
Nonparametric Components in Multivariate Discrete Choice Models

Marlene Müller

Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin

marlene.mueller@gmx.de

<http://ise.wiwi.hu-berlin.de/~marlene>

Plan

- Aim of the Analysis
- Multivariate Model
- Multivariate GEE
- Semiparametric Approach
- Conclusions

Aim of the Analysis

How do covariates for household i influence the possession of different assets?

$$Y_{i1} = \mathbf{1}[\text{savings books}],$$

$$Y_{i2} = \mathbf{1}[\text{stocks, bonds}],$$

$$Y_{i3} = \mathbf{1}[\text{life insurance, private pensions}],$$

$$Y_{i4} = \mathbf{1}[\text{own flat/ house}],$$

where $\mathbf{1}$ denotes the indicator function. (= 1 if household i possesses the corresponding goods).

Data

The INSEE data set "Les Actifs Financiers 1991" covers data of 9530 French households. The about 4000 variables can be roughly grouped as:

<i>"Actifs"</i>	bank accounts, savings books, savings for future house ownership, stocks and bonds, life insurances and private pension plans, own flats/houses, grounds, professional goods (buildings, grounds, firms),
<i>Household variables</i>	household composition, age, education, profession, current occupation,
<i>Ressources</i>	income
<i>Biography</i>	socio-economic data on parents, heritages,
<i>Bank credits</i>	data on bank credits taken by the household.

	household types			
	(S) single	(C) couple	(S+K) single+children	(C+K) family
Y_1 savings book	0.73	0.81	0.56	0.72
Y_3 stocks & bonds	0.25	0.36	0.08	0.22
Y_5 insurance/pension	0.37	0.50	0.35	0.46
Y_6 real estate	0.40	0.64	0.27	0.67
X_1 female	0.54	0.08	0.89	0.08
X_2 Île de France	0.24	0.18	0.23	0.14
X_3 "cadres"	0.20	0.29	0.07	0.20
X_4 "intermediaires"	0.23	0.25	0.25	0.24
X_5 "employees"	0.31	0.13	0.51	0.13
X_6 age	46.88 (14.52)	51.78 (14.27)	35.59 (5.65)	36.40 (5.63)
X_7 income	98666.77 (70631.16)	174098.00 (100252.14)	85215.88 (48924.68)	172339.81 (108547.09)
n	834	1102	168	1907

Table 1. Descriptive Statistics

Marginal Model

e.g. Probit

$$P(Y_i = 1|X = x_i) = E(Y_i|X = x_i) = \Phi(\eta(x_i))$$

where $\Phi(\bullet)$ denotes the Gaussian cdf

special case of **generalized linear model (GLM)**

$$E(Y_i|X = x_i) = G(\eta(x_i)), \quad \eta(x_i) = \beta^T x_i.$$

Likelihood & Quasi-Likelihood

likelihood \Leftrightarrow exponential family

$$f(y_i, \theta_i, \phi) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} - c(y_i, \phi) \right\}$$

maximum-likelihood

$$\sum_i \log f(Y_i, \theta_i, \phi) \rightarrow \max \iff \sum_i \theta'_i \{Y_i - b'(\theta_i)\} \stackrel{!}{=} 0$$

$$\stackrel{\text{GLM}}{\iff} \sum_i D_i V^{-1}(\mu_i) \{Y_i - \mu_i\} \stackrel{!}{=} 0$$

where in GLM $EY_i = \mu_i = G(\eta(x_i))$ and $Var(Y_i) = a(\phi)V(\mu_i)$ and D_i only depends on $G(\bullet)$, $\eta(\bullet)$ and x_i

Quasi-/Pseudo-Likelihood

$$EY_i = \mu_i, \quad \text{Var}(Y_i) = a(\phi)V(\mu_i)$$

$$\underset{\text{GLM}}{\iff} \sum_i D_i V^{-1}(\mu_i) \{Y_i - \mu_i\} \stackrel{!}{=} 0$$

\Rightarrow consistency, if $EY_i = \mu_i$ correctly specified

\Rightarrow “Estimating Equation”

Generalized Estimating Equations

- if expectation $EY_i = G(\beta^T x_i)$ correctly specified, then the pseudo-ML is also consistent (Gourieroux, Monfort & Trognon 1984)
- in particular, consistency of $\hat{\beta}$ is not influenced by misspecified $V(\bullet) \Rightarrow$ "Working Matrix"
- multivariate: "Generalized Estimating Equations" (GEE) (Liang & Zeger 1986, Zeger & Liang 1986)
- to estimate variance/covariance structure consistently, expectation and variance/covariance need to be specified correctly

Multivariate Model

- conditional model:

$$E(Y_{ij} | Y_{ik, k \neq j}, X_{ij}) = f_j(Y_{ik, k \neq j}, X_{ij}),$$

influence of X_{ij} might be hidden because of using $Y_{ik, k \neq j}$

- marginal model:

$$E(Y_{ij} | X_{ij}) = f_j(X_{ij}).$$

covariance structure within the Y_{ij} needs to be estimated

- **here we consider**

$$E(Y_{ij} | X_i) = f_j(X_i),$$

where $j = 1, \dots, r$ (dimension) and $i = 1, \dots, n$ (sample size)

Covariance structure

latent variable

$$Y_{ij} = \mathbf{1}[Y_{ij}^* > 0],$$

where Y_{ij}^* denotes the utility of household i to have asset j

model for Y_{ij}^*

$$Y_{ij}^* = \eta_j(X_i) + \varepsilon_{ij}$$

with

$$E\varepsilon_{ij} = 0, \quad \text{Var}(\varepsilon_{ij}) = 1, \quad \text{Corr}(\varepsilon_{ij}, \varepsilon_{ik}) = \rho_{jk}^*.$$

Association Measures

tetrachoric correlation between the Y^* :

$$\rho_{jk}^* = \text{Corr}(Y_{ij}^*, Y_{ik}^*),$$

alternatives

- correlation between the responses

$$\rho_{jk} = \text{Corr}(Y_{ij}, Y_{ik}),$$

- odds-ratio:

$$\omega_{jk} = \frac{P(Y_{ij} = 1, Y_{ik} = 1) \cdot P(Y_{ij} = 0, Y_{ik} = 0)}{P(Y_{ij} = 1, Y_{ik} = 0) \cdot P(Y_{ij} = 0, Y_{ik} = 1)}.$$

independence $\rightarrow \omega_{jk} = 1$; $\log \omega_{jk} \in (-\infty, \infty)$.

