GfK Verein

WORKING PAPER SERIES

USING LOGIT ON BIG DATA – FROM ITERATIVE METHODS TO ANALYTICAL SOLUTIONS

Birgit Stoltenberg

WORKING PAPER /// NO. 3 / 2016

Using logit on big data – from iterative methods to analytical solutions

Birgit Stoltenberg*†

Abstract— *Numerical optimization for logit models is often time consuming, at least for big data. In this paper I combine several analytical approaches to solve logit models in a practical context including (sign) restrictions and constraints. The application to a household panel model compares the two approaches on several dimensions.*

Keywords— *logit; analytical solution; (sign) restriction; household panel modeling; GfK-BrandSimulator®*

1 Introduction

Logit models (also known as logistic regression models) are widely used in market research. For example, they are applied to define the functional relationship between stated purchase intentions or preferences and the actual probability of purchase. This is necessary in conjoint and modeling projects. One property of logit is that the resulting logit probabilities range between zero and one, as required for a probability. Depending on the number of choice options to model, binomial logit or multinomial logit is used.
For the binomial logit, consider respondent n ($n = 1, ..., N$) who chooses from i alternatives ($i \in \{0, 1\}$) in period or choice situation t ($t = 1, ..., T$). The probability p of n for choice i in period t will be:

$$p_{nit} = \frac{1}{1 + e^{-V_{nit}}} \qquad V_{nit} = \alpha + \beta x_{nit} \tag{1}$$

Multinomial logit models the choice between several alternatives i ($i = 1, ..., J$) and is calculated like this:

$$p_{nit} = \frac{e^{V_{nit}}}{\sum_{j=1}^{J} e^{V_{njt}}} \qquad V_{nit} = \alpha + \beta x_{nit} \tag{2}$$

Traditionally, estimation of a logit involves maximizing the likelihood function. This is typically done by numerical optimization (see Train, 2009, p.37). F.ex., within GfK for a household panel model with several multinomial logits and one binomial logit, numerical procedures are used to find the optimal coefficients today. Furthermore some coefficients are restricted and for the binomial logit there are even constraints, i.e. additional conditions on the dependent variable (DV).
But numerical algorithms – like Nelder-Mead, Golden Section Search, Newton-Raphson or Iteratively Weighted Least Squares (IWLS) – are iterative procedures with the following drawbacks: They are time consuming, depend on initial coefficients and, unfortunately, there is no guarantee for finding the global maximum.
On the other hand, they also have advantages: It is quite simple to add (sign) restrictions for the estimated coefficients. And it is even possible to add constraints. To achieve this, we use ordinary least squares (OLS) instead of maximum likelihood as objective function and simply add the constraints to the OLS.
To overcome the numerical optimization Lipovetsky developped and described an analytical closed-form solution for binomial logits (Lipovetsky, 2014). In this paper I test whether this new method is also applicable to real data and whether restrictions and constraints can be included in the new method.
The rest of the working paper is organized as follows: In section 2 Lipovetsky's method 'quasi-analytical solution' (QAS) to overcome the numerical optimization is described. Section 3 is about whether (sign) restrictions are also possible in a linear regression environment. Section 4 describes how to handle constraints in combination with QAS. In section 5 these methods are applied to an existing household panel model, including a validation. There is a short summary in section 6.

*Corresponding author, birgit.stoltenberg@gfk-verein.org

†GfK Verein

Citation: Stoltenberg, B.(2016), Using logit on big data – from iterative methods to analytical solutions, GfK Working Paper Series, No. 3 / 2016

2 The quasi-analytical solution

Lipovetsky has developped an analytical closed-form solution for binomial logits by categorical predictors. In his paper he also showed ways to generalize the method to continuous predictors – as we find them in our database. First, his approach is summarized briefly, followed by a description of the method recommended for the real data.

