefficient is nearer to that of column 2 and the lower is nearer to column 1. We do not find the same pattern for input factor $\ln K$. In this case Stoye's $\hat{\beta}_{\ln K;u}$ deviates more from that in column 2 than in column 1. And for $\hat{\beta}_{\ln K;l}$ we find the opposite result. Stoye's intervals ($\Delta\hat{\beta}_{\ln L} = \hat{\beta}_{\ln L;u} - \hat{\beta}_{\ln L;l}$; $\Delta\hat{\beta}_{\ln K} = \hat{\beta}_{\ln K;u} - \hat{\beta}_{\ln K;l}$) are shorter than that with or without outliers. In other words, the estimates are more precise.

Table 5 Unconditional and conditional DiD estimates with company-level pact (CLP) effects; dependent variable: logarithm of sales—$\ln Y$

	Unconditional	Conditional
$\ln L$		0.9423
		(166.03)
$\ln K$		0.2211
		(53.37)
CLP	3.1152	0.0951
	(35.91)	(2.36)
D2009	0.0597	0.0216
	(2.25)	(1.54)
CLP $*$ D2009	−0.3029	0.0400
	(−2.90)	(0.84)
n	31,985	20,490
R^2	0.101	0.841

Note: t-values in parentheses

The next tables present estimates of alternative methods in order to determine causal effects. First, the difference-in-differences (DiD) approach is estimated. Results can be found in Table 5. The coefficient of the interaction variable CLP $*$ D2009 in column 1 is significantly different from zero. This means that sales between firms with a company-level pact (CLP), adopted in 2009, and those without such a pact differ between 2009 and the years before (2006–2008). The adoption of a CLP in the year of the Great Recession is combined with lower sales than in the years before if an unconditional DiD specification is used. In column 2 the sign changes and the effect of the interaction variable is insignificant if an extended CDF is estimated. This approach is preferred because in the former the influence of the input factors is partially added to the causal effect. Now, no influence of the adoption of a CLP on sales in 2009 can be detected. One could argue that the estimates in column 1 lead more than that in column 2 to significant results because the sample in the former is larger. This argument is not compelling. If we draw a random sample of 63.83 percent so that in column 1 the sample size is $n = 20{,}489$ the interaction effect is -0.2939 and the significance is preserved ($t = -2.26$). If CLPs change labor and capital productivity we should not incorporate $\ln L$ and $\ln K$ in a conditional DiD. In other words, in this case we should not control for these variables before treatment.

Table 6 Estimates of CDFs with CLP effects using matching procedures; dependent variable: logarithm of sales—ln Y

	No matching	MM	NNM
$\ln L$	0.9420	0.9362	0.9533
	(166.03)	(47.75)	(63.32)
$\ln K$	0.2212	0.1938	0.2007
	(53.42)	(15.12)	(19.70)
CLP	0.1231	0.1928	0.0496
	(5.22)	(1.31)	(1.46)
n	20,490	1,806	3,346
R^2	0.840	0.838	0.849

Note: MM—Mahalanobis metric matching, NNM—nearest neighbor matching, t-values in parentheses. Matching variables are profit situation, year in which the establishment was founded, introduction of new products, further training, average working time, working time accounts, opening clause

Alternative methods to determine causal effects are matching procedures. These are suggested when there does not exist control over the assignment of treatment conditions, when in the basic equation $y = X\beta + \alpha D + u$ the dichotomous treatment variable D and the disturbance term u correlate, when the ignorable treatment assignment assumption is violated. In the example of the CDF it is questioned that this condition is fulfilled for CLPs. As an alternative the Mahalanobis metric matching (MM) without propensity score and the nearest neighbor matching (NNM) with caliper are applied, presented in Table 6, column 2 and 3, respectively. In the latter method non-replacement is used. That is, once a treated case is matched to a non-treated case, both cases are removed from the pool. The former method allows that one control case can be used as a match for several treated cases. Therefore, the total number of observations in the nearest neighbor is larger than that in column 2. We find that the CLP effect on sales is insignificant in both cases but the CLP coefficient of MM estimates exceeds by far that of NNM. The estimates of the partial elasticities of production are very similar in the three estimates in Table 6. The insignificance of the CLP effect confirms the result of column 2 of Table 5. If the DiD estimator of column 2 in Table 5 is applied after matching the causal effect is—not unexpected—also insignificant. The probvalue is 0.182 if the MM procedure is used and 0.999 under the NNM procedure.

The previous estimates have demonstrated that company-level pacts (CLP) have no statistically significant influence on output. We cannot be sure that this result is also true for subgroups of firms. One way to test this is to conduct quantile estimates. As presented in Sect. 2.2 four methods can be applied to determine quantile treatment effects (QTE). The CLP effects on sales can be found in Table 7 where the results of five quantiles ($q = 0.1, 0.3, 0.5, 0.7, 0.9$) are presented. In contrast to the previous estimations most CLP effects are significant in the columns 1–4 of Table 7. Firpo considers the simplest case without control variables under the assumption that the adoption of a company-level pact is exogenous. The estimated coefficients in column 1 (F) seem oversized. The same follows from the Frölich-Melly approach, where CLP is instrumented by a short work time dummy (column 3—F-M). Other available instruments like opening clauses, collective bargaining, works councils or research and development within the firm do not evidently change the results. One reason for the overestimated coefficients can be neglected determinants of the output that correlate with CLP. Estimates of column 2 (K-B) and 4 (A-A-I) support this hypothesis.

From the view of expected CLP coefficients the conventional quantile estimator, the Koenker-Bassett approach, with $\ln L$ and $\ln K$ as regressors seems best. However, the ranking of the size of the coefficients within column 2 seems unexpected. The smaller the quantile the larger is the estimated coefficient. This could mean that CLPs are advantageous for small firms. However, it is possible that small firms with advantages in productivity due to CLPs have relative high costs to adopt a CLP. In this case the higher propensity of large firms to introduce a CLP is consistent with higher productivity of small firms.