Multivariate GEE

$$\sum_i \mathcal{D}_i^T \mathcal{V}_i^{-1} \mathcal{S}_i = 0, \quad \text{where } \mathcal{D}_i = \mathcal{J}_i \mathcal{X}_i, \quad \mathcal{J}_i = \text{diag} \left(\frac{\partial \mu_{ij}}{\partial \eta_{ij}} \right)$$

Fisher Scoring algorithm:

$$\begin{aligned} \beta_{new} &= \beta + \left(\sum_i \mathcal{D}_i^T \mathcal{V}_i^{-1} \mathcal{D}_i \right)^{-1} \sum_i \mathcal{D}_i^T \mathcal{V}_i^{-1} \mathcal{S}_i \\ &= \left(\sum_i \mathcal{X}_i^T \mathcal{W}_i \mathcal{X}_i \right)^{-1} \sum_i \mathcal{X}_i^T \mathcal{W}_i \mathcal{Z}_i \end{aligned}$$

with

$$\mathcal{W}_i = \mathcal{J}_i \mathcal{V}_i^{-1} \mathcal{J}_i, \quad \mathcal{Z}_i = \mathcal{X}_i \beta + \mathcal{J}_i^{-1} \mathcal{S}_i$$

Example: Bivariate GEE

Sei $\sigma_{i12} = \sigma(X_i, \beta_1, \beta_2, \gamma) = \text{Cov}(Y_{i1}, Y_{i2})$, then

$$\sum_i \mathcal{D}_i^T \mathcal{V}_i^{-1} \mathcal{S}_i = 0$$

where

$$\mathcal{D}_i^T = \begin{pmatrix} \frac{\partial}{\partial \beta_1} \mu_{i1} & \frac{\partial}{\partial \beta_1} \mu_{i2} & \frac{\partial}{\partial \beta_1} \sigma_{i12} \\ \frac{\partial}{\partial \beta_2} \mu_{i1} & \frac{\partial}{\partial \beta_2} \mu_{i2} & \frac{\partial}{\partial \beta_2} \sigma_{i12} \\ 0 & 0 & \frac{\partial}{\partial \gamma} \sigma_{i12} \end{pmatrix},$$

and

$$\mathcal{S}_i = \begin{pmatrix} Y_{i1} - \mu_{i1} \\ Y_{i2} - \mu_{i2} \\ S_{i12} - \sigma_{i12} \end{pmatrix}, \quad S_{i12} = (Y_{i1} - \mu_{i1})(Y_{i2} - \mu_{i2}).$$

”Working” Covariance Structure

$$\mathcal{V}_i = a(\phi) \{\text{diag}(\text{Var}(\mathcal{S}_i))\}^{1/2} \mathcal{R} \{\text{diag}(\text{Var}(\mathcal{S}_i))\}^{1/2}$$

simplest model: independence $\Rightarrow \mathcal{R} = \mathcal{I}$

$$\mathcal{V}_i = a(\phi) \begin{pmatrix} V(\mu_{i1}) & 0 & 0 \\ 0 & V(\mu_{i2}) & 0 \\ 0 & 0 & V(\mu_{i1})V(\mu_{i2}) \end{pmatrix},$$

Fisher Scoring algorithm

$$\begin{pmatrix} \beta_1 \\ \beta_2 \\ \gamma \end{pmatrix}_{new} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \gamma \end{pmatrix} + \left(\sum_i \mathcal{D}_i^T \mathcal{V}_i^{-1} \mathcal{D}_i \right)^{-1} \left(\sum_i \mathcal{D}_i^T \mathcal{V}_i^{-1} \mathcal{S}_i \right).$$

Probit GEE Correlations

	(S)	(C)	(S+K)	(C+K)
ρ_{12}^*	0.20	0.27	-0.28	0.20
ρ_{13}^*	0.26	0.09	-0.10	0.19
ρ_{14}^*	0.05	0.12	-0.06	0.05
ρ_{23}^*	0.43	0.29	0.00	0.25
ρ_{24}^*	0.16	0.15	0.48	0.05
ρ_{34}^*	0.10	0.14	0.11	-0.01

Table 2. Estimated correlations using probit for different household types.

Semiparametric Approach

$$E(Y|x, t) = G(\eta(x, t)), \quad G(\bullet) \text{ known (z.B. } G = \Phi)$$

- estimating GLM/GEE:
weighted LS using the adjusted dependent variable z
(Nelder & Wedderburn 1972)
- estimating a semiparametric GEE:
semiparametric algorithm using the adjusted dependent variable z
(cf. Wild & Yee (1996))

Generalized Linear Model (GLM)

$$\star \eta(x, t) = \alpha + \beta^T x + \delta^T t$$

Generalized Partial Linear Model (GPLM)

$$\star \eta(x, t) = \beta^T x + m(t)$$

Generalized Additive Model (GAM)

$$\star \eta(x, t) = \alpha + \beta^T x + \sum_{j=1}^q m_j(t_j)$$

Household Portfolio Allocation

$n = 834$, Singles

$$Y_{i1} = \mathbf{1}[\text{savings books}],$$

$$Y_{i2} = \mathbf{1}[\text{stocks \& bonds}],$$

$$Y_{i3} = \mathbf{1}[\text{insurance/pension}],$$

$$Y_{i4} = \mathbf{1}[\text{real estates}],$$

possible nonparametric components:

$$T_i = \text{age, income,}$$

linear covariables:

