Tutz, Edlich, Bäumer:

Comparison between local estimates for multi-categorical varying-coefficent models

Sonderforschungsbereich 386, Paper 151 (1999)

Comparison between local estimates for multi-categorical varying-coefficent models

Gerhard Tutz
Institut für Statistik
LMU München

Silke Edlich
Institut für quantitative Methoden
Technische Universität Berlin

Christoph Bäumer
Gesellschaft für Innovationstechnik mbH
Berlin

Summary

Varying coefficient models with discrete values of the effect modifier may be estimated by maximum likelihood or weighted least square techniques. We compare bias reduction methods for both estimates as well as the performance of the estimates as compared to each other.

1 Introduction

In varying-coefficient models the coefficients may vary across the so-called effect modifier. Models of this type which are based on the rather wide class of generalized linear models have been considered by Hastie & Tibshirani (1993).

In the following the focus is on categorial regression models where the response is multinomially distributed. With response variable y taking values from $\{1, \ldots, k\}$ the model to be considered has the form

$$\pi(x, u) = h(Z(x)\beta(u)) \qquad (1)$$

where $\pi(x,u)' = (P(y = 1|x,u), \ldots, P(y = q|x,u)),\ q = k - 1$, is the vector of response probabilities, $Z(x)$ is a design matrix composed from covariates x and $\beta(u)$ is a parameter vector which may vary across the effect modifier u. That means the predictor $\eta = Z(x)\beta(u)$ is linear for fixed u but the dependence on x *and* u is nonlinear. The objective of this paper is to investigate small sample properties of two types of estimates and compare their performance. The first estimate is based on the local likelihood principle (Tibshirani & Hastie, 1987), the second estimate is a weighted least squares estimate which is less time consuming than the local likelihood estimator. For both estimates some form of bias correction seems advisable. In particular for the weighted least squares estimate alternative variants are investigated.

2 The estimates

Both estimates that are considered here are local estimates. The estimate of $\beta(u)$ is based on observations which are obtained in the neighbourhood of u, but the influence of observations is weighted down with increasing distance from u.
Let the observations be given in a grouped form by $(p_{ti}, x_{ti}, u_t),\ t = 1, \ldots, T,\ i = 1, \ldots, m_t$, where $p'_{ti} = (p_{ti1}, \ldots, p_{tiq})$ is the vector of relative frequencies observed at (x_{ti}, u_t). Here, x_{ti} is a discrete vector of covariates and u_t is the effect modifier which may be discrete or continuous. Let n_{ti} denote the number of observations at fixed (x_{ti}, u_t) and $\sum_{i=1}^{m_t} n_{ti}$ denote the number of observations taken at the value u_t of the effect modifier. The model is given by

$$\pi_{ti} = h(Z_{ti}\beta_t) \quad \text{or} \quad g(\pi_{ti}) = Z_{ti}\beta_t, \tag{2}$$

$t = 1, \ldots, T,\ i = 1, \ldots, m_t$ with the response vector $\pi'_{ti} = (P(y = 1|x_{ti}, u_t), \ldots, P(y = q|x_{ti}, u_t))$, design matrices $Z_{ti} = Z(x_{ti})$ and parameter vector $\beta_t = \beta(u_t)$. The link function $g = (g_1, \ldots, g_q)$ is the inverse of the response function $h = (h_1, \ldots, h_q)$, i.e. $g = h^{-1}$.
For the estimation of β_t the weights which determine the influence of observations y_{si} are given by

$$w_\gamma(u_t, u_s) = K\left(\frac{u_t - u_s}{\gamma}\right) / K(0),$$

where K is a unimodal, symmetric kernel function and $\gamma > 0$ is a smoothing parameter. Observations at u_t will obtain the weight $w_\gamma(u_t, u_t) = 1$ whereas observations at $u_t \neq u_s$ will get weight $w_\gamma(u_t, u_s) < 1$.

2.1 Local likelihood estimator

The basic idea is to maximize the local likelihood instead of the full likelihood. For the estimation of $\beta_t = \beta(u_t)$ the kernel of the local likelihood is given by

$$\begin{aligned} l_\gamma(\beta_t) = & \sum_{s=1}^{T} w_\gamma(u_t, u_s) \sum_{i=1}^{m_s} n_{si} \sum_{r=1}^{q} \{p_{sir} \log(h_r(Z_{si}\beta_t)) \\ & + (1 - \sum_{j=1}^{q} p_{sij}) \log(1 - \sum_{j=1}^{q} h_j(Z_{si}\beta_t))\}. \end{aligned} \tag{3}$$

That means β_t is considered to be the underlying parameter and all observations are used, but observations which are close to u_t are emphasized. Maximization of (3) implies to solve the local score equation

$$s_\gamma(\beta_t) = \sum_{s=1}^{T} w_\gamma(u_t, u_s) \sum_{i=1}^{m_s} Z'_{si} D_{si}(\beta) \Sigma_{si}^{-1}(\beta)(p_{si} - h(Z_{si}\beta_t)) = 0.$$

where $D_{si}(\beta_t) = \partial h(Z_{si}\beta_t)/\partial\eta$ stands for the derivative and $\Sigma_{si}(\beta_t) = \{\text{Diag}(\pi_{si}) - \pi_{si}\pi'_{si}\}/n_{si}$ is the covariance matrix of p_{si} if $\pi_{si} = h(Z_{si}\beta_t)$ is the underlying probability.

Under weak conditions the local likelihood estimate is consistent and asymptotically normally distributed

$$\hat{\beta}_t - \beta_t \sim N(0, V_t)$$

with $V_t = \sum_{s=1}^{T} w_\gamma(u_t, u_s) F_s(\beta_t)$ where F_s is the local Fisher matrix

$$F_s(\beta_t) = \sum_{i=1}^{m_s} Z'_{si} D_{si}(\beta_t) \Sigma_{si}^{-1}(\beta_t) D_{si}(\beta_t)' Z_{si}.$$

The local likelihood maximum (lml) estimate with bias correction which is to be preferred is given by

$$\hat{\beta}_t^c = \hat{\beta}_t + V_t^{-1} s_\gamma(\hat{\beta}_t).$$

For a derivation of the bias correction and the asymptotic behaviour see Tutz & Kauermann (1995).