The coefficients of the Abadie-Angrist-Imbens approach, a combination of Frölich-Melly's and Koenker-Bassett's model, are also large but not so large as in column 1 and 3.

Possibly, all estimates in column 1–4 of Table 7 are biased and inconsistent. This is the case when CLP and non-CLP firms fundamentally differ due to unobserved variables. To avoid this problem the QTE and the matching approaches are combined. Based on the matching of Table 6 the QTE analogously to column 1–4 in Table 7 can be estimated. In column 5 and 6 only two combinations are presented, namely MM + K-B and MM + A-A-I. We find that the ranking and the size of the coefficients are plausible in column 5. The sizes of the coefficients in column 6 are smaller than in column 4 but the identified causal effects seems still too high. The most important result is the following: the CLP effects are significant for higher quantiles, i.e. for $q = 0.9$ in column 5 and for $q = 0.7$ and $q = 0.9$ in column 6. However, the median estimators ($q = 0.5$) of CLP effects in column 5 and 6 that can be compared with the estimates of column 2 in Table 6 are insignificant. Quantile estimators highlight information that cannot be revealed by other treatment methods, i.e. in Tables 5 and 6. The estimations of the other six combinations (MM + F, MM + F-M, NNM + F, NNM + K-B, NNM + F-M, NNM + A-A-I)—not presented in the tables—are less plausible. The ranking of the size of coefficients is inconsistent in the light of theoretical and practical experience.

Table 7 Quantile estimates of CLP effects; dependent variable: logarithm of sales—ln Y

Quantile	F	K-B	F-M	A-A-I	MM + K-B	MM + A-A-I
$q = 0.1$	2.9957	0.2236	5.3012	1.2092	−0.1064	0.9776
	(38.94)	(6.76)	(20.42)	(3.10)	(−0.87)	(1.06)
$q = 0.3$	3.3242	0.1836	5.8227	1.1615	0.0715	0.7140
	(54.67)	(7.15)	(23.67)	(3.11)	(0.46)	(0.62)
$q = 0.5$	3.1325	0.1526	6.3549	1.2000	0.1793	0.6736
	(54.19)	(6.31)	(24.58)	(2.57)	(1.09)	(1.37)
$q = 0.7$	2.9312	0.1036	6.8703	1.2479	0.2270	0.8072
	(56.91)	(4.07)	(26.14)	(2.09)	(1.54)	(2.18)
$q = 0.9$	2.3203	−0.0176	7.8119	1.6549	0.4523	1.4242
	(34.18)	(−0.37)	(20.12)	(1.36)	(3.36)	(2.92)
n	31,985	20,490	20,909	13,496	1,806	1,206

Note: F—Firpo; K-B—Koenker/Bassett; F-M—Frölich/Melly; A-A-I—Abadie/Angrist/Imbens, MM—Mahalanobis metric matching, control variables are ln L and ln K, t-values in parentheses

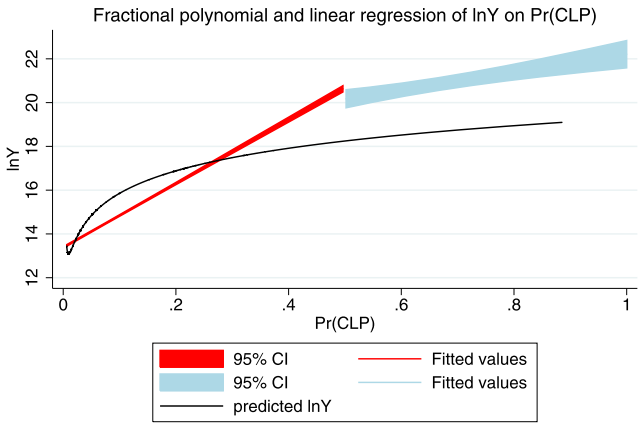


Fig. 1 Regression discontinuity of CLP probability

The final discussed treatment method in Sect. 2.2 is the regression discontinuity (RD) design. This approach exploits information of the rules determining treatment. The probability of receiving a treatment is a discontinuous function of one or more variables where treatment is triggered by an administrative definition or an organizational rule.

In a first example using a sharp RD design it is analyzed whether at an estimated probability of 0.5 that a company-level pact (CLP) exists a structural break on logarithm of output (ln Y) is evident. For this purpose a probit model is estimated with profit situation, working-time account, total wages per year and works council as determinants of CLP. All coefficients are significantly different from zero—not in the tables. The estimated probability Pr(CLP) is then plotted against ln Y based on a fractional polynomial model over the entire range ($0 < \Pr(\text{CLP}) < 1$) and on two linear models split into $\Pr(\text{CLP}) <= 0.5$ and $\Pr(\text{CLP}) > 0.5$. The graphs are presented in Fig. 1.

A structural break seems evident. Two problems have to be checked: First, is the break due to a nonlinear shape, and second, is the break significant? The answer to the first question is yes, because the shape over the range $0 < \Pr(\text{CLP}) < 1$ is obviously nonlinear when a fractional polynomial is assumed. The answer to the second question is given by a t-test—cf. Sect. 2.2—based on

$$
\begin{aligned}
y &= \gamma_0 + \gamma_1 D\_\Pr(\text{CLP}) + \gamma_2\big(\Pr(\text{CLP}) - \overline{\Pr(\text{CLP})}\big) \\
&\quad + \gamma_3 D\_\Pr(\text{CLP}) \cdot \big(\Pr(\text{CLP}) - \overline{\Pr(\text{CLP})}\big) + u \\
&=: \gamma_0 + \gamma_1 D\_\Pr(\text{CLP}) + \gamma_2 c\Pr(\text{CLP}) \\
&\quad + \gamma_3 D\_\Pr(\text{CLP}) \cdot c\Pr(\text{CLP}) + u,
\end{aligned}
$$

where

$$
D\_\Pr(\text{CLP}) = \begin{cases} 1 & \text{if } \Pr(\text{CLP}) \le 0.5 \\ 0 & \text{otherwise.} \end{cases}
$$

The null that there is no break has to be rejected ($\hat{\gamma}_1 = -3.96$; $t = -6.87$; probvalue $= 0.000$) as can be seen in Table 8.