2.1 Method description

Lipovetsky starts with the analytical solution for a binomial logit with one dichotomous predictor x_{it} and the binary outcome y_{it}. Without loss of generality only one respondent is modeled, so compared to Equation (1) we leave out the n:

$$p_{it} = \frac{1}{1+e^{-V_{it}}} \qquad V_{it} = \alpha + \beta x_{it}. \tag{3}$$

Equation (3) can easily be transformed into its linear-link form:

$$\ln \frac{p_{it}}{1-p_{it}} = V_{it} = \alpha + \beta x_{it}. \tag{4}$$

The likelihood function of Equation (3) is normally used to estimate coefficients α and β:

$$L = \prod_{t=1}^{T} \prod_{i \in 0,1} p_{it}^{y_{it}} (1-p_{it})^{1-y_{it}} \tag{5}$$

Predictor X takes a value x_0 or x_1 in each t-th period, and the results of Equation (3) will also be of two values: p_0 and p_1. Results will be summarized in a (2×2) contingency table of counts of all combinations with a total of T counts (Table 1).

Table 1: Contingency 2×2 table of counts for predictor X and outcome Y with a total of T cases

		Binary event y=0	Binary event y=1	Row-totals
values	x_0	T_{00}	T_{01}	$T_{0.}$
	x_1	T_{10}	T_{11}	$T_{1.}$
Column-totals		$T_{.0}$	$T_{.1}$	T

Lipovetsky's approach is based on the observation that event frequencies – calculated on contingency tables of counts from the data – coincide with the estimated probabilities. His idea is to first calculate the probabilities and their log odds, and then calculate the coefficients analytically via a linear-link regression model.
To 'prove' this idea, he uses the counts of the contingency table in the binomial likelihood function (5), so he gets:

$$L = p_0^{T_{01}} (1-p_0)^{T_{00}} p_1^{T_{11}} (1-p_1)^{T_{10}} \tag{6}$$

with p_0 and p_1 being the probabilities in the points x_0 and x_1. To calculate the estimates, we have to partially derive Equation (6) and put the equations to zero. This leads to solutions

$$\hat{p_0} = \frac{T_{01}}{T_{00}+T_{01}}, \qquad \hat{p_1} = \frac{T_{11}}{T_{10}+T_{11}}. \tag{7}$$

The estimated values of p_0 and p_1 correspond to the empirical definition of probability when the relative frequency found by the counts can serve as the sampling probability estimator for a large sample size T.
To summarize the (2×2) case: The linear-link regression model leads to coefficients and therefore also to standard errors, t-statistics and R^2. Formulas for these can be found in Lipovetsky's paper (see Lipovetsky, 2014, pp.40-41). The (2×2) likelihood solution is already known (see Greene, 2012, p.797), but the detailed derivation of formulas makes it easy to expand the approach from binomial logit and one dichotomous predictor to:

- one categorical predictor (with K levels) resulting in a $(K \times 2)$ contingency table
- two categorical predictors (with K_1 and K_2 levels) resulting in a $((K_1 * K_2) \times 2)$ contingency table
- several categorical predictors (with $K_1, ..., K_n$ levels) with $M := (K_1 * ... * K_n)$ resulting in an $(M \times 2)$ contingency table
- multinomial logit can be transformed into binomial logit by transforming y_j into $y_j \in \{0,1\}$ with the additional levels being stacked, e.g. a three-level y_j with levels a, b and c will be transformed to y_a, y_b and y_c each with two levels $\in \{0,1\}$, stacked together
- continuous predictors have to be discretized into groups of discrete ordinal levels, the number of level has to be choosen depending on the dataset

2.2 Application on real data

What steps have to be executed when we apply the new method to real data? The first step is to find a reasonable contingency table of size $((M \ll N) \times 2)$. Of course M has to be much smaller than N, otherwise the contingency table would have about the same size of the original data, but with summarized information and with – obviously – a lot of missings. Cells with 1 or 0 have to be transformed to $(1-\varepsilon)$ or $(0+\varepsilon)$, using e.g. $\varepsilon = 10^{-5}$.
Note, that Lipovetsky himself (see Lipovetsky, 2014, p.44) tested a transformation of an $(N \times B)$ X database into an $(M \times B)$ $\tilde{X}$ database which I used for large categories like chocolate bars or softdrinks.
In a second step, the 'estimated' probabilities are calculated out of the contingency table and put in the linear-link function, as results at the left side of Equation (4). The third step is solving the linear-link regression by the standard $(X'X)^{-1}X'y$ formula (see Fahrmeir et al., 1996, p.98).
Why do we call the new method 'quasi'-analytical

solution? The 'quasi' comes into the solution in the first step, while defining the contingency table and by choosing an adequate ε. Steps two (calculation of probabilities) and three (solving a linear regression) are totally analytical.