$$X_i = \text{female, Île de France, occupation}$$

Model to Estimate

marginal

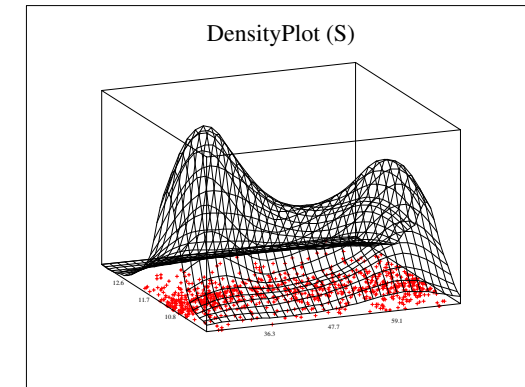
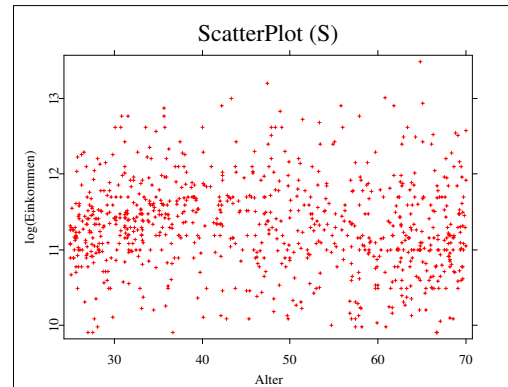
$$E(Y_{ij}|x_i, t_i) = P(Y_{ij} = 1|x_i, t_i) = \Phi\{\beta_j^T x_i + m_j(t_i)\}$$

covariance structure = tetrachoric correlations

$$\rho_{jk}^* = \text{Corr}(\varepsilon_{ij}, \varepsilon_{ik})$$

Exploring the Data

Singles (S)



Families (C+K)

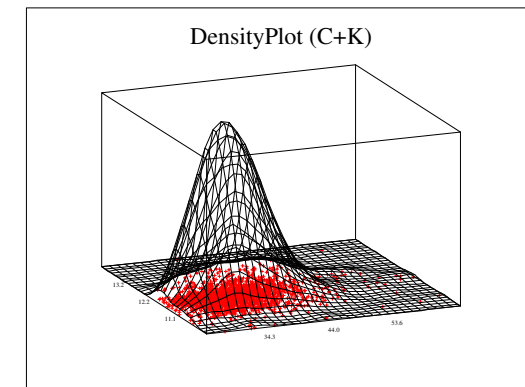
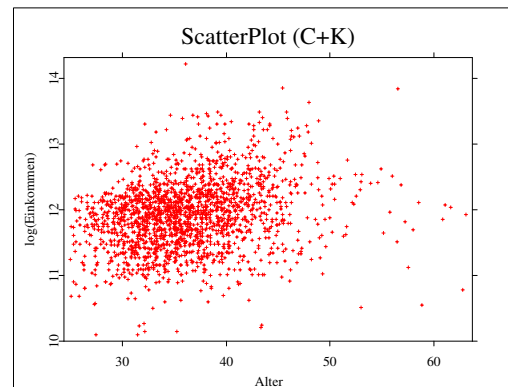


Figure 1. Scatterplot and bivariate density estimates for age and log(income).

Marginal Coefficients

e.g. singles (S) and savings books, $n = 834$

	Variable	Linear	Part. Linear $m(t_1)$	Part. Linear $m(t_1, t_2)$
Y ₁ savings books	intercept	2.371 (1.97)	–	–
	female	0.295 (2.84)	0.292 (2.80)	0.285 (2.71)
	Île de France	-0.043 (-0.37)	-0.052 (-0.44)	-0.042 (-0.36)
	"cadres"	0.859 (4.49)	0.889 (4.53)	0.824 (4.43)
	"intermediaires"	0.403 (2.72)	0.339 (2.29)	0.340 (2.33)
	"employees"	0.138 (1.05)	0.109 (0.84)	0.113 (0.87)
	age	0.004 (1.22)	0.005 (1.52)	–
	log(income)	-0.211 (-1.96)	–	–

Table 3. Coefficients (t values) for probit fit and partial linear probit models.

Significant Effects

	(S)	(C)	(S+K)	(C+K)
Y_1 savings books	female occupation (income)			-female $\hat{\beta}$ le income
Y_2 stocks & bonds	occupation age income	occupation age income		$\hat{\beta}$ le occupation income
Y_3 life insurance/pension	female $\hat{\beta}$ le (occupation) age income	age income	age	$\hat{\beta}$ le (occupation) (income)
Y_4 real estates	$\hat{\beta}$ le (occupation) age income	$\hat{\beta}$ le age income	age (income)	$\hat{\beta}$ le (-occupation) age income

Table 4. Significance for the coefficients of the parametric fit.

Nonparametric Curves

(S)