The estimates in Table 8 cannot tell us whether the output jump in $\Pr(\text{CLP}) = 0.5$ is a general phenomenon or whether the Great Recession in 2008/09 is responsible. To test this the combined method of RD and DiD—derived in Sect. 2.2—is employed and the results are presented in Table 9. The estimates show that the output jump does not significantly change between 2006/2007 and 2008/2010. The influence of $D\_\Pr(\text{CLP}) \cdot T$ and that of $D\_\Pr(\text{CLP}) \cdot c\Pr(\text{CLP}) \cdot T$ on ln Y is insignificant. Therefore, we conclude that the break is of general nature.

Two further examples are presented in Fig. 2 and 3. The Institut für Mittelstandsforschung defines small firms

Table 8 Testing for structural break of CLP effects between Pr(CLP) ≤ 0.5 and Pr(CLP) > 0.5

	Coef.	Std.err.	t	$P > \|t\|$
D_Pr(CLP)	−3.9608	0.5765	−6.87	0.000
cPr(CLP)	4.3413	0.8390	5.17	0.000
D_Pr(CLP) · cPr(CLP)	11.3838	0.8437	13.49	0.000
const	18.4375	0.5764	31.99	0.000

Table 9 Testing for differences in structural break of CLP effects between Pr(CLP) ≤ 0.5 and Pr(CLP) > 0.5 in 2006/07 and 2008/10

	Coef.	Std.err.	t	$P > \|t\|$
T	0.0130	1.3118	0.01	0.992
D_Pr(CLP)	−4.1045	1.1191	−3.67	0.000
cPr(CLP)	3.9314	1.6795	2.34	0.019
D_Pr(CLP) · cPr(CLP)	11.6383	1.6884	6.89	0.000
D_Pr(CLP) · T	0.0392	1.3119	0.03	0.976
cPr(CLP) · T	0.2801	1.9520	0.14	0.886
D_CLP · cPr(CLP) · T	−0.0662	1.9623	−0.03	0.973
const	18.5422	1.1190	16.57	0.000

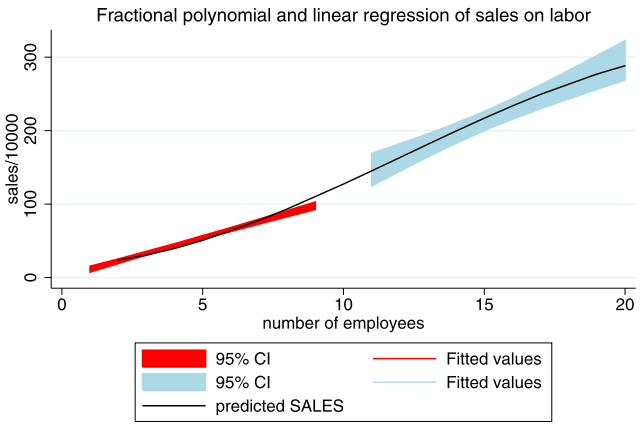


Fig. 2 Regression discontinuity of small firms

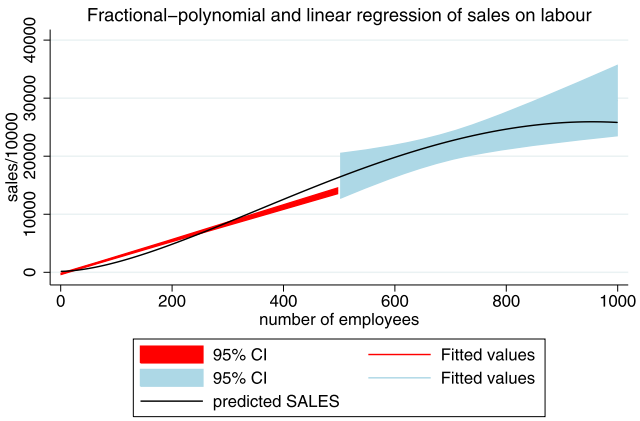


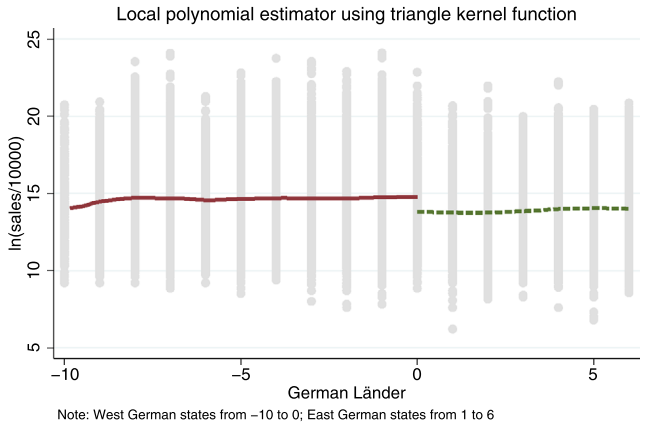
Fig. 3 Regression discontinuity of middle-size firms

as such that have less than 10 employees and until 1 million Euro sales per year. The analogous definition of middle-size firms is less than 500 employees and until 50 million Euro sales per year. A sharp regression discontinuity design is applied to test whether the first and the second part of the definition are consistent. In other words, based on a Cobb-Douglas production function with only one input factor, the number of employees, it is tested whether there exists a structural break for small firms between 9 and 10 employees at a 1 million sales border. We find for small firms in Fig. 2 that there seems to be a sales break around 1 million Euro per year.

