Web appendix for

Smooth quantile based modeling of brand sales, price and promotional effects from retail scanner panels,

Journal of Applied Econometrics

Harry Haupt*, Kathrin Kagerer† and Winfried J. Steiner‡

1 Computational issues

All of the computations in the paper are implemented using the open-source software `R` (version 2.15.0, 32-bit). B-spline basis functions are generated using the function `splineDesign` from the `spline` package. For the estimations of the conditional mean with monotonicity constraint

$$\alpha_{j-1}(\vartheta) - \alpha_j(\vartheta) \geq 0, \quad \text{for } j = -(k-1), \ldots, m,$$

[equation (5) in the paper] (i.e. for the constrained B-spline model and the smoothing spline model) we employ the function `pcls` from the `mgcv` package (Wood 2012). This function is based on an optimization process using quadratic programming with respect to the monotonicity constraint which is imposed by means of an inequality constraint. For quantile

*Prof. Dr. Harry Haupt, Department of Statistics, University of Passau, Innstr. 27, 94032 Passau, Germany. Tel. +49 851 509-2560

Corresponding author: harry.haupt@uni-passau.de

†Dr. Kathrin Kagerer, Faculty of Business, Economics and Management Information Systems, University of Regensburg, Universitätsst. 31, 93053 Regensburg, Germany.

kathrin.kagerer@wiwi.uni-regensburg.de

‡Prof. Dr. Winfried J. Steiner, Department of Marketing, Clausthal University of Technology, Julius-Albert-Str. 2, 38678 Clausthal-Zellerfeld, Germany, Tel. +49 5323 72-7650.

winfried.steiner@tu-clausthal.de

regression, routines from the `quantreg` package from Koenker (2012) are used, based on the respective works of Koenker et al. (1994), He and Shi (1994, 1998), He et al. (1998), He and Ng (1999), Koenker and Mizera (2004), and Koenker (2005, 2010). Monotonicity constraints can be easily implemented in the optimization process via inequality constraints (see e.g. Koenker and Ng 2005), where we apply the option `method="fnc"` for the `rq`-function of the `quantreg` package. A detailed description of the R-functions we use is given in Table 1.

model	LS estimations	QR estimations	chosen parameters
1: parametric	`ls` (base) `summary` (base)	`rq` (quantreg) `summary.rq` (quantreg)	-
2: B-spline	`ls` (base) `summary` (base) `splineDesign` (splines)	`rq` (quantreg) `summary.rq` (quantreg) `splineDesign` (splines)	knot sequence: equidistant, $m = 4$
3: monotone B-spline	`pcls` (mgcv) `splineDesign` (splines) bootstrapped std. err.	`rq` (quantreg, method="fnc") `summary.rq` (quantreg) `splineDesign` (splines)	knot sequence: equidistant, $m = 4$ inequ. restr.: constr. matrix: $\left(\begin{smallmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{smallmatrix}\right)$, $\mathbf{C} = \begin{pmatrix} -1 & 1 & 0 & \\ 0 & -1 & 1 & \\ & & \ddots & \ddots \end{pmatrix}$
4: smoothing spline	`pcls` (mgcv) `splineDesign` (splines) bootstrapped std. err.	`rqss` (quantreg) `qss` (quantreg, constr.="D") `summary.rqss` (quantreg)	knot sequence: distinct obs. inequ. restr.: see 3 smoothing parameter: $\lambda = 1$

Table 1: Applied R-functions (and package they are contained in, partly with function option) for model estimations.

2 Model Specification and Data

We illustrate the quantile regression methods outlined in Section 2 of the paper in an empirical application using weekly store-level scanner data from Dominick's Finer Foods, a large retail chain in the Chicago metropolitan area. The data is publicly available at `http://research.chicagobooth.edu/marketing/databases/dominicks`, provided by the James M. Kilts Center, University of Chicago Booth School of Business. In particular, the data set includes weekly unit sales, retail prices and display activities on the store level for nine brands in the refrigerated orange juice category (package size 64 oz) and contains a balanced (i.e. no product introductions occurred during the considered time span) panel data set for a cross-section of 46 stores of the retail chain over a time span of 88 weeks resulting in a total of 4048 observations. Among the nine brands are three premium brands, five national brands and the supermarket's own store brand. For those nine brands, Table 2 provides summary statistics for average weekly prices and market shares (pooled across stores). Differences in quality across the three tiers are well represented by higher (lower) average prices for higher (lower) quality tier brands. For brands that comprise several sub-brands (e.g., with different levels of pulp, with/without Calcium) those sub-brands were merged in advance to prevent multicollinearity problems. In this case, prices represent sales-weighted averages of the sub-brands' prices. In the following, we exemplarily study the national brand Minute Maid.