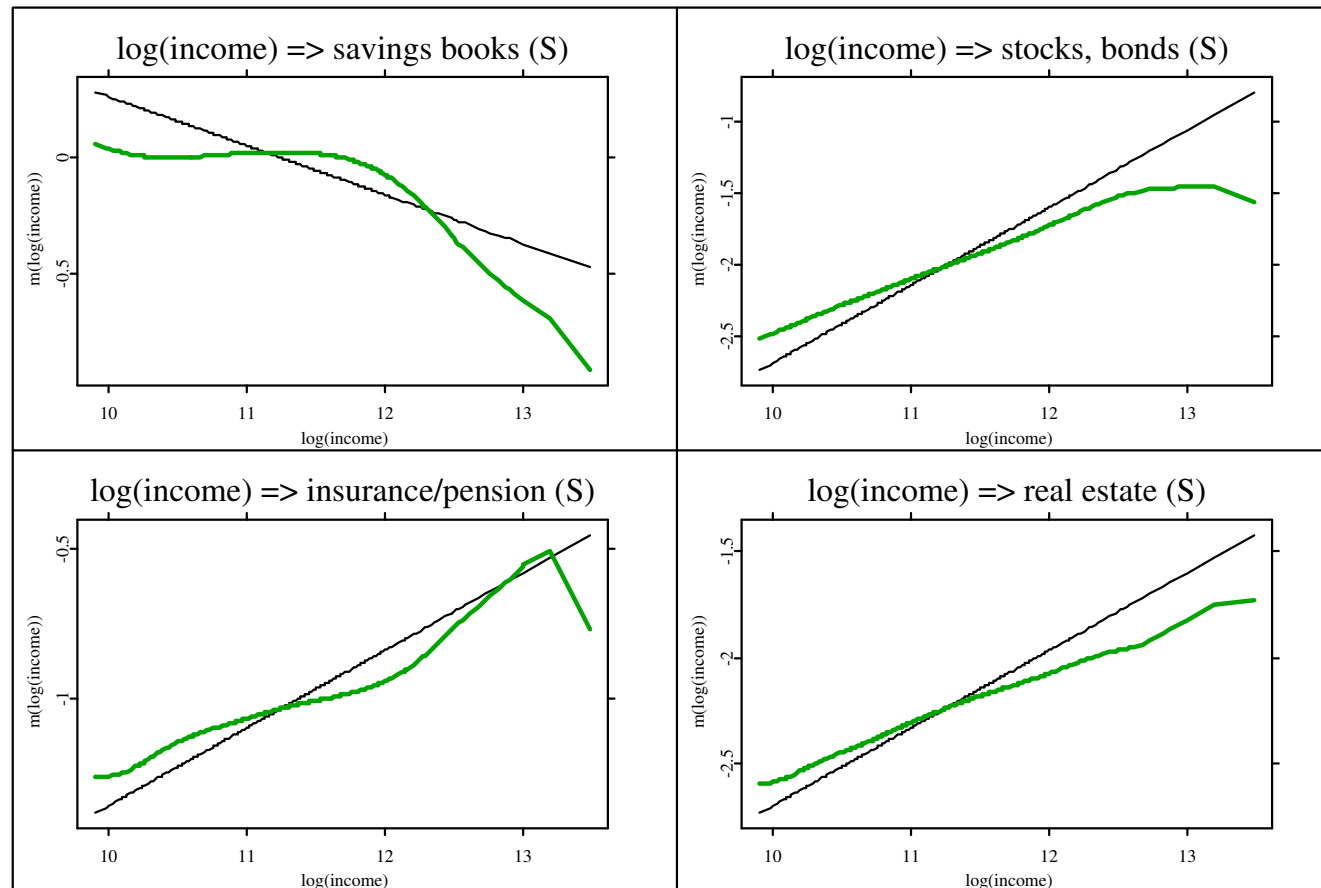


Figure 2. Effect of $\log(\text{income})$. Nonparametric curves and linear fits.

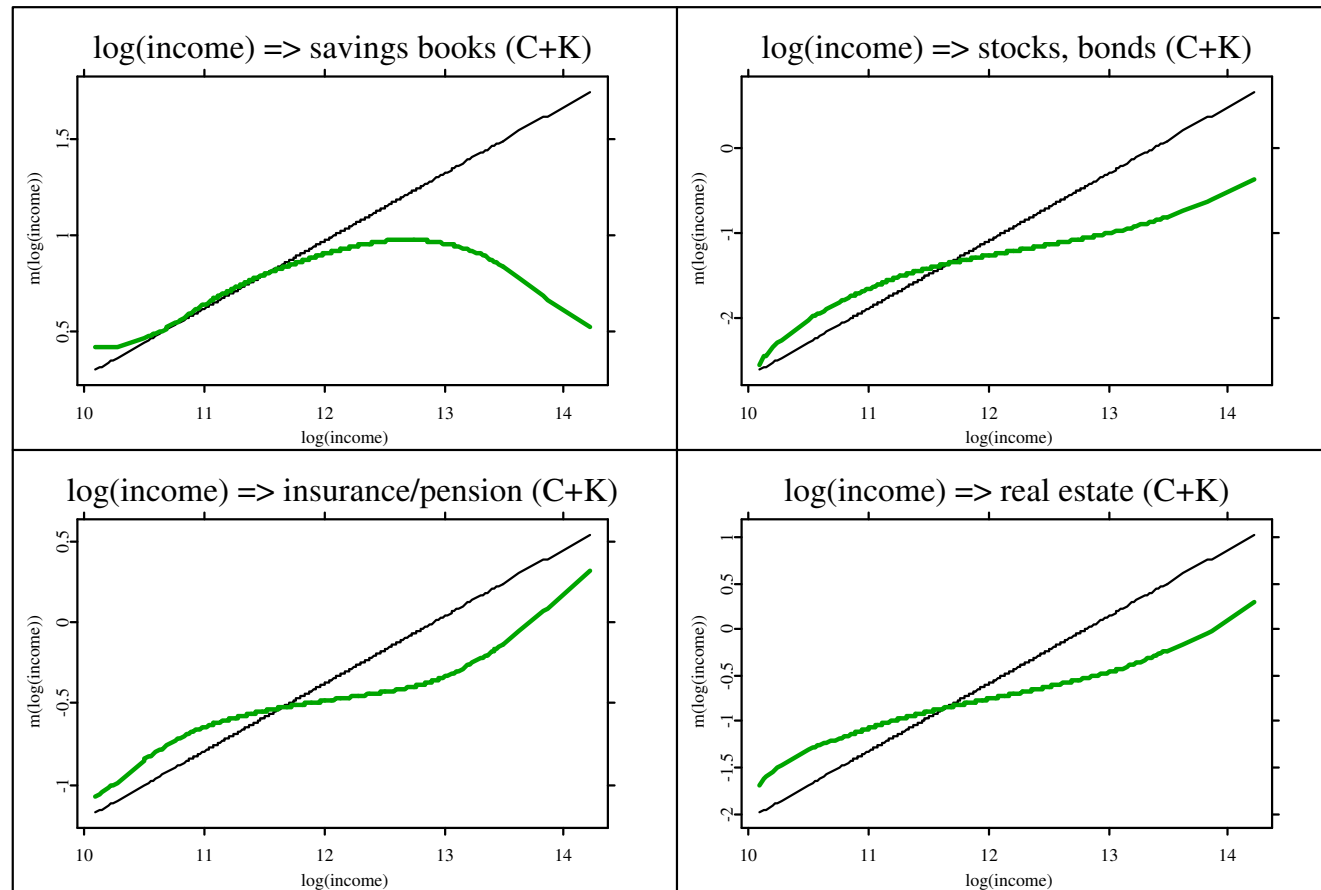
(C+K)

Figure 3. Effect of $\log(\text{income})$. Nonparametric curves and linear fits.