2.2 Locally weighted least squares estimator

The common least squares estimate for β_t (e.g. Grizzle, Starmer & Koch, 1969) is based on minimization of the criterion

$$\sum_{i=1}^{m_t} (g(p_{ti}) - Z_{ti}\beta_t)' C_{ti}(p_{ti})^{-1} (g(p_{ti}) - Z_{ti}\beta_t) \tag{4}$$

with $C_{ti}(\tilde{p}) = [\partial g(\pi_{ti})/\partial\pi'] (\text{Diag}\,\pi_{ti} - \pi_{ti}\pi'_{ti}) [\partial g(\pi_{ti})/\partial\pi] / n_{ti}$ representing an approximation of $\text{cov}(g(p_{ti}))$. In (4) only observations at fixed measurement points u_t are used. The locally weighted least squares estimate minimizes

$$\sum_{s=1}^{T} w_\gamma(u_t, u_s) \sum_{i=1}^{m_s} (g(p_{si}) - Z_{si}\beta_t)' C_{si}(\pi_{si})^{-1} (g(p_{si}) - Z_{si}\beta_t)$$

where all observations are used but with varying weights. The estimate is given by

$$\hat{\beta}_t^{LS} = \left\{ \sum_{s=1}^{T} \sum_{i=1}^{m_s} w_\gamma(u_t, u_s) Z'_{si} C_{si}(p_{si})^{-1} Z_{si} \right\}^{-1} \sum_{s=1}^{T} \sum_{i=1}^{m_s} w_\gamma(u_t, u_s) Z'_{si} C_{si}(p_{si})^{-1} g(p_{si}) \tag{5}$$

where $C_{si}(\pi_{si})$ is replaced by the empirical covariance matrix $C_{si}(p_{si})$.
The asymptotic behaviour of $\hat{\beta}_t^{LS}$ is the same as for the local likelihood estimator β_t^{ML} (see Tutz & Kauermann, 1996). However, asymptotic properties suggest differing bias correction schemes. The first correction scheme may be seen directly from (5). By the approximation $Eg(p_{si}) \approx Z_{si}\beta_s$ one obtains

$$E\hat{\beta}_t^{LS} \approx \left\{ \sum_{s=1}^{T} \sum_{i=1}^{m_s} w(u_t, u_s) Z'_{si} C_{si}(p_{si})^{-1} Z_{si} \right\}^{-1} \sum_{s=1}^{T} \sum_{i=1}^{m_s} w(u_t, u_s) Z'_{si} C_{si}(p_{si})^{-1} Z_{si} \beta_s.$$

If β_s is replaced by the estimate $\hat{\beta}_s$, the corresponding bias is given by

$$\begin{aligned} \hat{b}_{t,1} &= \left\{ \sum_{s=1}^{T} \sum_{i=1}^{m_s} w(u_t, u_s) Z'_{si} C_{si}(p_{si})^{-1} Z_{si} \right\}^{-1} \sum_{s=1}^{T} \sum_{i=1}^{m_s} w(u_t, u_s) Z'_{si} C_{si}(p_{si})^{-1} Z_{si} (\hat{\beta}_s - \hat{\beta}_t) \\ &= \left\{ \sum_{s=1}^{T} \sum_{i=1}^{m_s} w(u_t, u_s) Z'_{si} C_{si}(p_{si})^{-1} Z_{si} \right\}^{-1} \sum_{s=1}^{T} \sum_{i=1}^{m_s} w(u_t, u_s) Z'_{si} C_{si}(p_{si})^{-1} (Z_{si}\hat{\beta}_s - g(p_{si})). \end{aligned} \tag{6}$$

Thus one obtains the bias corrected estimate

$$\hat{\beta}_{t,1}^c = \hat{\beta}_t - \hat{b}_{t,1}.$$

The second method of bias reduction is based on the asymptotic expansion of $\hat{\beta}_t^{LS}$ (see Tutz & Kauermann, 1996). It reduces the bias by the order $O(n^{-\alpha})$ by using the approximation

$$\hat{b}_{t,2} = A_t^{-1} \bar{B}_t - A_t^{-1} \bar{A}_t A_t^{-1} B_t$$

with

$$A_t = \frac{n_{ti}}{n_t} \sum_i w(u_t, u_t) Z'_{ti} \bar{C}_{ti}^{-1} Z_{ti}, \qquad \bar{A}_t = \sum_{s \neq t} w(u_t, u_s) \frac{n_s}{n_t} \sum_i \frac{n_{si}}{n_s} Z'_{si} \bar{C}_{si}^{-1} Z_{si},$$

$$B_t = \frac{n_{ti}}{n_t} \sum_i w(u_t, u_t) Z'_{ti} \bar{C}_{ti}^{-1} g(p_{ti}), \qquad \bar{B}_t = \sum_{s \neq t} w(u_t, u_s) \frac{n_s}{n_t} \sum_i \frac{n_{si}}{n_s} Z'_{si} \bar{C}_{si}^{-1} g(p_{si}).$$

Using $\hat{\beta}_{t,2}$ directly yields very high variance of the estimate. Thus we consider a limited influence bias correction by using a tuning constant f_δ which depends on a threshold $\delta > 0$. The corresponding bias corrected estimate is given by

$$\hat{\beta}_{t,2}^c = \hat{\beta}_t - f_\delta \hat{b}_{t,2}.$$

where

$$f_\delta = \begin{cases} 1 & \text{if} \quad ||\hat{\beta}_{t,2}||/||\hat{\beta}_t|| < \delta \\ \delta ||\hat{\beta}_t||/||\hat{\beta}_{t,2}|| & \text{otherwise.} \end{cases}$$

That means if the correction is comparatively small, it is actually used, but if the correction is above threshold δ a downweighted version is used. One obtains for $\delta \to \infty$ $f_\delta = 1$. The disadvantage of this procedure is that an additional threshold δ which determines the tuning constant f_δ has to be chosen.