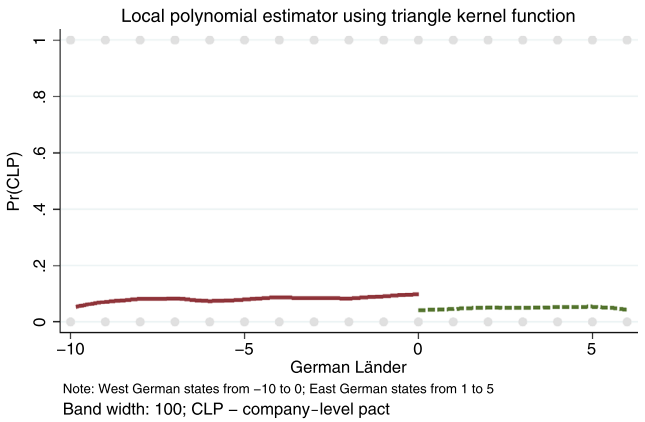
The t-test analogously to the first example yields weak significance ($\hat{\gamma}_1 = -13.8667$; $t = -1.61$; probvalue = 0.107). The same procedure for middle-size firms—see Fig. 3—leads to following results.

Apparently, there exists a break. However, the first part of the definition of middle-size firms from the Institut für Mittelstandsforschung is not compatible with the second part. The break of sales at 500 employees is not 50 million Euro per year but around 150 million Euro. Furthermore, the visual result might be due to a nonlinear relationship as the fractional polynomial estimation over the entire range suggests. The t-test does not reject the null ($\hat{\gamma}_1 = -8977$; $t = -0.54$; probvalue = 0.588). The conclusion from Fig. 2 and 3 is that the graphical representation without the polynomial shape as comparison course and without testing for a structural break can lead to a misinterpretation.

The final example uses a fuzzy regression discontinuity design. It is analyzed whether the CLP effects on the logarithm of sales ($\ln Y = \ln(\text{sales}/10000)$) differ between the East and West German federal states. The graphical representation can be found in Figs. 4a and 4b. The former shows



(a)



(b)

Fig. 4 (**a**) Regression discontinuity of ln(sales). (**b**) Regression discontinuity of treatment CLP

Table 10 Fuzzy regression discontinuity between East and West German federal states (GFS)—Wald test for structural break of company-level pact (CLP) effects on sales; jump at GFS > 0; dependent variable: logarithm of sales—ln *Y*

Variable	Coef.	Std.err.	*z*
ln *Y* jump	−0.8749	0.1234	−7.09
CLP jump	−0.0571	0.0138	−4.13
Wald estimator	15.3165	3.5703	4.29

Note: GFS = −10 Berlin(West); −9 Schleswig-Holstein; −8 Hamburg; −7 Lower Saxony; −6 Bremen; −5 North Rhine-Westphalia; −4 Hesse; −3 Rhineland-Palatinate; −2 Baden-Württemberg; −1 Bavaria; 0 Saarland; 1 Berlin(Ost); 2 Brandenburg; 3 Mecklenburg-West Pomerania; 4 Saxony; 5 Saxony-Anhalt; 6 Thuringia

the disparities in the level of sales per year and the latter those of Pr(CLP)—here measured by the relative frequency of firms with a CLP to all firms in a German federal state.

Although clear differences are detected for both characteristics (ln *Y*, Pr(CLP)) we cannot be sure that these disparities are significant and whether the CLP effects are smaller or larger in West Germany. This is checked by a Wald test in Table 10. We find that the CLP effects on ln *Y* ($-0.8749/-0.0571 = 15.3165$) are significantly higher in the West German federal states ($z = 4.29$). When the interpretation is focussed on the dummy "East Germany" as an instrument of a dummy "CLP" we should note that the former is not a proper instrument because the output ln *Y* differs between East and West Germany independent of a CLP.

4 Summary

Many reasons like heteroskedasticity, clustering, basic probability of qualitative regressors, outliers and only partially identified parameters may be responsible that estimated standard errors based on classical methods are biased. Applications show that the estimates under suggested modifications do not always deviate so much from that of the classical methods.

The development of new procedures is ongoing. Especially, the field of treatment methods were extended. It is not always obvious which method is preferable to determine the causal effect. As the results evidently differ it is necessary to develop a framework that helps to decide which method is most appropriated under typically situations. We observe a tendency away from the estimation of average effects. The focus is shifted to distribution topics. Quantile analysis helps to investigate differences between subgroups of the population. This is important because economic measures have not the same influence on heterogeneous establishments and individuals. A combination of quantile regression with matching procedure can improve the determination of the causal effects. Further combinations of treatment methods seem helpful. Difference-in-differences estimates should be linked with matching procedures and regression discontinuity designs. And also regression discontinuity split to quantiles can lead to new insights.

Executive summary

Empirical economics is governed by econometric methods since many years. During the last 20 years contents and major questions have strongly changed in this field. Therefore methods were modified and completely new methods were developed. In comparison to conventional approaches attention is paid to peculiarities of the data, to the specification of the estimating approach, to unobserved heterogeneity, to endogeneity and causal effects. Real data are often not compatible with the assumptions of classical methods. If the latter are used, this can lead to a misinterpretation of the results. We have to ask, whether the results are correct. Is it really possible to interpret the estimated effects as causal or are these only statistical artifacts, which are irrelevant or even

counterproductive for policy measures? In order to avoid this, the practitioner has to be familiarwith the wide range of existing methods for the empirical investigations. The user has to know the assumptions of the methods and whether the application allows adequate conclusions at given information. It is necessary to check the robustness of the results by alternative methods and specifications.