Now I test whether the QAS can handle some typical characteristics of household panel data. There are two special cases where standard logit models show some difficulties: First, it could be that a household has only one choice option in its choice set. With the traditional objective function, these households will always get a probability of one ($p = 1$), independent of the estimated coefficient. Therefore, these households are removed because they bring no additional information into the objective function. But in small categorys where households do not buy that often or stick to one favourite choice option, these households could be a significant part of all households. In the contingency table these households bring in some new information and are therefore better represented by the QAS approach.
Second, in household panel modeling, no experimental design is created upfront as it is in conjoints, but the choice sets consist of real available choice options. Therefore it is possible that the choice design consists of several separated parts with no connection. In this case we had to fix one coefficient for each isolated island in order to optimize the remaining coefficients with numerical optimization (see Train, 2009, pp. 20-21). In the QAS approach these islands are already connected in the contingency table.
Up to this point, I demonstrated how to overcome some drawbacks of numerical optimization: the calculation of the results is no longer time consuming, we do not need initial coefficients and we will find the global maximum. But are (sign) restrictions and constraints still possible?

3 (Sign) restrictions

Why do we need (sign) restrictions? As we would like to do predictive analytics with our logit, we need face validity. E.g., for 'what if' scenarios in marketing mix modelling it is necessary that price gets a negative sign. Therefore we want to restrict the sign of some coefficients. Additionally, we have some utility values and want to guarantee that high loyalty utilities are greater than low loyalty utilities.
Through calculating probabilities from the contingency table and using them as input in the linear-link function we are in the environment of a linear regression and (in case of no restrictions) will just calculate the coefficients by the known $(X'X)^{-1}X'y$ formula. In case of restrictions we have to apply a trick. Again Lipovetsky, together with Conklin (2015), has published a paper where they compared different measures to analytically calculate preditors' relative importance as it is known from the Shapley value regression (SVR). They searched for a measure to overcome the computational burden of SVR. If both is needed – predictors' importance and regression coefficients – the measure of Gibson (1962) and R. Johnson (1966) (GJ) beats the other measures. So for the question at hand GJ is the one to take. The idea behind the GJ measure is the decomposition of R^2, using an orthonormal matrix approximation to the data. Furthermore, Lipovetsky and Conklin (2015) have improved the orthonormal approximation of GJ.β to GJ.β^* to become independent of singular values close to zero.
Lipovetsky and Conklin searched for a measure to calculate predictors' importance, they found one and additionally found that "the regression coefficients can also be adjusted to reach the best data fit and to be meaningful and interpretable" (see Lipovetsky and Conklin, 2015, p.1). Our solution has the additional benefit that the coefficients are also interpretable as predictors' importances and therefore are robust to multicollinearity.
For calculating GJ.β^* coefficients, the following 5 steps have to be executed:

I. calculate OLS.β:
$\text{OLS}.\beta = (X'X)^{-1}X'y$

II. calculate GJ.β which is the sqare root calculated by singular value decomposition. They have the meaning of pair correlations and simultaneously coefficients of regression:
$\text{GJ}.\beta = (X'X)^{1/2}\text{OLS}.\beta$

III. (optional) change signs in GJ.β, e.g.
$\text{GJ.sign} = (\text{GJ}.\beta_1, ..., abs(\text{GJ}.\beta_j), ..., \text{GJ}.\beta_n)$

IV. improvement by Lipovetsky and Conklin to get independent of singular values close to zero by q:
$$q = \frac{X'y\text{GJ.sign}}{\text{GJ.sign}'X'X\text{GJ.sign}}$$

V. result GJ.β^*
$\text{GJ}.\beta^* = q * \text{GJ.sign}$

As we can see from step III, the resulting coefficients are always positive. So we have to transform the input matrix to get any restrictions (coefficient $\gtrless$ value a, $\text{coefficient}_i \gtrless \text{coefficient}_j$).
In this section I demonstrated that QAS is able to handle (sign) restrictions by performing steps I. to V. instead of the standard regression formula. But how does the procedure change in case of constraints?