3 Comparison between estimators

For the evaluation of the estimators the criterion in the following is the expected squared error loss. Focusing on the probability one considers

$$L_\pi = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{m_t} \sum_{i=1}^{m_t} \sum_{j=1}^{k} (\pi_{tij} - \hat{\pi}_{tij})^2,$$

where $\pi'_{ti} = (\pi_{ti1}, \ldots, \pi_{tik})$ and $\hat{\pi}'_{ti} = (\hat{\pi}_{ti1}, \ldots, \hat{\pi}_{tik})$ are the true and the estimated response probabilities for covariates (x_{ti}, u_t), respectively. Estimators will be compared by the expectation EL. For the dichotomous case EL is the mean squared error (multiplied with the factor 2).

When the focus is on the varying coefficients one considers the integrated mean squared error (IMSE)

$$L_\beta = \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{p} (\beta_{ti} - \hat{\beta}_{ti})^2.$$

Estimators to be compared are

- local likelihood (LL)
- local likelihood with bias correction (LLB)
- locally weighted least squares (LS)
- locally weighted least squares with bias correction $b_{t,1}, b_{t,2}$ (LSB1, LSB2)

In simulation study A the underlying model is the dichotomous logit model

$$\pi_{ti} = \frac{\exp(\beta_{t0} + x_i\beta_{t1})}{1 + \exp(\beta_{t0} + x_i\beta_{t1})}, \quad t = 1, \dots, 21, \quad i = 1, 2$$

with dichotomous $x \in \{0, 1\}$ and varying coefficients

$$\beta_{t0} = 0.75 \sin(2\pi(t-1)/20)$$
$$\beta_{t1} = -0.75 + (0.075 + \exp(-(t-10)^2/10)).$$

For the 100 simulated data sets at each point (t, i), $t = 1, \dots, 21$, $i = 1, 2$, $n_{ti} = N$ observations are drawn.
Since in application, the smoothing parameter is not known it has to be chosen data adaptively. The cross validation criterion to be minimized is

$$\text{CV}(\gamma) = \sum_{t=1}^{T} \sum_{i=1}^{m_t} \frac{n_{ti}}{n} \sum_{j=1}^{k} (\pi_{tij} - \hat{\pi}_{tij}^{-(ti)})^2$$

where $\pi_{tij}^{-(ti)}$ stands for the estimate where the observation (p_{ti}, x_{ti}, u_t) has been omitted.
In simulation study B the same model is used but with

$$\begin{aligned} \beta_{t0} &= 0.4((t-11)/3)^2 \\ \beta_{t1} &= -0.75 + (0.075t + \exp(-(t-10)^2/2)). \end{aligned}$$

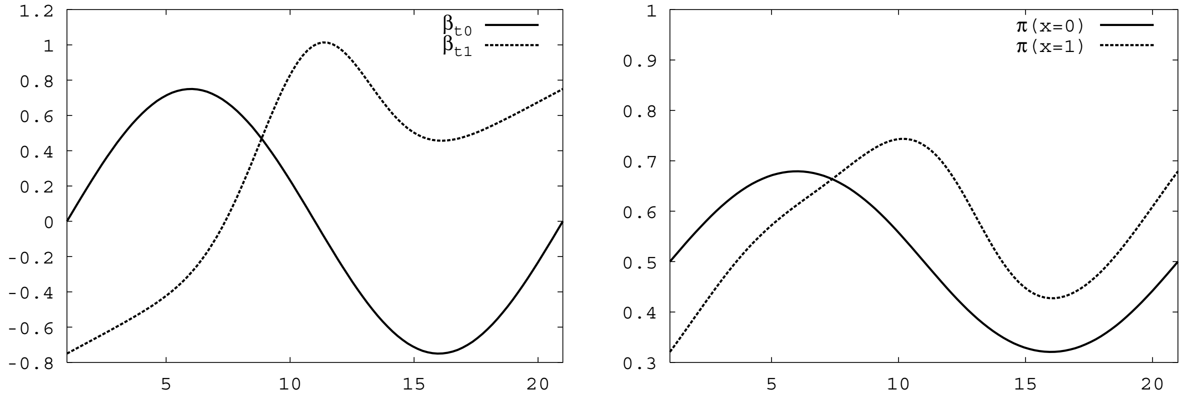


Figure 1: Variation of β_{t0}, β_{t1} (left) and the corresponding probabilities (right) for simulation study A

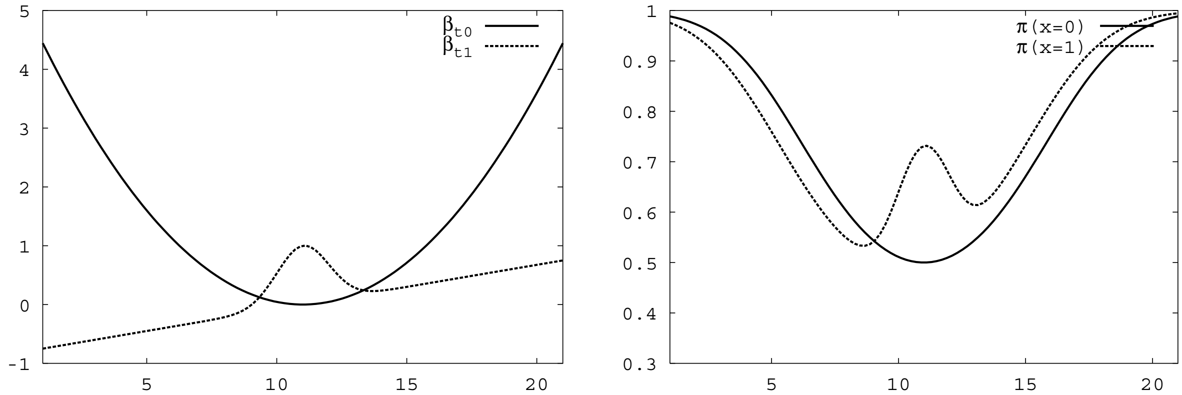


Figure 2: Variation of β_{t0}, β_{t1} (left) and the corresponding probabilities (right) for simulation study B

Figure 1 and 2 show the variation of β_{t0}, β_{t1} and the corresponding probability for simulation study A and B. The essential difference between the simulation studies is that in study B the probabilities are chosen to be above 0.5 for both populations whereas in study A the probabilities vary around 0.5.