Nonparametric Surfaces

(S)

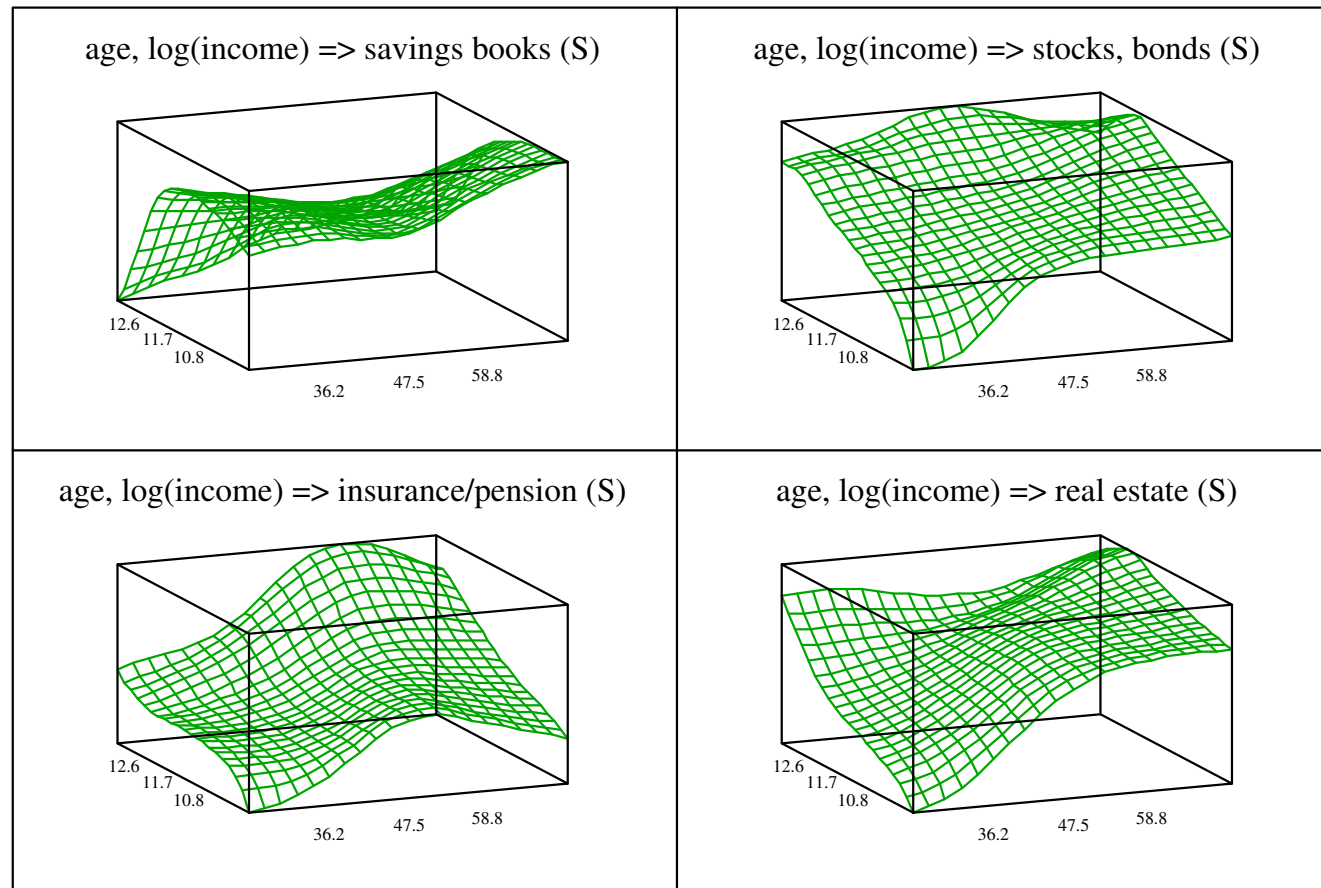


Figure 4. Effect of age and log(income).

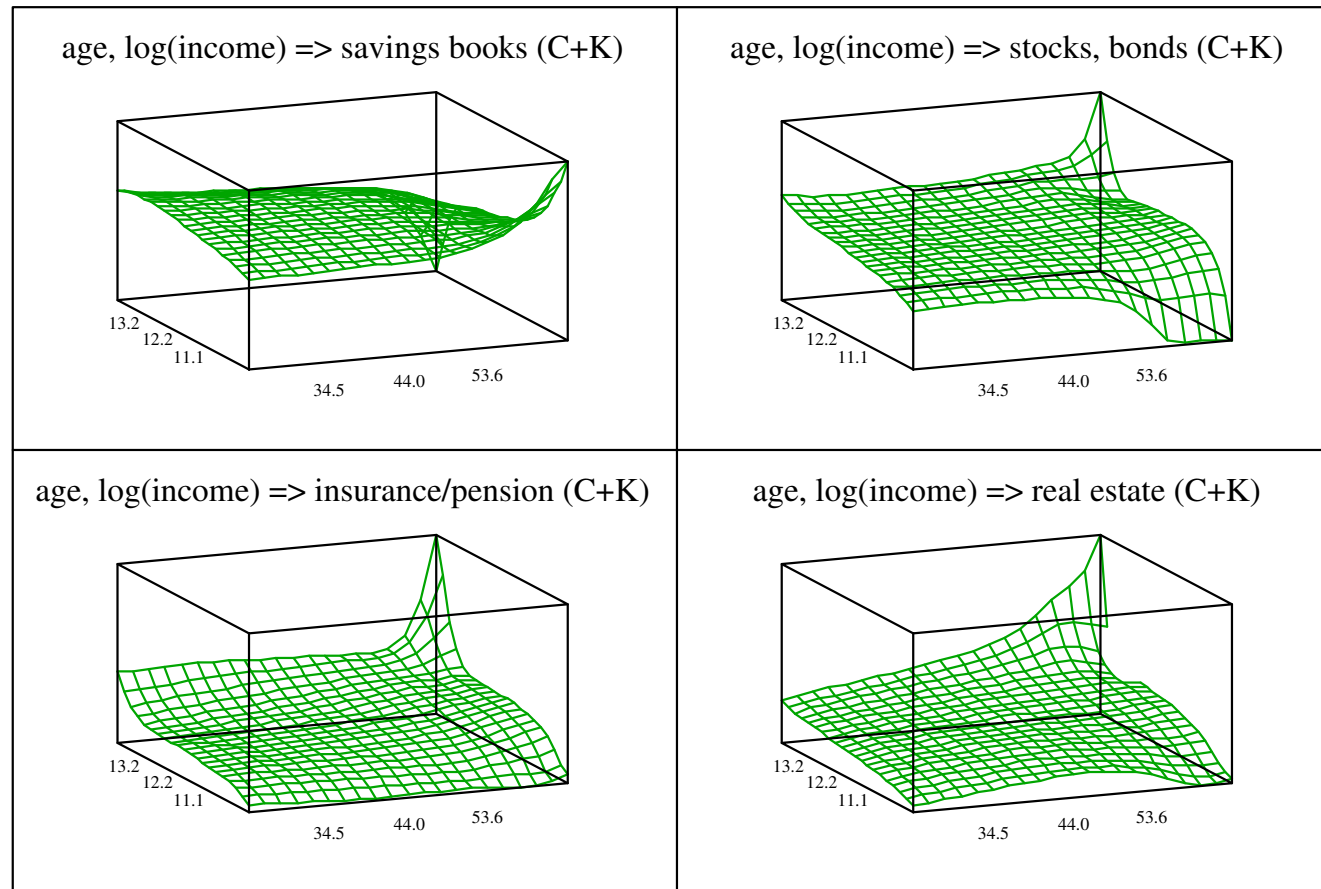
(C+K)

Figure 5. Effect of age and log(income).