4 Constraints on the DV

A logit OLS objective function including constraints might look like this:

$$\begin{aligned} \text{minimize} \quad & \sum_{t=1}^{T}\sum_{n=1}^{N}(p_{nt} - y_{nt})^2 \\ \text{subject to} \quad & \sum_{n=1}^{N} p_{n_t} = \sum_{n=1}^{N} y_{n_t} \qquad t = (1, ..., T) \end{aligned} \tag{8}$$

Compared to the restricted coefficients from the section before, constraints could be handled as restrictions

on summary statistics of predicted values. Think of the following example: In modeling whether a household buys in one week, the weekly sum of purchase acts should be achieved as well. In logit models with numerical optimization the original OLS is augmented to:

$$\text{OLS} = \sum_{t=1}^{T}\sum_{n=1}^{N}(p_{nt} - y_{nt})^2 + \sum_{t=1}^{T}\left(\sum_{n=1}^{N}(p_{nt} - y_{nt})\right)^2 \quad (9)$$

In order to include this additional information into the QAS approach, we have to stack the constraint information with the original dataset and calculate the coefficients on the extended dataset. But how can we calculate the additional information? This is done mathematically.
We have to transform Equation (9) into a (linear-link) LS. Using Equation (4) and $y_{nt} := \ln \frac{p_{nt}}{1-p_{nt}}$ we get:

$$\text{(linear-link) LS} = \sum_{t=1}^{T}\sum_{n=1}^{N}(y_{nt} - \alpha - \beta x_{nt})^2 + \quad (10)$$

$$+\sum_{t=1}^{T}\left(\sum_{n=1}^{N}(y_{nt} - \alpha - \beta x_{nt})\right)^2 \quad (11)$$

The question now is how to transform Equation (11) to stack it with the data. Therefore the sum over all n has to be placed inside the brackets:

$$\sum_{t=1}^{T}\left(\frac{\sum_{n=1}^{N} y_{nt}}{N} - \alpha - \beta\frac{\sum_{n=1}^{N} x_{nt}}{N}\right)^2 \quad (12)$$

Now we see what we have to stack: From Equation (10) we see that we have $N * T$ lines with y_{nt} and x_{nt} within the dataset. From Equation (12) we have to add T lines with $\frac{\sum_{n=1}^{N} y_{nt}}{N}$ and $\frac{\sum_{n=1}^{N} x_{nt}}{N}$. This is the basis for calculating the QAS. In case of having constraints and (sign) restrictions we have to calculate GJ.β^* on the augmented dataset.
Why do we have these constraints on our logit OLS? In linear regression the following constraint: the total sum of y_{nt} equals the total sum of $\alpha + \beta x_{nt}$ because the mean of the dependent variable equals the mean of the values estimated by the model. But for non-linear logit this is not the case and therefore a more detailed constraint like Equation (11) is also not automatically solved.

In our household panel model we want to predict whether a household buys in time period t. In some categories, like detergents or cleaners, most household buy the category less frequently. If this is the case, we additionally have to modify the resulting coefficients like King and Zeng (see King and Zeng, 2001, p.144) suggested. They described a slightly modified Bayesian adjustment for the intercept:

$$\alpha^* = \alpha - \ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right] \quad (13)$$

In more frequently bought categories like chocolate bars or softdrinks this Bayesian adjustment leads to no further improvement, but it has no disadvantage either. The adjustment α^* in Formula (13) can be combined with data augmentation and GJ.β^* calculation.
We now have a complete theoretical solution for solving logit models analytically including (sign) restrictions and constraints. In the next section I apply this solution to an existing household panel model.