ML estimate and bias correction

Figure 3 and 4 show the mean squared error (averaged over all t) for local sample sizes $n_{ti} = 4$ and $n_{ti} = 10$ as well as the bias for various values of the smoothing parameter. For $\gamma \to \infty$ all the observations are used for estimating β_t with the consequence that bias and IMSE are quite high. For $\gamma \to 0$ the bias is rather low since neighbourhoods are small. However, with smaller values of γ the variance increases yielding an increase in IMSE.

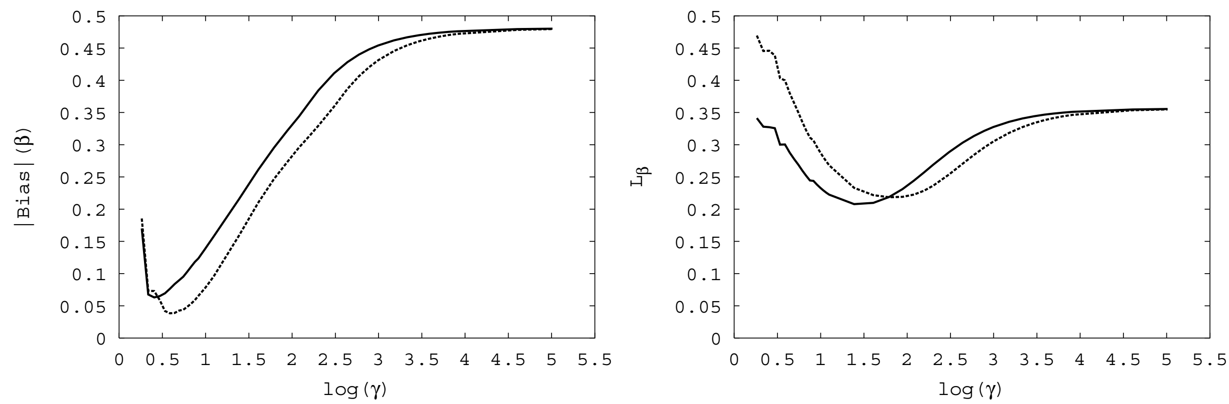


Figure 3: Bias (left) and mean squared error (right) for the estimation of β_t with local sample size $n_{ti} = 4$, study A (solid line is the ML estimate and dashed line the bias corrected estimate)

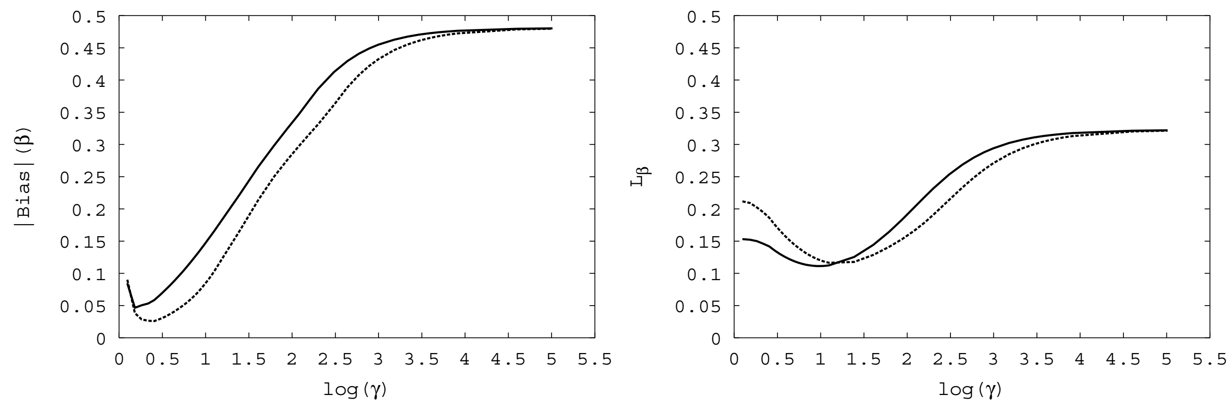


Figure 4: Bias (left) and mean squared error (right) for the estimation of β_t with local sample size $n_{ti} = 10$, study A (solid line is the ML estimate and dashed line the bias corrected estimate)

The proposed bias reduction of ML estimates seems to work well for all smoothing parameters. However, it has the side effect that the variance of the estimates is increased yielding a mean squared error that is not always superior. It is seen that the bias corrected estimates yield better performance in terms of the mean squared error for larger smoothing parameters. However, if smoothing is very low the variance is strongly increased with the effect that the mean squared error is higher for the estimate with bias correction. The turning point is quite close to the optimal smoothing parameter. Since the construction of confidence bands is based on nearly unbiased estimates the bias reduction is often to be considered as more important than the loss of variance.
It is seen that additive bias reduction works quite well. The bias is strongly reduced. The advantages of additive bias reduction as compared to polynomial fitting are investigated in citeNKau-etal:98. The loss in variance is low if instead of the optimal smoothing parameter slight oversmoothing is applied.

WLS estimate and bias correction

Bias correction for the WLS estimate shows the same effects as correction for the ML estimates. The bias is decreased but variability is increased. Figure 5 and 6 show the bias and mean squared error for $n_{ti} = 4$ and $n_{ti} = 10$ of study A. The corresponding pictures of study B are quite similar. It is seen that for the second method of bias correction the turning constant is very important. For large values of δ the bias is strongly reduced but IMSE gets quite large for low smoothing. With sensible choice of the tuning constant, e .g . $\delta = 0.125$ in this case, the performance in particular for low smoothing is distinctly superior to the first method of bias correction.