This paper presents a selective review of econometric methods and demonstrates by applications that the methods work. In the first part, methodological problems to standard errors and treatment effects are discussed. First, heteroskedasticity- and cluster-robust estimates are presented. Second, peculiarities of Bernoulli distributed regressors, outliers and only partially identified parameters are revealed. Approaches to the improvement of standard error estimates under heteroskedasticity differ in the weighting of residuals. Other procedures use the estimated disturbances in order to create a larger number of artificial samples, to obtain better estimates. And again others use nonlinear information. Cluster robust estimates try to solve the Moulton problem. Too low standard errors between observations within clusters are adjusted. This objective is only partially successful. We should be cautious if we compare the effects of dummy variables on an endogenous variable, because the more the mean of dummies deviates from 0.5 the higher are the standard errors. Outliers, i.e. unusual observations that are due to systematic measurement errors or extraordinary events may have enormous influence on the estimates. The suggested approaches to detect outliers vary relating to the measurement concept and do not necessarily demonstrate whether outliers should be accounted for in the empirical analysis. New methods for partially identified parameters may be helpful in this context. Under uncertainty the degree of precision, whether outliers should be eliminated, can be increased.

Four principles to estimate causal effects are in the focus: difference-in-differences (DiD) estimators, matching procedures, quantile treatment effects (QTE) analysis and regression discontinuity design. The DiD models distinguish between conditional and unconditional approaches. The range of the popular matching procedures is wide and the methods evidently differ. They aim to find statistical twins, to homogenize the characteristics of observations from the treatment and the control group. Until now, the application of QTE analysis is relatively rare in practice. Four types of models are important in this context. The user has to decide whether the treatment variable is exogenous or endogenous and whether additional control variables are incorporated or not. Regression discontinuity (RD) designs separate between sharp and fuzzy RD methods. It is distinguished whether an observation is assigned to the treatment or to the control group directly by an observable continuous variable or indirectly via the probability and the mean of treatment, respectively, conditional on this variable.

In the second part of the paper the different methods are applied to estimates of Cobb-Douglas production functions using IAB establishment panel data. Some heteroskedasticity-consistent estimates show similar results while cluster-robust estimates differ strongly. Dummy variables as regressors with a mean near 0.5 reveal as expected smaller variances of the coefficient estimators than others. Not all outliers have a strong effect on the significance. Methods of partially identified parameters demonstrate more efficient estimates than traditional procedures.

The four discussed treatment effects methods are applied to the question whether company-level pacts have a significant effect on the production output. Unconditional DiD estimators and estimates without matching display significantly positive effects. In contrast to this result we cannot find the same if conditional DiD or matching estimates based on the Mahalanobis metric are applied. This outcome has more precisely formulated under quantile regression. The higher the quantile the more is the tendency to positive and significant effects. Sharp regression discontinuity estimates display a jump at the probability 0.5 that an establishment has a company-level pact. No specific influence can be detected during the Great Recession. Fuzzy regression discontinuity estimates reveal that the output effect of company-level pacts is significantly lower in East than in West Germany. A combined application of the four principles determining treatment effects lead to some interesting new insights. We determine joint DiD and matching estimates as well as that ofthe former together with regressions discontinuity designs. Finally, matching is interrelated to quantile regression.

Kurzfassung

Empirische Wirtschaftsforschung wird schon seit vielen Jahren ganz wesentlich von ökonometrischen Methoden getragen. In den letzten 20 Jahren haben sich Inhalte und Fragestellungen in der empirischen Wirtschaftsforschung stark verändert. Dies hat dazu geführt, dass viele Methoden modifiziert oder völlig neue entwickelt wurden. Gegenüber traditionellen Ansätzen wird verstärkt auf die Besonderheiten der Daten, auf die Spezifikation des zu schätzenden Ansatzes, auf unbeobachtete Heterogenität, auf Endogenität und auf Kausaleffekte geachtet. Reale Daten sind ganz überwiegend nicht vereinbar mit den Annahmen klassischer Methoden. Werden letztere trotzdem eingesetzt, so sind damit häufig Fehlinterpretationen der Ergebnisse verbunden. Zu fragen ist, wie sicher die getroffenen Aussagen sind. Können die Schätzergebnisse tatsächlich kausal interpretiert werden oder haben sich lediglich rein statistische Zusammenhänge ergeben, die für Handlungsanweisungen irrelevant oder gar kontraproduktiv sind? Um dies zu verhindern,

muss der Praktiker für seine empirischen Untersuchungen mit dem Spektrum vorhandener Methoden vertraut sein. Er muss wissen, welche Annahmen den jeweiligen Methoden zugrunde liegen und ob deren Anwendung bei gegebener Information geeignete Aussagen zulassen. Er sollte durch den Einsatz vergleichbarer Methoden die Robustheit der Ergebnisse überprüfen.