Effect of age

(S)

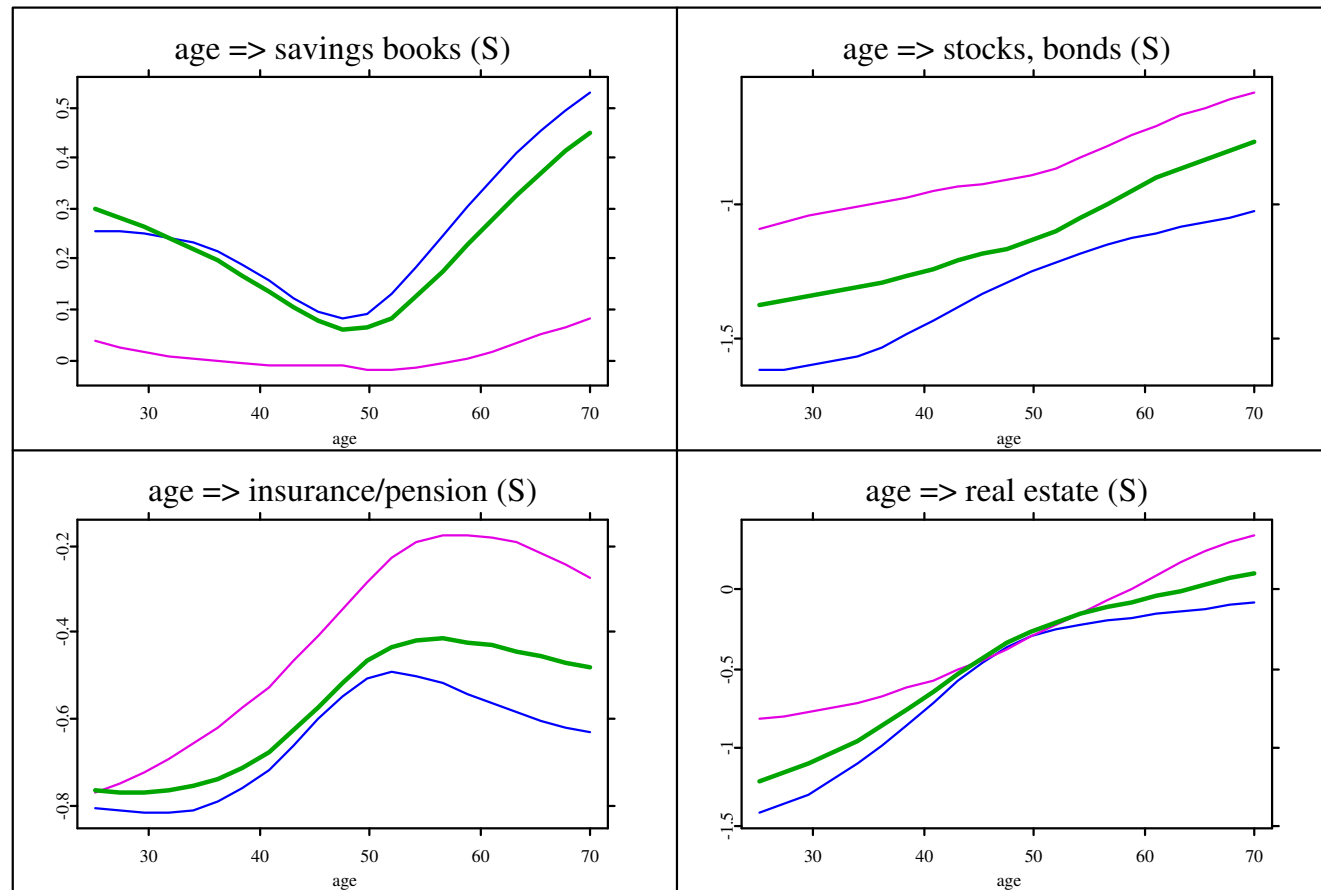


Figure 6. Effect of age. (Income fixed.)