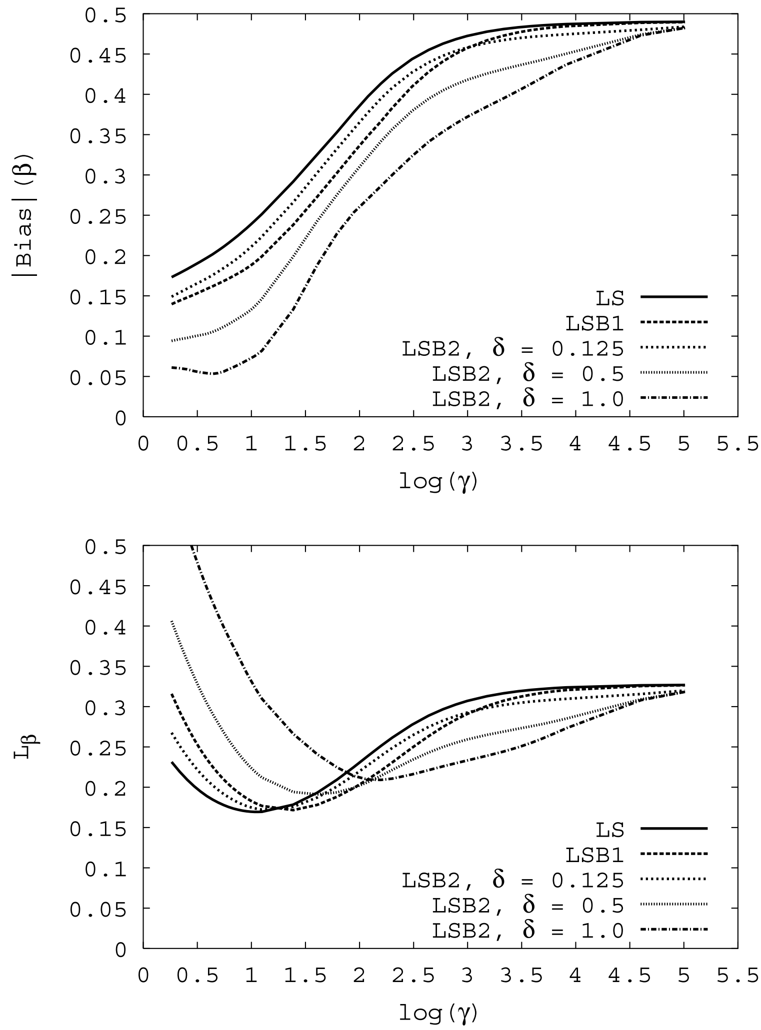


Figure 5: Bias (above) and mean squared error (below) for estimation of β_t with $n_{ti} = 4$, study A

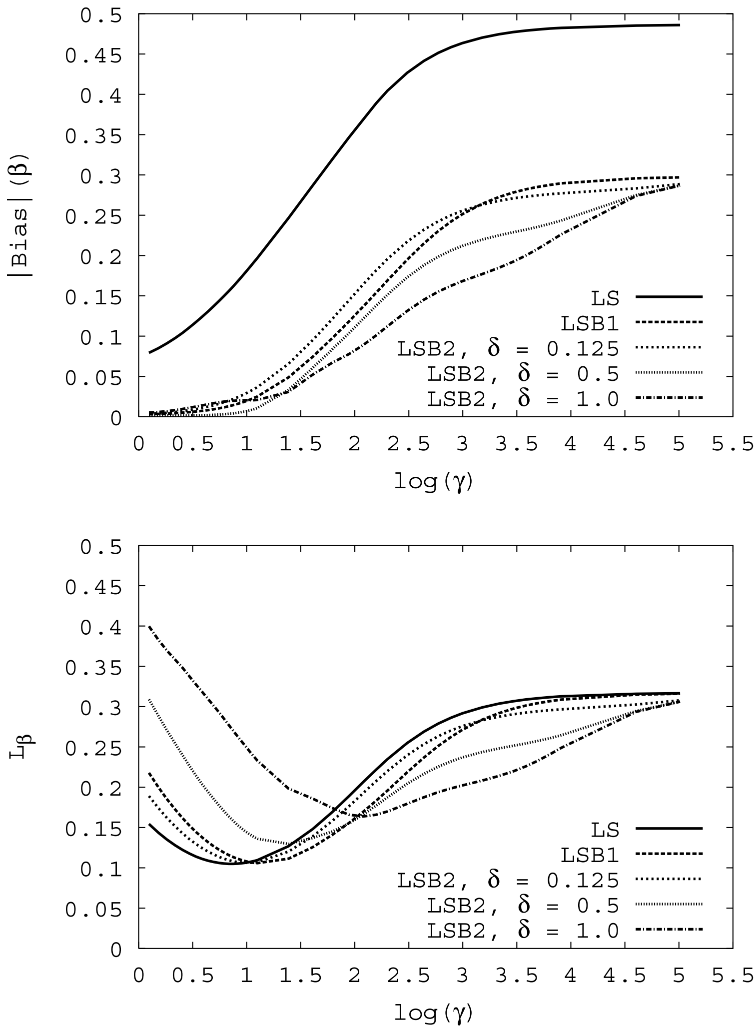


Figure 6: Bias (above) and mean squared error (below) for estimation of β_t with $n_{ti} = 10$, study A

Comparison between ML and WLS

ML and WLS estimates have differing individual smoothing parameters. Thus for comparison their performance is measured at their individual optimal smoothing parameters. Tables 1 to 4 give the corresponding IMSEs for the estimation of β_t andπ_{ti}, where LSB2 uses tuning constant $\delta = 0.125$.