Einen Überblick über selektiv ausgewählte ökonometrische Methoden zu liefern und anhand von Anwendungen deren Arbeitsweise aufzuzeigen, ist Anliegen dieses Beitrags. Behandelt werden methodische Probleme zu Standardfehlern und Treatment-Effekten. Zunächst geht es um heteroskedastie- und cluster-robuste Schätzungen. Es folgt die Erörterung von Problemen bei bernoulliverteilten Regressoren, Ausreißern und partiell identifizierten Parametern. Vorgeschlagene Ansätze zur Verbesserung der Standardfehler bei Vorliegen von Heteroskedastie unterscheiden sich in der Gewichtung der Residuen. Andere Verfahren nutzen die geschätzten Störgrößen aus, um künstlich eine größere Anzahl von Stichproben zu erzeugen, um auf deren Basis eine bessere Schätzung der Standardfehler zu erhalten oder machen sich vorhandene Nichtlinearitäten zunutze. Clusterrobuste Schätzungen zielen darauf ab, das Moulton-Problem zu lösen. Zu geringe Standardfehler bei Vorliegen von in Clustern zusammengefassten ähnlichen Beobachtungen werden korrigiert. Dies gelingt in den vorgeschlagenen Ansätzen nur unvollständig. Ein bisher nicht erörtertes Phänomen, dass Dummy-Variablen als Regressoren zu höheren Standardfehlern führen, je mehr ihr Mittelwert von 0.5 entfernt ist, mahnt zur Vorsicht beim Vergleich hinsichtlich der Präzision des Einflusses verschiedener [0; 1]-Regressoren. Ausreißer, d. h. ungewöhnliche Beobachtungen, die vor allem auf systematische Messfehler oder ungewöhnliche Ereignisse zurückzuführen sind, können erhebliche Auswirkungen auf die Schätzergebnisse haben. Die vorgeschlagenen Ansätze zur Aufdeckung von Ausreißern variieren hinsichtlich des Messkonzeptes und liefern nicht zwangsläufig Hinweise darauf, ob diese bei der empirischen Analyse zu berücksichtigen sind. Neuere Ansätze für nur partiell identifizierte Parameter können hier hilfreich sein. Erhöhen sie doch den Präzisionsgrad bei Unsicherheit, ob Ausreißer zu entfernen sind oder nicht.

Bei den Verfahren zur Bestimmung von Treatment-Effekten stehen vier Prinzipien im Fokus: Differenz-von-Differenzen-Schätzer, Matching-Verfahren, Analyse von Treatment-Effekte bei Quantilsregressionen und Regression-Discontinuity-Ansätze. Bei den Differenz-von-Differenzen-Schätzern ist zu unterscheiden, ob zusätzliche Kontrollvariablen zu berücksichtigen sind oder nicht. Das Spektrum der in neuerer Zeit sehr beliebten Matching-Verfahren, die darauf abzielen Untersuchungsgruppe und Kontrollgruppe zu homogenisieren, um statistische Zwillinge herauszufiltern, ist einerseits recht umfangreich geworden und weist andererseits methodisch bedeutsame Unterschiede auf. Noch vergleichsweise selten ist bisher der Einsatz von Quantilsregressionen zur Erfassung heterogener Kausaleffekte. Methodisch zu unterscheiden ist dabei, ob die Treatmentvariable als exogen oder endogen aufgefasst wird und ob weitere Kontrollvariablen Berücksichtigung finden oder nicht. Bei den Regression-Discontinuity-Ansätzen ist zu unterscheiden, ob die Zuordnung zur Treatment- oder Kontrollgruppe allein auf Basis einer beobachteten kontinuierlichen Variablen erfolgt oder auch nicht beobachtete Variablen herangezogen werden.

Die zunächst rein auf die Methodik abgestellte Diskussion der verschiedenen Verfahren wird im zweiten Teil dieses Beitrags um Anwendungen auf Cobb-Douglas-Produktionsfunktionen unter Verwendung von IAB-Betriebspaneldaten ergänzt. Verschiedene heteroskedastie-konsistente Schätzverfahren führen zu ähnlichen Resultaten für die Standardfehler. Cluster-robuste Schätzungen weisen deutlichere Abweichungen auf. Dummy-Variable als Regressoren mit einem Mittelwert in der Nähe von 0.5 führen zu kleineren Varianzen der Koeffizientenschätzer als Dummies mit niedrigeren oder höheren Mittelwerten. Nicht alle Ausreißer haben einen starken Einfluss auf die Signifikanz. Neuere Methoden zur Behandlung des Problems nur partiell identifizierter Parameter führen zu effizienteren Schätzungen als traditionelle Verfahren.

Die vier diskutierten Treatment-Effekt-Verfahren werden angewandt auf die Frage, ob betriebliche Bündnisse einen signifikanten Effekt auf den Produktionsoutput haben. Im Gegensatz zu unbedingten Differenz-von-Differenzen-Schätzern und Schätzern ohne Matching ergeben sich bei bedingten Differenz-von-Differenzen-Schätzern oder Matching-Schätzern auf Basis der Mahalanobis-Metrik positive, aber nur insignifikante Effekte. Das letztere Ergebnis muss im Rahmen der Quantils-Treatmenteffekt-Analyse spezifiziert werden. Je höher das betrachtete Quantil ist, umso eher besteht eine Tendenz zu positiv signifikanten Effekten. Eine einfache Regression-Discontinuity-Analyse zeigt einen Strukturbruch bei einer Wahrscheinlichkeit von 0.5, dass ein Betrieb ein betriebliches Bündnis vereinbart hat. Keine speziellen Effekte lassen sich während der großen Rezession 2008/09 ausmachen. Fuzzy Regression-Discontinuity-Schätzungen offenbaren, dass der Outputeffekt betrieblicher Bündnisse in Ostdeutschland signifikant niedriger liegt als in Westdeutschland. Eine kombinierte Anwendung der vier Grundprinzipien zur Ermittlung von Kausaleffekten führt zu interessanten neuen Erkenntnissen. So werden unter anderem Differenz-von-Differenzen Schätzer mit Matching-Verfahren verknüpft. Erstere werden auch in Verbindung mit Regressions-Discontinuity erörtert und letztere in Verbindung mit Quantilsregressionen.