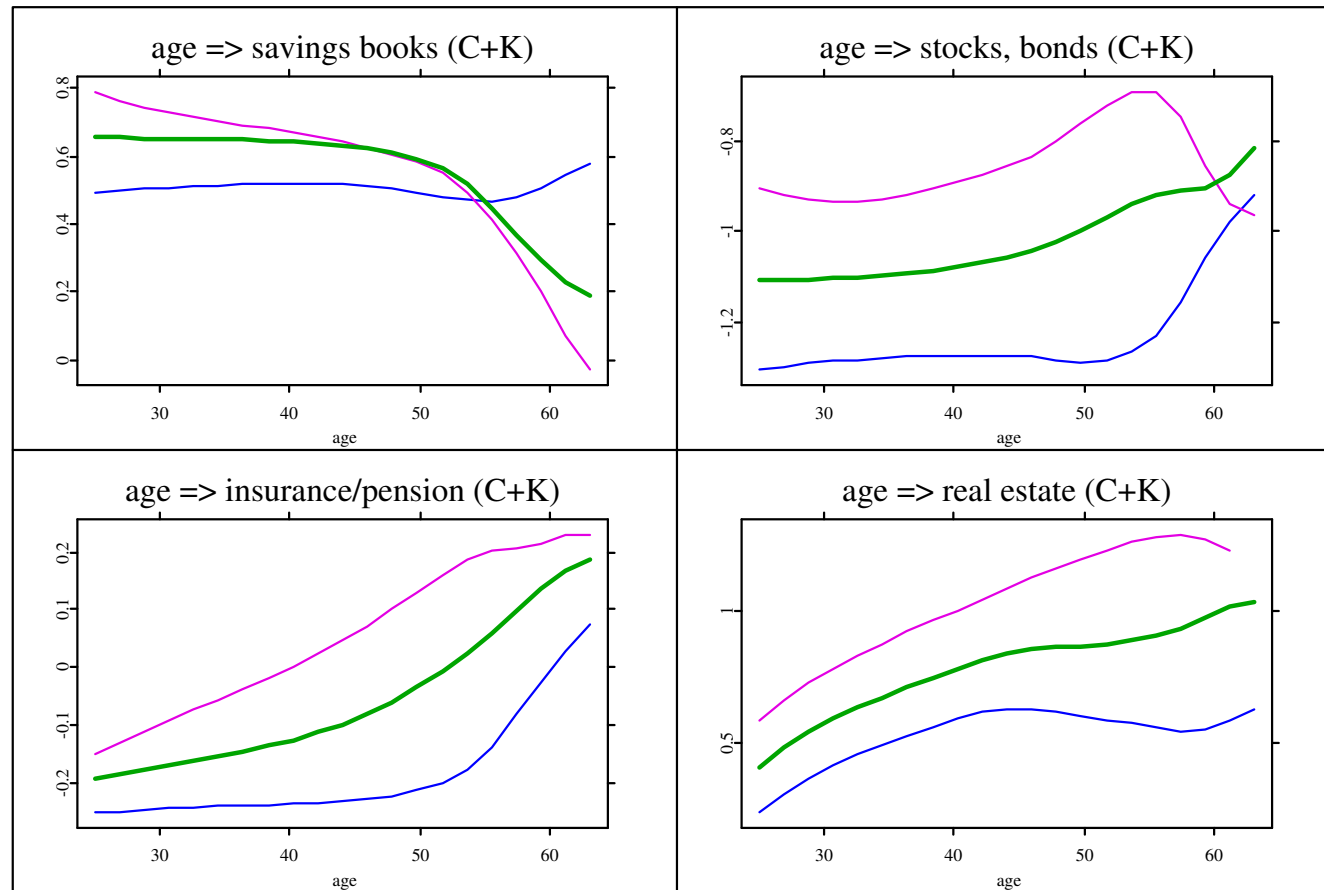
(C+K)

Figure 7. Effect of age. (Income fixed.)

Effect of log(income)

(S)

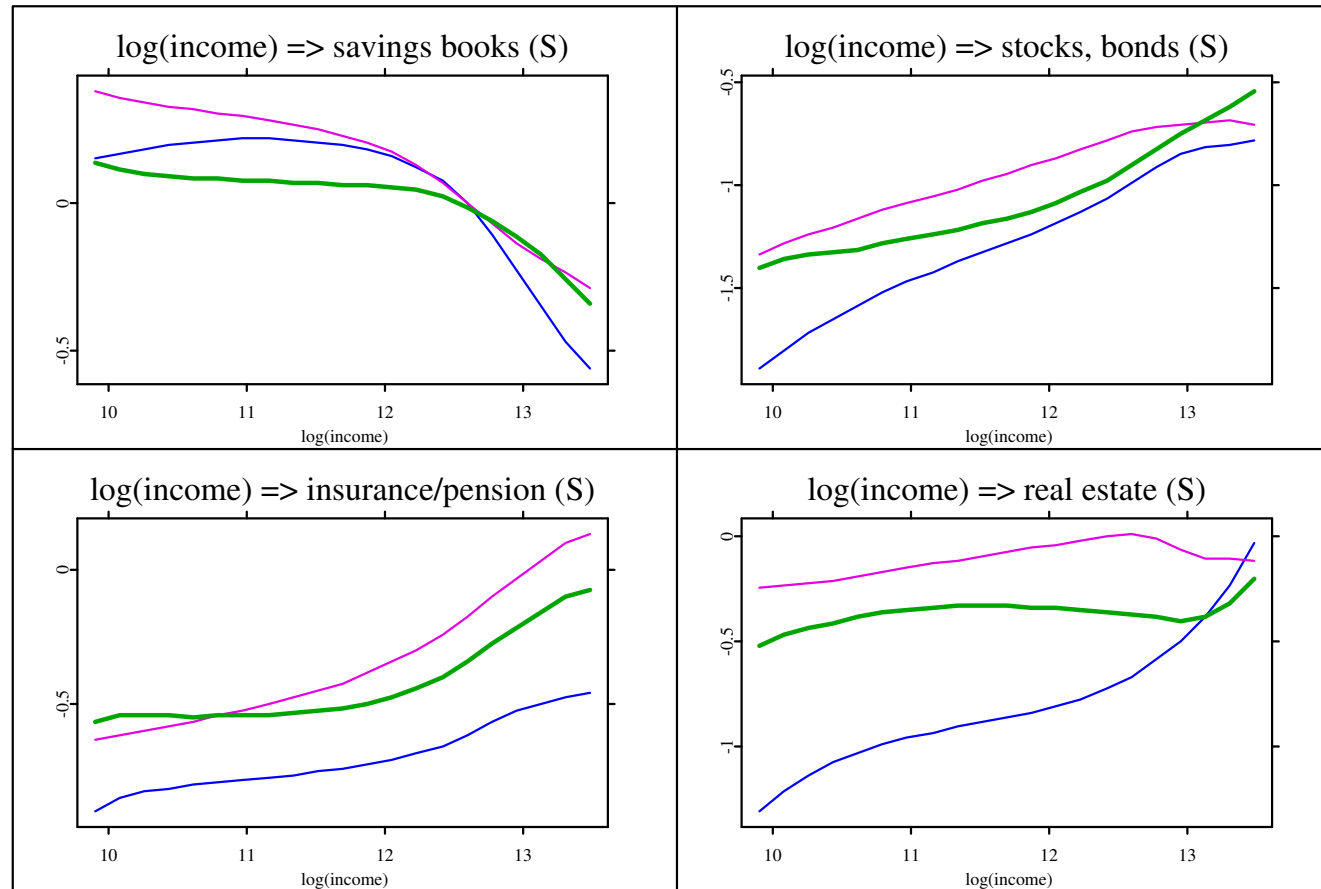


Figure 8. Effect of income. (Age fixed.)