n_{ti}	LL	LLB	LS	LSB1	LSB2
$n_{ti} = 2$	0.0343	0.0360	0.0240	0.0243	0.0242
$n_{ti} = 4$	0.0207	0.0211	0.0174	0.0172	0.0177
$n_{ti} = 10$	0.0108	0.0104	0.0105	0.0100	0.0103

Table 1: Mean squared error for the estimates of the underlying probability, study A

n_{ti}	LL	LLB	LS	LSB1	LSB2
$n_{ti} = 2$	0.3500	0.34936	0.2348	0.2391	0.2392
$n_{ti} = 4$	0.2079	0.21848	0.1694	0.1716	0.1723
$n_{ti} = 10$	0.1113	0.11640	0.1050	0.1058	0.1062

Table 2: Mean squared error for the estimates of the parameter β, study A

n_{ti}	LL	LLB	LS	LSB1	LSB2
$n_{ti} = 2$	0.0255	0.0237	0.0423	0.0403	0.0391
$n_{ti} = 4$	0.0174	0.0167	0.0257	0.0238	0.0236
$n_{ti} = 10$	0.0050	0.0042	0.0119	0.0107	0.0108

Table 3: Mean squared error for the estimates of the underlying probability, study B

n_{ti}	LL	LLB	LS	LSB1	LSB2
$n_{ti} = 2$	0.8441	0.8447	1.2680	1.2197	1.2170
$n_{ti} = 4$	0.7298	0.6394	0.9308	0.8793	0.8812
$n_{ti} = 10$	0.3624	0.3070	0.5605	0.5174	0.5188

Table 4: Mean squared error for the estimates of the parameter β, study B

Table 1 and 2 shows that the performance of the weighted least square estimates is distinctly superior to the performance of the ML estimations variants. This holds especially for low sample sizes. This is surprising since the ML estimate is an iteratively reweighted WLS estimator which for small sizes often is expected to yield better performance. The WLS estimator actually is smoothed in two ways: smoothing by neighbourhood information and smoothing in the case of relative frequencies one or zero. Since for $p_{si} \in \{0,1\}$ the inverse logistic function $g(p_{si})$ does not exist, relative frequencies in these cases are computed by $\tilde{p}_{sir} = (n_{sir}+1)/(n_{si}+k)$. The effect is shrinkage of p_{si} towards 0.5. Estimators which shrink towards 0.5, e.g. kernel estimators and Bayes estimators, often have advantages over unshrinked estimates. In particular if the underlying probabilities vary around 0.5 shrinkage seems to be helpful. Although there is bias at some peaks which are away from 0.5, the variation of the estimates is reduced resulting in much lower variance of the estimator. This is demonstrated in Figure 7 where the variances for ml and weighted least squares estimates are compared for the extreme case $n_{ti} = 2$.

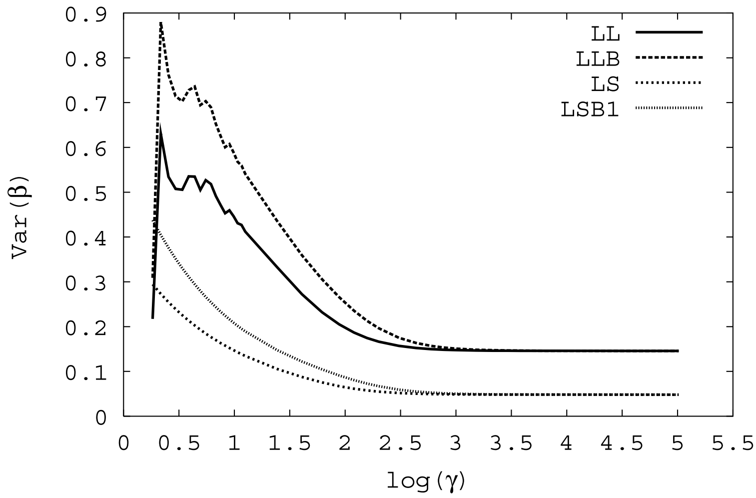
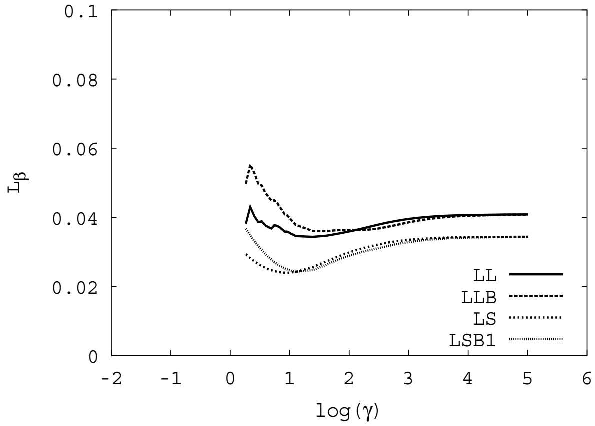


Figure 7: Variance of ML and WLS estimates with $n_{ti} = 2$, study A

In most data sets we considered, there was some variation of probabilities around 0.5. Nevertheless this will not be the case in all data sets. Thus simulation study B has been deliberately constructed in a way that all probabilities are above 0.5 with the effect that shrinkage always means stronger bias. From Table 3 and 4 it is seen that then the performance of the ML estimator is superior to that of the WLS estimator. A comparison of the mean squared error functions where bias and variance come together is given in Figure 8. It is obvious from Figure 8 that for study B the ML estimator has lower MSE than the WLS estimators. However, the ML estimator is very unstable in the neighbourhood of the minimum, meaning that often it does not exist. In study A where the probabilities are varying around 0.5 the WLS estimator clearly outperforms the ML estimate.

A further issue is stability of the estimators for data-based smoothing parameters.