Acknowledgements I wish to thank an anonymous reviewer for his constructive suggestions and the participants of the Nutzerkonferenz in Nürnberg for helpful comments.

Appendix

Table 11 OLS estimates of Cobb-Douglas functions with artificial dummies (DV.) as regressor; dependent variable: logarithm of sales

	$\hat{\beta}_{\ln L}$	Std.err.	$\hat{\beta}_{\ln K}$	Std.err.	$\hat{\beta}_{DV.}$	Std.err.
$\overline{DV1} = 0.1692$	0.9464	0.0043	0.2223	0.0032	0.0470	0.0128
$\overline{DV2} = 0.2952$	0.9453	0.0043	0.2223	0.0032	0.0808	0.0105
$\overline{DV3} = 0.3672$	0.9446	0.0043	0.2224	0.0032	0.0923	0.0099
$\overline{DV4} = \mathbf{0.5388}$	0.9434	0.0043	0.2225	0.0032	0.1334	**0.0096**
$\overline{DV5} = 0.6301$	0.9432	0.0043	0.2226	0.0032	0.1285	0.0100
$\overline{DV6} = 0.7190$	0.9438	0.0043	0.2226	0.0032	0.1124	0.0107
$\overline{DV7} = 0.8360$	0.9449	0.0043	0.2226	0.0032	0.0979	0.0130
$\overline{DV8} = 0.9445$	0.9448	0.0043	0.2226	0.0032	0.1599	0.0210
$\overline{DV9} = 1.0000$	0.9472	0.0043	0.2225	0.0032	0.0000	–

Table 12 OLS estimates of Cobb-Douglas functions with an artificial dummy (D.) determined from a rectangular distributed random variable as regressor. Results are average values of 300 estimates; dependent variable: logarithm of sales

	$\hat{\beta}_{D.}$	Std.err.
$\overline{D1} = 0.1$	−0.0177	0.0182
$\overline{D2} = 0.2$	−0.0040	0.0135
$\overline{D3} = 0.3$	−0.0065	0.0118
$\overline{D4} = 0.4$	−0.0148	0.0110
$\overline{D5} = \mathbf{0.5}$	−0.0105	**0.0108**
$\overline{D6} = 0.6$	−0.0111	0.0110
$\overline{D7} = 0.7$	−0.0134	0.0118
$\overline{D8} = 0.8$	−0.0073	0.0135
$\overline{D9} = 0.9$	0.0086	0.0180

Note: IAB Establishment Panel 2006–2010; $n = 34{,}308$

Table 13 OLS estimates of Cobb-Douglas functions with company-level pact dummy (CLP) as regressor, decreasing shares of $n(\text{CLP} = 1)/n$; dependent variable: logarithm of sales

	$\hat{\beta}_{\text{CLP}}$	Std.err.	t
$\overline{\text{CLP}} = \mathbf{0.0693}$	0.1231	**0.0236**	5.22
$\overline{\text{CLP}} = 0.0624$	0.1209	0.0246	4.92
$\overline{\text{CLP}} = 0.0533$	0.1299	0.0259	5.02
$\overline{\text{CLP}} = 0.0477$	0.1131	0.0275	4.11
$\overline{\text{CLP}} = 0.0407$	0.1006	0.0295	3.41
$\overline{\text{CLP}} = 0.0336$	0.1005	0.0322	3.12
$\overline{\text{CLP}} = 0.0273$	0.1429	0.0356	4.01
$\overline{\text{CLP}} = 0.0207$	0.1446	0.0403	3.39
$\overline{\text{CLP}} = 0.0135$	0.1357	0.0486	2.79
$\overline{\text{CLP}} = 0.0067$	0.1887	0.0671	2.80

Note: IAB Establishment Panel 2006–2010; $n = 31{,}985$. In the first line the estimation with the original sample and $\overline{\text{CLP}} = 0.0693$ is presented. Next, only 90 % of the firms with CLP = 1, where $\overline{\text{CLP}} = 0.0624$, are considered. The random selection of the CLP firms is based on a rectangular distribution of the CLP firms. The determination of the following lines is analogous to that of the second line

Table 14 OLS estimates of Cobb-Douglas functions with works council dummy (WOCO) as regressor, decreasing shares of $n(\text{WOCO} = 1)/n$—randomly determined based on a rectangular distribution; dependent variable: logarithm of sales

	$\hat{\beta}_{\text{WOCO}}$	Std.err.	t
$\overline{\text{WOCO}} = \mathbf{0.3045}$	0.4076	**0.0136**	29.50
$\overline{\text{WOCO}} = 0.2747$	0.3573	0.0136	26.32
$\overline{\text{WOCO}} = 0.2440$	0.3140	0.0136	23.13
$\overline{\text{WOCO}} = 0.2132$	0.2784	0.0137	20.29
$\overline{\text{WOCO}} = 0.1829$	0.2418	0.0141	17.11
$\overline{\text{WOCO}} = 0.1523$	0.2102	0.0148	14.20
$\overline{\text{WOCO}} = 0.1221$	0.1904	0.0159	11.99
$\overline{\text{WOCO}} = 0.0920$	0.1842	0.0177	10.43
$\overline{\text{WOCO}} = 0.0605$	0.1888	0.0208	9.07
$\overline{\text{WOCO}} = 0.0305$	0.1730	0.0281	6.16

Note: IAB Establishment Panel 2006–2010; $n = 34{,}217$

Table 15 Different CDF estimates, t-values in parentheses; dependent variable: logarithm of sales—$\ln Y$

	OLS	Cluster-robust	Fixed effects
$\ln L$	0.9472	0.9472	0.4096
	(222.12)	(126.29)	(35.84)
$\ln K$	0.2225	0.2225	0.0195
	(70.11)	(43.04)	(9.72)
const	9.0811	9.0810	13.2449
	(333.20)	(215.20)	(302.82)
n	34,308	34,308	34,308
R^2	0.843	0.843	0.070