		average prices ($)			average market shares (%)		
brand	abbreviation	[min, max]	mean	sd	[min, max]	mean	sd
Florida Natural	flona	[1.56, 3.16]	2.85	0.33	[0.57, 47.79]	4.55	6.12
Tropicana Pure	tropu	[1.58, 3.50]	2.94	0.50	[3.24, 55.26]	14.18	11.24
Minute Maid Premium	mmprm	[1.58, 3.50]	2.94	0.51	[1.24, 20.28]	5.40	4.13
Minute Maid	mmaid	[1.29, 2.88]	2.22	0.39	[3.26, 85.54]	20.18	20.41
Citrus Hill	cithi	[1.08, 2.78]	2.29	0.31	[1.28, 75.19]	7.51	10.90
Florida Gold	flogo	[1.31, 2.78]	2.17	0.35	[0.20, 34.26]	2.68	4.52
Tree Fresh	trefr	[1.06, 2.44]	2.13	0.27	[0.80, 39.78]	4.01	5.04
Tropicana	tropi	[1.49, 2.70]	2.19	0.33	[2.06, 71.24]	20.08	21.82
Dominick's	domin	[0.99, 2.44]	1.75	0.39	[1.38, 78.24]	21.41	20.28

Table 2: Summary statistics for average weekly prices and market shares (pooled across stores, i.e. summary of 88 weekly observations which are each pooled from 46 store observations).

To account for the price-quality tier structure of the data, one could capture cross-price effects at the tier-level (e.g., Steiner et al. 2007). This would yield a more parsimonious model on the one hand, but may affect the estimates of own-price elasticities and/or the predictive performance of a model due to that aggregation on the other hand. One solution for this trade-off is to model at least separate cross-price effects for the direct competitors of Minute Maid in its national brand tier (Florida Gold, Tree Fresh, Tropicana, Citrus Hill). In addition, it is reasonable to model the cross-price effect of Minute Maid Premium separately, as a Minute Maid manager who is responsible for both Minute Maid brands would be especially concerned with this effect. The cross-price effect for the other two premium brands (Florida Natural, Tropicana Pure) is instead captured jointly by defining $\texttt{cross\_premium}_{s,t}$ to represent the price of the respective cheaper brand in week t and store s. Using this minimum rule to aggregate the cross-prices of Florida Natural and Tropicana

Pure seems to be a natural choice since a national brand buyer is usually attracted by price cuts in the premium brand tier which is well-reflected by the lowest price of premium brands rather than a mean price, for instance. We therefore include the following cross-price covariates in our model: `cross_premium`, `cross_mmprm`, `cross_cithi`, `cross_flogo`, `cross_trefr`, `cross_tropi`, `cross_domin` (please compare Table 2). Given this cross-price specification, a standard parametric model for analyzing store-level data can be stated as follows:

$$\begin{aligned}\log(\texttt{sales})^b_{s,t} = {} & \beta^b_0 + \beta^b_1 \log(\texttt{price})^b_{s,t} + \beta^b_2 \log(\texttt{cross\_premium})^b_{s,t} + \beta^b_3 \log(\texttt{cross\_mmprm})^b_{s,t} \\ & + \beta^b_4 \log(\texttt{cross\_cithi})^b_{s,t} + \beta^b_5 \log(\texttt{cross\_flogo})^b_{s,t} + \beta^b_6 \log(\texttt{cross\_trefr})^b_{s,t} \\ & + \beta^b_7 \log(\texttt{cross\_tropi})^b_{s,t} + \beta^b_8 \log(\texttt{cross\_domin})^b_{s,t} \\ & + \beta^b_9 \texttt{display}^b_{s,t} + \beta^b_{10} \texttt{end\_99}^b_{s,t} + \beta^b_{11} (\texttt{display} \cdot \texttt{end\_99})^b_{s,t} + \beta^b_{12} \texttt{holiday}_t \\ & + \beta^b_{13} \texttt{summer}_t + \beta^b_{14} \texttt{fall}_t + \beta^b_{15} \texttt{winter}_t + \sum_{o=2}^{46} \delta^b_o \texttt{store}\, o_s + u^b_{s,t}, \qquad (A, 1)\end{aligned}$$