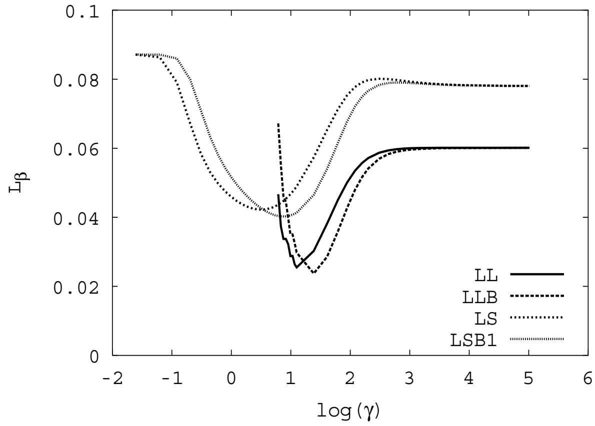


Figure 8: Integrated mean squared error, study A (left) and study B (right), $n_{ti} = 2$

WLS estimators with correction for the relative frequencies zero or one always exist whereas ML estimators often do not exist if neighbourhood smoothing is low. Figure 9 shows the number of successful estimators from 100 simulations. It is seen that the ML estimator is very unstable if the smoothing parameter is close to $\log(\gamma) = 1$. This is noteworthy since $\log(\gamma) \approx 1$ indicates the range of smoothing parameters where the loss is minimal (see Figure 3). Thus ML estimates are unstable close to the optimal smoothing parameter.
In Figure 10 the mean squared error is given after selection of the smoothing parameter by cross validation. The x-axis gives the cross-validation criteria, the y-axis shows the resulting IMSE. It is seen that the ML estimate has stronger variation in terms of IMSE. This is also seen from Figure 12 which gives the distribution of the IMSE estimated from the simulation results by a kernel density estimate. For study B where the ML estimate dominates the effect is turned around. Now the ML estimate show weaker variation than the WLS estimators. Thus the estimator that performs best has also lower variation.

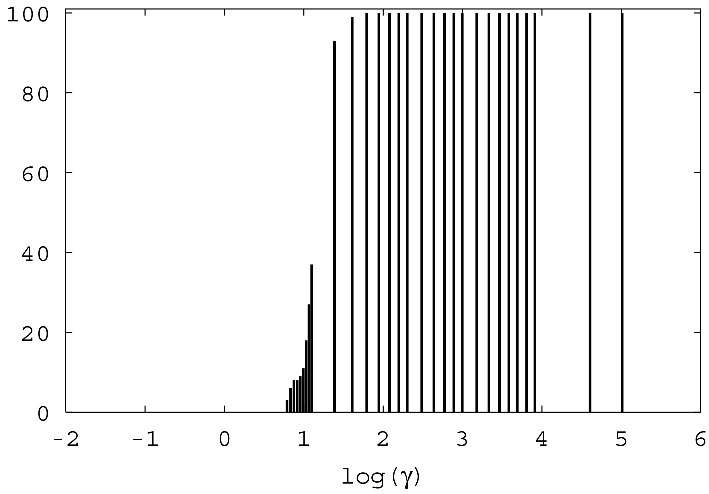


Figure 9: Number of successful ML estimators from 100 simulations, $n_{ti} = 2$, study B

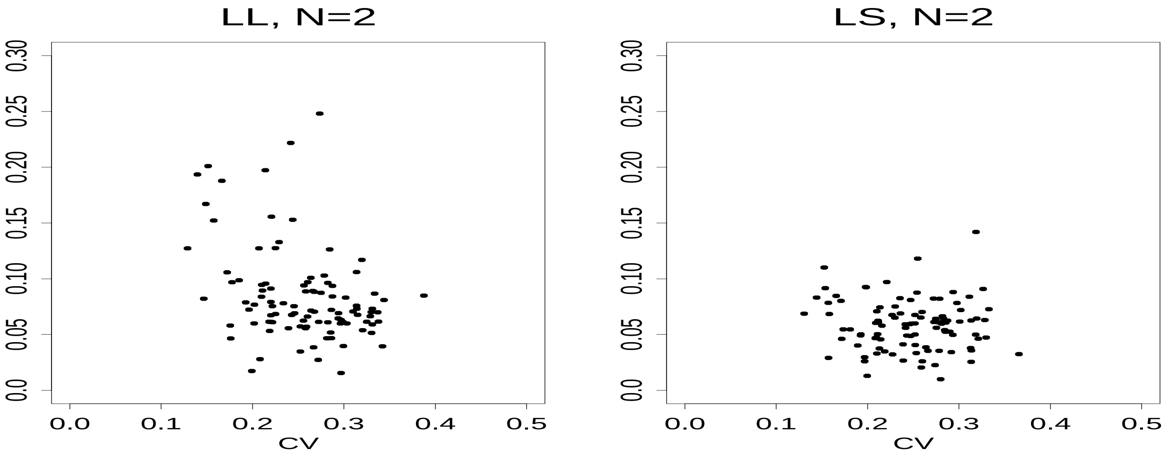


Figure 10: Actual Integrated mean squared error against smoothing parameter resulting from cross validation for ML estimate (left) and WLS estimate (right), study A, $n_{ti} = 2$

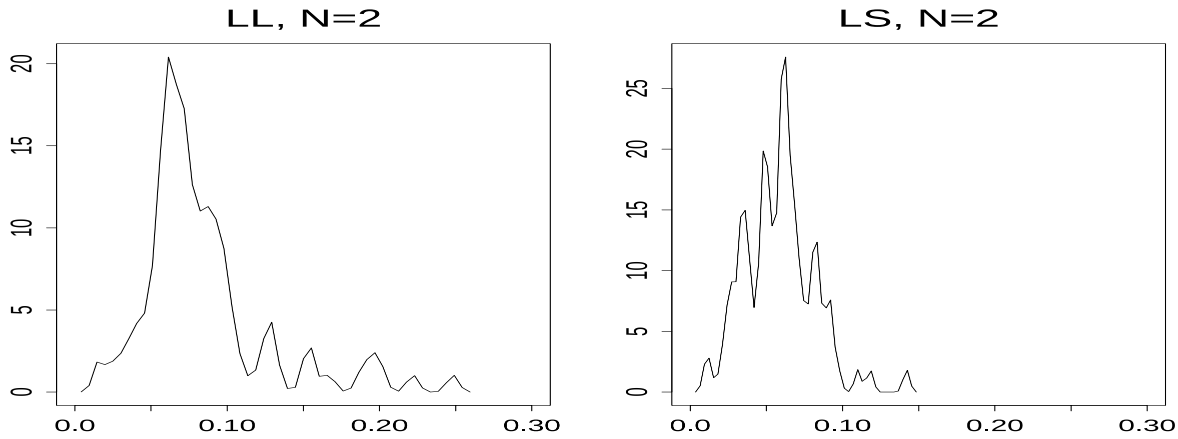


Figure 11: Integrated mean squarred error after cross validation for ML estimate (left) and WLS estimate(right), study A, $n_{ti} = 2$