5 Application on a household panel model

The aim of the household panel model – named GfK-BrandSimulator® – is to predict the next year's purchase acts and volumes of a category in order to simulate different pricing and promotion strategies (Wildner and Scherübl, 2006). This is achieved by a mathematical model which consists of several sub models: two multinomial logits, one binomial logit and one truncated poisson model. The multinomial logits have no constraints, but sign restrictions and other restrictions. With these sub models the households are choosing brands and retailers for their purchases. They are labeled 'WHERE' for the choice of a retailer and 'WHAT' for the choice of the brand. The binomial logit works with sign restrictions and constraints. We use it for modeling whether the household buys in one week or not. It is labeled 'WHEN'. A combination of these three models is used to predict purchase acts.
The truncated poisson model, for which QAS is not applicable, is used ontop of purchase acts to predict purchase volume. Because QAS is not applicable to truncated poisson sub model, I concentrate on estimating purchase acts sub models and on predicting purchase acts for the comparison of the different optimizations.
So there are three models for testing the QAS. And additional to the two multinomial logits a combination of both into one multinomial logit can be tested. Up to now such a combined sub model was impossible to calculate due to the number of coefficients to be estimated. But in an analytical process, the number of coefficients does not matter that much. The new model combines the retailer and brand choice and will be labeled sub model 'WHERE & WHAT' ('W&W').
X-variables for the 'WHERE' and 'WHAT' models are mainly loyalty (towards retailers and brands respectively) and marketing mix information (on prices, promotions and distribution). An advantage of the combined model 'W&W' is, that it is able to include also loyalty (towards retailers and brands) and marketing mix. But as it is only one multinomial logit (compared to the product of two multinomial logits in case of 'WHERE' and 'WHAT') the marketing mix goes into the model only once. This is an enormous advantage.
X-variables for the 'WHEN' model are seasonal, inventory and consumption variables as well as marketing mix information (on prices, promotions and distribution) in-

cluding lagged coefficients for marketing mix.
The new methods are applied to three datasets from the German GfK household panel with the characteristics shown in Table 2.

Table 2: Characteristics of three datasets from the German GfK household panel

number of	**Fabric Softener**	**Chocolate Bars**	**Soft-drinks**
purchase acts	17,987	113,869	188,372
households	3,688	8,492	10,782
weeks	53	53	53
retailers	12	29	24
brands	11	31	61

All three datasets consist of one year of data (53 weeks) for the model estimation. To have data also for a validation, I selected households which have bought the category at least twice in two years. 'Fabric softener' is a rather small dataset with 11 brands and about 3,700 households buying the category almost 18,000 times. With a larger number of brands and a more than doubled number of households the categories 'chocolate bars' and 'softdrinks' cover over 100,000 purchase acts. In Table 3 and Table 4 the dimensions of the X matrices are listed.

Table 3: Dimensionality of the X matrices: number of rows

number of rows	**Fabric Softener**	**Chocolate Bars**	**Soft-drinks**
WHERE	47,779	373,704	863,285
WHAT	47,931	419,872	602,588
W&W	124,409	1,876,344	2,340,625
WHEN	195,464	450,076	571,446

In Table 3 the number of rows for the combined model really explodes. Looking at the number of rows for the 'W&W' model with over 2 million rows we come in the range of big data.

Table 4: Dimensionality of the X matrices: number of columns

number of columns	**Fabric Softener**	**Chocolate Bars**	**Soft-drinks**
WHERE	52	120	307
WHAT	57	157	100
W&W	107	274	404
WHEN	13	13	11

The number of columns correspond to the number of estimated coefficients. In Table 4 the number of coefficients for 'W&W' is almost the sum of the coefficients for 'WHERE' and 'WHAT'. The number of coefficients for the multinomial logits are so large due to the number of loyalty coefficients.
The model results are compared with the following measures:

- $R^2_{\text{Efron}} = 1 - \dfrac{\sum (y_n - \hat{y_n})^2}{\sum (y_n - \bar{y})^2}$
  (Efron, 1978)
- mean absolute error: $\text{MAE} = \dfrac{1}{N} \sum_{n=1}^{N} |y_n - \hat{y_n}|$
  (see Fair, 1984, p.261)
- average probability for hits:
  $\text{av.hit.prob} = \dfrac{\sum y_n * \hat{y_n}}{\sum y_n}$
  (see Rossi and Allenby, 1993, p.180)
- estimation time

The measures focus on different aspects. In R^2 differences between observed and predicted probability are squared. MAE works directly with the differences and av.hit.prob bases upon predicted probabilites for the alternatives chosen. All measures have in common the range $[0; 1]$. So they are easily interpretable and comparable.
R^2 and MAE will be calculated on the household level database (called: individual R^2/ MAE) as well as on a more aggregated database – summed over households – depending on the sub model (called: aggregated R^2/ MAE). The dimensions of the database are displayed in Tables 5 and 6. The measure av.hit.prob focuses on chosen alternatives. With aggregated purchase acts (sum or mean) this feature would be lost. So we will judge the quality of the results based on five measures (individual and aggregated R^2, individual and aggregated MAE, av.hit.prob). A sixth measure estimation time will complement the evaluation.