where for brand b (here: Minute Maid) in store s and week t, $\texttt{sales}^b_{s,t}$ are the unit sales, $\texttt{price}^b_{s,t}$ is the observed own-price, $\texttt{cross\_premium}^b_{s,t}, \ldots, \texttt{cross\_domin}^b_{s,t}$ are cross-prices (as outlined above), $\texttt{display}^b_{s,t}$ is a dummy indicating whether a display is used, $\texttt{end\_99}^b_{s,t}$ is a dummy indicating whether the own-price ends in .99 cents, $\texttt{holiday}_t$ is a dummy indicating whether a holiday is in week t, $\texttt{summer}_t$, $\texttt{fall}_t$, and $\texttt{winter}_t$ are seasonal dummies indicating whether week t belongs to the summer season, the fall season, or the winter season, respectively, $\texttt{store}\, o_s$ are store dummies (for stores 2 to 46) indicating whether the observations refer to store s, and $u^b_{s,t}$ is an error term. The store dummies are used to account for cross-sectional heterogeneity in the baseline sales of Minute Maid across the different stores (e.g., due to different store sizes), while the seasonal and holiday effects accommodate time-specific shifts in baseline sales for Minute Maid. This log-log or multiplicative model is the most widely used parametric sales response model to relate brand sales to marketing instruments (Hanssens et al. 2001).

We study three semiparametric models (denoted as models $(A, 2)$, $(A, 3)$, and $(A, 4)$), where the parametric part $(\beta^b_0 + \beta^b_1 \log(\texttt{price})^b_{s,t})$ is replaced by the B-spline functional

$$\sum_{j=-(k-1)}^{m} \alpha^b_j B^{\boldsymbol{\kappa},k}_j (\log(\texttt{price})^b_{s,t})$$

in order to estimate own-price response flexibly, respectively. Models $(A, 2)$ and $(A, 3)$ have in common that they use a cubic ($k = 4$) B-spline basis with an equidistant knot sequence $\boldsymbol{\kappa}$ with $m = 4$ interior knots (cf. Section 2.4 in the paper). However, models $(A, 2)$ and $(A, 3)$ differ in that the latter has an additional monotonicity constraint imposed on the nonparametric price term to obtain a monotonically decreasing own-price effect and hence a negative own-price elasticity. Finally, model $(A, 4)$ is estimated using a linear smoothing spline with an additional monotonicity constraint imposed on the nonparametric price term (cf. Section 2.4 in the paper)[1].

In Section 5 of the paper where we also study the predictive performance of models with proxy cross-prices approximated from historical data (instead of ex post observed cross-prices considered in models $(A, 1)$ to $(A, 4)$), we mimic the situation of a manager who wants to predict the unit sales for her/his brand, but does not have complete information on future competitive pricing ex ante. In analogy, the proxy-based alternatives are referred to as models $(B, 1)$ to $(B, 4)$. Table 3 provides an overview of all models considered for estimation and prediction in the following.

covariate versions	functional forms			
	parametric	unconstrained B-spline	constrained B-spline	constrained smoothing spline
observed cross-prices	$(A, 1)$	$(A, 2)$	$(A, 3)$	$(A, 4)$
proxy cross-prices	$(B, 1)$	$(B, 2)$	$(B, 3)$	$(B, 4)$

Table 3: Models studied.

[1]We estimated models $(A, 1)$ to $(A, 4)$ with several different cross-price specifications, including flexibly estimated cross-price effects using B-splines and smoothing splines. Details on estimation results for these models can be obtained from the authors upon request.

References

Hanssens, D. M., Parsons, L. J., and Schultz, R. L. (2001). *Market Response Models: Econometric and Time Series Analysis*. Chapman and Hall, London.

He, X. and Ng, P. T. (1999). Quantile splines with several covariates. *Journal of Statistical Planning and Inference*, 75(2):343–352.

He, X., Ng, P. T., and Portnoy, S. (1998). Bivariate quantile smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):537–550.

He, X. and Shi, P. (1994). Convergence rate of B-spline estimators of nonparametric conditional quantile functions. *Journal of Nonparametric Statistics*, 3(3&4):299–308.

He, X. and Shi, P. (1998). Monotone B-spline smoothing. *Journal of the American Statistical Association*, 93(442):643–650.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Koenker, R. (2010). Additive models for quantile regression: An analysis of risk factors for malnutrition in india. In Vinod, H. D., editor, *Advances in Social Science Research Using R*, volume 196 of *Lecture Notes in Statistics*, chapter 2, pages 23–33. Springer.

Koenker, R. (2012). *quantreg: Quantile Regression*. R package version 4.79.

Koenker, R. and Mizera, I. (2004). Penalized triograms: total variation regularization for bivariate smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):145–163.

Koenker, R. and Ng, P. T. (2005). Inequality constrained quantile regression. *Sankhya: The Indian Journal of Statistics*, 67(2):418–440.

Koenker, R., Ng, P. T., and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, 84(4):673–680.

Steiner, W. J., Brezger, A., and Belitz, C. (2007). Flexible estimation of price response functions using retail scanner data. *Journal of Retailing and Consumer Services*, 14(6):383–393.

Wood, S. N. (2012). *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation*. R package version 1.7-16.