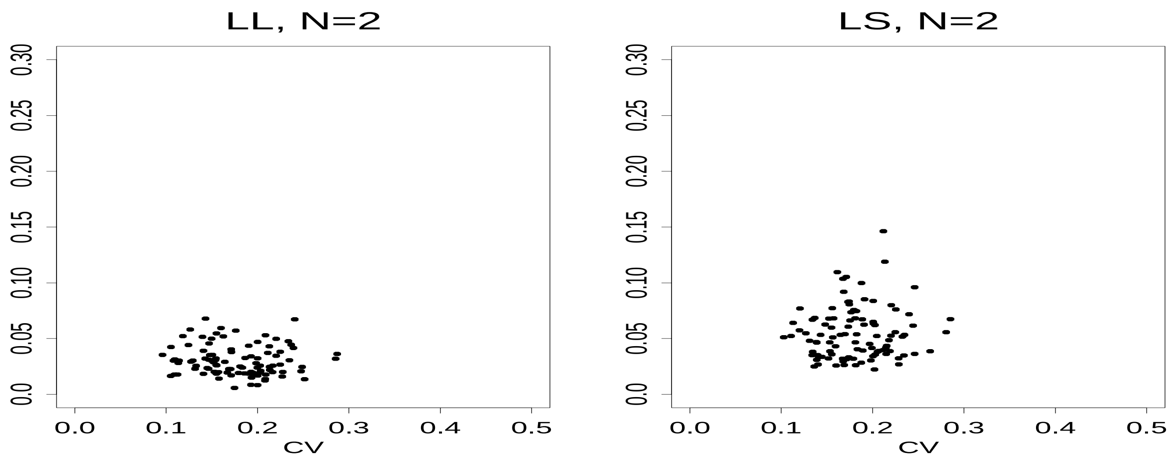


Figure 12: Actual mean squared error against smoothing parameter resulting from cross validation for ML estimate (left) and WLS estimate (right), study B, $n_{ti} = 2$

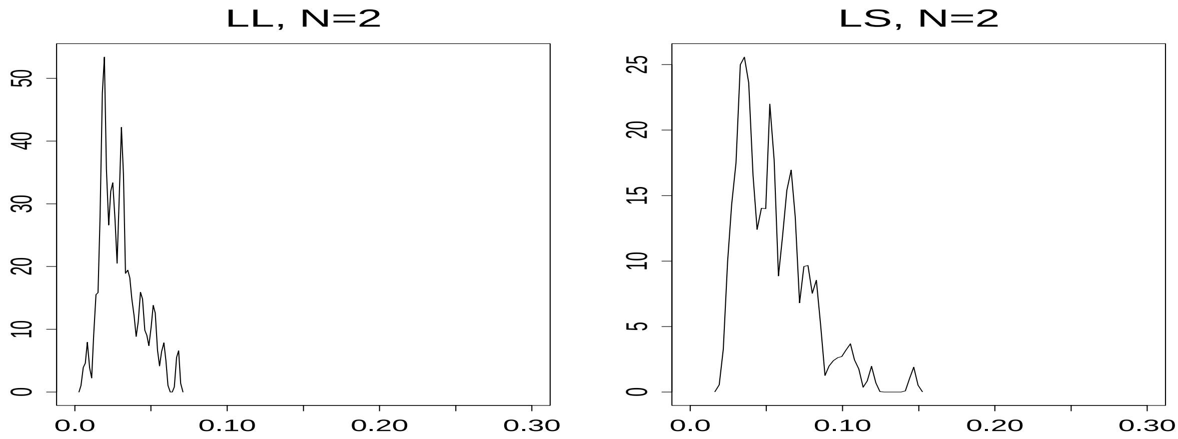


Figure 13: Integrated mean squarred error after cross validation for ML estimate (left) and WLS estimate(right), study B, $n_{ti} = 2$

4 Concluding remarks

Additive bias correction is a strong device to reduce the bias for ML estimates as well as WLS estimates. With respect to the quadratic loss undersmoothing has to be avoided whereas oversmoothing in combination with bias reduction yields good results.

Comparison of ML and WLS estimates has several aspects. A disadvantage of ML estimates is that estimates often do not exist in a range that is close to the optimal smoothing parameter whereas WLS approaches always yield estimates. Which one is better, ML of WLS, depends on the underlying structure. If all of the probabilities are below or above 0.5 the shrinkage towards 0.5 makes the ML estimate inferior whereas in cases where the probabilities vary around 0.5 the WLS estimate performs superior to the ML estimate.

References

Grizzle, J. E., Starmer, C. F., and Koch, G. G. (1969). Analysis of categorical data by linear models. *Biometrika 28*, 137–156.

Hastie, T. and Tibshirani, R. (1993). Varying–coefficient models. *J. Roy. Statist. Soc. B 55*, 757–796.

Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *J. Amer. Statist. Assoc. 82*, 559–568.

Tutz, G. and Kauermann, G. (1995). On bias correction for local likelihood estimation in varying coefficient models. Forschungsberichte des Fachbereichs Informatik 95–5, Technische Universit"at Berlin.

Tutz, G. and Kauermann, G. (1996). Locally weighted least squares in categorical varying-coefficient models. (submitted).