Note: IAB Establishment Panel 2006–2010

References

Abadie, A., Angrist, J., Imbens, G.: Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. Econometrica **70**, 91–117 (2002)

Ai, C., Norton, E.C.: Interaction terms in logit and probit models. Econ. Lett. **80**, 123–129 (2003)

Angrist, J., Pischke, J.-S.: Mostly Harmless Econometrics—an Empiricist's Companion. Princeton University Press, Princeton (2009)

Belsley, D.A., Kuh, E., Welsch, R.E.: Regression Diagnostics—Identifying Influential Data and Sources of Collinearity. Wiley, New York (1980)

Cameron, A.C., Miller, D.L.: Robust inference with clustered data. In: Ullah, A.C., Giles, D.E.A. (eds.) Handbook of Empirical Economics and Finance, pp. 1–28 (2010)

Chernozhukov, V., Hong, H., Tamer, E.: Estimation and confidence regions for parameter sets in econometric models. Econometrica **75**, 1243–1284 (2007)

Cribari-Neto, F., da Silva, W.D.: A new heteroskedasticity-consistent covariance matrix estimator for the linear regression model. AStA Adv. Stat. Anal. **95**, 129–146 (2011)

Cribari-Neto, F., Souza, T.C., Vasconcellos, K.L.P.: Inference under heteroskedasticity and leveraged data. Commun. Stat., Theory Methods **36**, 1877–1888 (2007)

Firpo, S.: Efficient semiparametric estimation of quantile treatment effects. Econometrica **75**, 259–276 (2007)

Frölich, M., Melly, B.: Unconditional Quantile Treatment Under Endogeneity. (2012) *mimeo*

Goldstein, H.: Multilevel Statistical Models, Kendall's Library of Statistics, 3rd edn. Arnold, London (2003)

Guo, S., Fraser, M.W.: Propensity Score Analysis. Sage Publications, Thousand Oaks (2010)

Hadi, A.S.: Identifying multiple outliers in multivariate data. J. R. Stat. Soc. B **54**, 761–771 (1992)

Hamermesh, D.S.: The craft of labormetrics. Ind. Labor Relat. Rev. **53**, 363–380 (2000)

Imbens, G.W., Manski, C.F.: Confidence intervals for partially identified parameters. Econometrica **72**, 1845–1857 (2004)

Koenker, R., Bassett, G.: Regression quantiles. Econometrica **46**, 33–50 (1978)

Krämer, W.: The cult of statistical significance—what economists should and should not do to make their data talk. J. Appl. Soc. Sci. Stud. **131**, 455–468 (2011)

Leamer, E.E.: Sensitivity analyses would help. Am. Econ. Rev. **75**, 308–313 (1985)

MacKinnon, J.G., White, H.: Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. J. Econom. **29**, 305–325 (1985)

Moulton, B.R.: Random group effects and the precision of regression estimates. J. Econom. **32**, 385–397 (1986)

Moulton, B.R.: Diagnostic tests for group effects in regression analysis. J. Bus. Econ. Stat. **6**, 275–282 (1987)

Moulton, B.R.: An illustration of a pitfall in estimating the effects of aggregate variables on micro units. Rev. Econ. Stat. **72**, 334–338 (1990)

Puhani, P.: The treatment, the cross difference, and the interaction term in nonlinear 'difference-in-differences' models. Econ. Lett. **115**, 85–87 (2012)

Raudenbush, A.S., Bryk, S.W.: Hierarchical Linear Models, 2nd edn. Sage Publications, Thousand Oaks (2002)

Romano, J.P., Shaikh, A.M.: Inference for the identified set in partially identified econometric models. Econometrica **78**, 169–211 (2010)

Rosenbaum, P.R., Rubin, D.P.: Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am. Stat. **39**, 33–38 (1985)

Stoye, J.: More on confidence intervals for partially identified parameters. Econometrica **77**, 1299–1315 (2009)

Wald, H.: The fitting of straight line if both variables are subject to error. Ann. Math. Stat. **11**, 284–300 (1940)

White, H.: A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica **48**, 817–838 (1980)

Wooldridge, J.M.: Cluster-sample methods in applied econometrics. Am. Econ. Rev. **93**(PaP), 133–138 (2003)

Woutersen, T.: A Simple Way to Calculate Confidence Intervals for Partially Identified Parameters. (2009) *mimeo*

Ziliak, S., McCloskey, D.: The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives. University of Michigan Press, Michigan (2008)

Olaf Hübler studied economics at the Free University of Berlin where he received his diploma in 1970. From 1970–1975 he was an assistant at the chair for public finance and in the team for undergraduate studies at the Technical University of Berlin. In 1974 he received his Ph.D. and in 1978 his post-doctoral degree (Habilitation) in economics and statistics. From 1979 to 1980 he was Senior Research Fellow at the International Institute of Management in Berlin. Since 1982 he was a Professor of Econometrics at the University of Hannover, 1999 Visiting Professor at the University ofStirling, Scotland. Since 1999 he is Research Fellow at IZA in Bonn and since 2005 Research Fellow at the IAB in Nuremberg. From 2003–2012 he was co-chair and chair of the German Statistical Association's section of Empirical Economics and Applied Econometrics and from 2006–2008 speaker of the DFG priority program "Flexibilisierungs potenziale bei heterogenen Arbeitsmärkten". Among others Olaf Hübler has published papers in numerous national and international refereed journals and is author of the textbooks "Ökonometrie" and "Einführung in die empirischeWirtschaftsforschung". He is member of the editorial board of Advances in Statistical Analysis since 2001. 2011 he was retired.