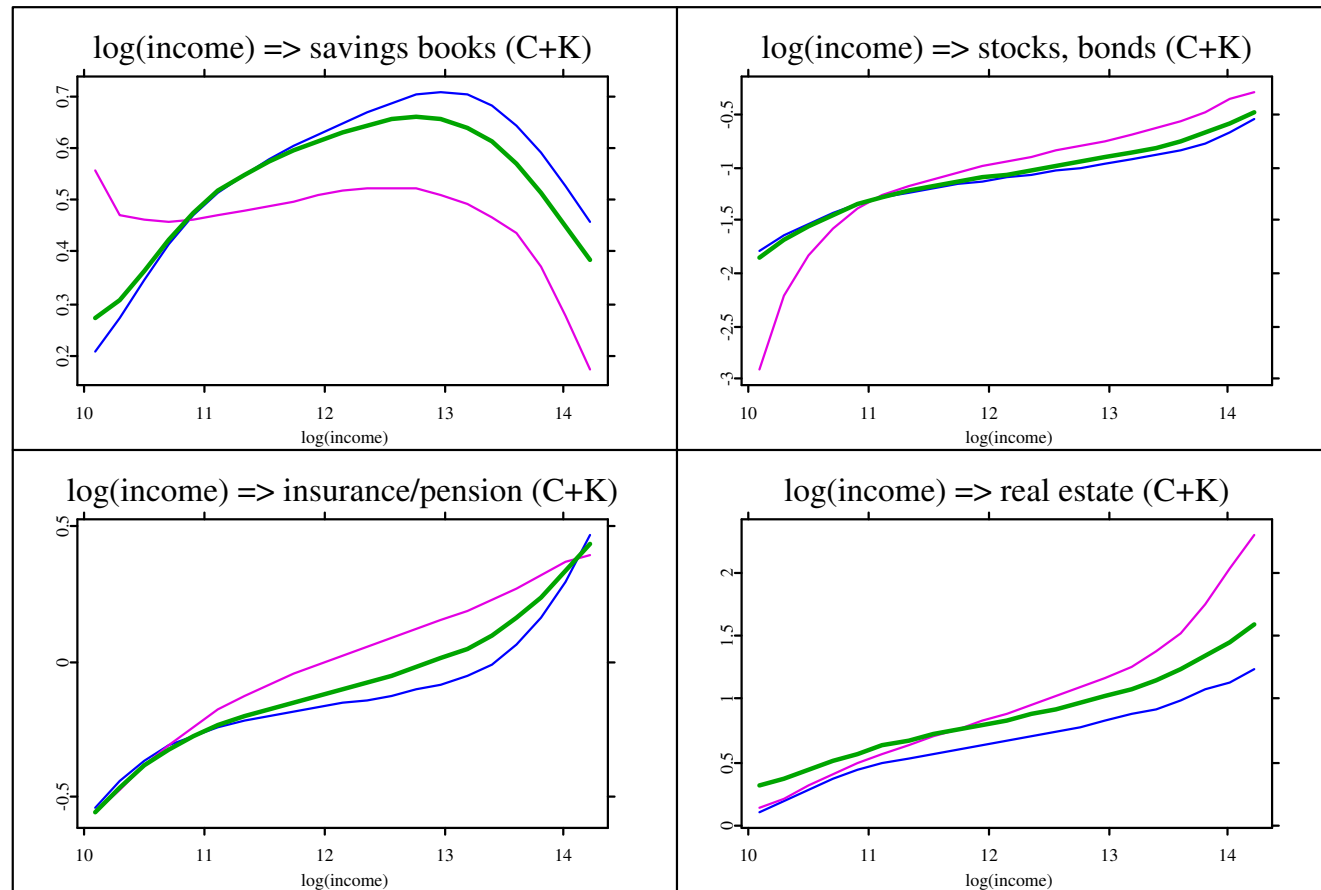
(C+K)

Figure 9. Effect of income. (Age fixed.)

Tetrachoric correlations

parametric

	(S)	(C)	(S+K)	(C+K)
ρ_{12}^*	0.20	0.27	-0.28	0.20
ρ_{13}^*	0.26	0.09	-0.10	0.19
ρ_{14}^*	0.05	0.12	-0.06	0.05
ρ_{23}^*	0.43	0.29	0.00	0.25
ρ_{24}^*	0.16	0.15	0.48	0.05
ρ_{34}^*	0.10	0.14	0.11	-0.01

semiparametric

	(S)	(C)	(S+K)	(C+K)
ρ_{12}^*	0.18	0.25	-0.30	0.20
ρ_{13}^*	0.28	0.12	-0.11	0.20
ρ_{14}^*	0.06	0.16	-0.06	0.08
ρ_{23}^*	0.42	0.31	-0.03	0.25
ρ_{24}^*	0.16	0.16	0.47	0.08
ρ_{34}^*	0.09	0.13	0.09	0.01

Table 5. Correlations for different household types.

Conclusion

- GEE permit the estimation of nonparametric components in multivariate generalized linear models
- a "working covariance matrix" reduces the computational effort
- outlook: semiparametric estimation of correlations

References

- Fahrmeir, L. & Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer.
- Gourieroux, C., Monfort, A. & Trognon, A. (1984). Pseudomaximum likelihood methods: Theory, *Econometrica* **52**: 681–700.
- Härdle, W., Mammen, E. & Müller, M. (1998). Testing Parametric versus Semiparametric Modelling in Generalized Linear Models. *Journal of the American Statistical Association*, **93**(444), 1461–1474.
- Liang, K. Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**: 13–22.
- Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models, *Journal of the Royal Statistical Society, Series A* **135**(3): 370–384.

- Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation, *Biometrics* **44**: 1033–1048.
- Severini, T. A. & Staniswalis, J. G. (1994). Quasi-likelihood estimation in semiparametric models, *Journal of the American Statistical Association* **89**: 501–511.
- Severini, T. A. & Wong, W. H. (1983). Generalized profile likelihood and conditionally parametric models, *Annals of Statistics* **20**: 1768–1802.
- Wild, C. J. & Yee, T. W. (1996). Additive Extensions to Generalized Estimating Equation Methods, *Journal of the Royal Statistical Society, Series B* **58**: 711–725.
- Zeger, S. L. & Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes, *Biometrics* **42**: 121–130.
- Zhao, L. P. & Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model, *Biometrika* **77**: 642–648.