Now we turn to model fit and validation. Let us first look at the model fit. In Table 5 the columns called numerical optimization (num.) refer to the current GfK household panel model GfK-BrandSimulator®. Because there is no equivalent numerical model for the combined QAS model ('W&W'), we combine – for comparison – the results of the single numerical sub models (product for R^2, MAE and av.hit.prob, sum for time).
Table 5 summarizes the results on model fit and shows several patterns:

- In all cases (all models, all datasets) the R^2 is higher for the numerical optimization. This is true for individual and aggregated R^2. The difference is quite obvious. We find the biggest difference for the 'WHEN' sub model.
- Individual MAE is very similar for both methods. Only for the 'WHEN' model QAS is always better then numerical optimization. But again on the aggregated level, the numerical model always beats QAS.

Table 5: Model fit for numerical optimization (num.) vs. QAS

			datasets					
			Fabric Softener		**Chocolate Bars**		**Soft-drinks**	
models	criteria	dim. of database[□]	num.	QAS	num.	QAS	num.	QAS
WHERE	R^2	h,w,r	0.325	0.294	0.242	0.236	0.311	0.305
		w,r	0.976	0.967	0.991	0.988	0.992	0.992
	MAE	h,w,r	0.316	0.304	0.253	0.250	0.270	0.276
		w,r	0.048	0.055	0.021	0.023	0.018	0.018
	av.hit.prob	h,w,r	0.580	0.596	0.410	0.418	0.502	0.492
	estimation time[▽]	–	0.400	0.008	7.100	0.235	7.483	0.267
WHAT	R^2	h,w,r,b	0.434	0.321	0.211	0.167	0.348	0.223
		w,r,b	0.915	0.846	0.950	0.829	0.966	0.809
	MAE	h,w,r,b	0.262	0.263	0.264	0.262	0.223	0.245
		w,r,b	0.109	0.140	0.060	0.087	0.066	0.111
	av.hit.prob	h,w,r,b	0.656	0.656	0.405	0.410	0.493	0.441
	estimation time[▽]	–	1.417	0.007	26.100	0.406	888.767	2.174
W&W	R^2	h,w,r,b	0.247	0.194	0.056	0.035	0.121	0.105
		w,r,b	0.867	0.802	0.866	0.842	0.897	0.781
	MAE	h,w,r,b	0.181	0.178	0.093	0.093	0.107	0.110
		w,r,b	0.057	0.066	0.024	0.026	0.035	0.042
	av.hit.prob	h,w,r,b	0.379	0.388	0.148	0.151	0.227	0.212
	estimation time[▽]	–	1.817	0.053	33.200	1.315	896.250	2.811
WHEN	R^2	h,w	0.064	0.021	0.130	0.102	0.256	0.233
		w	0.549	0.000	0.817	0.164	0.353	0.107
	MAE	h,w	0.150	0.140	0.273	0.254	0.278	0.250
		w	0.008	0.017	0.011	0.026	0.013	0.016
	av.hit.prob	h,w	0.862	0.888	0.783	0.833	0.868	0.955
	estimation time[▽]	–	10.950	0.009	53.750	0.022	32.117	0.022

[□] h: household; w: week; r: retailer; b: brand
[▽] estimation time measured in minutes

- On the other hand, in almost all cases, besides Softdrinks 'WHERE', 'WHAT' and 'W&W', the av.hit.prob favors QAS.

- And last but not least the estimation time is always lower for QAS. And the larger the dataset, the more substantial is the improvement in calculation time.

Summing up the results, we observe that for the individual measures we get about the same quality for both methods. Although R^2 is less for the QAS, MAE and av.hit.prob show that the quality of the new method is not worse than the numerical optimization – and the coefficients are estimated in much less time. But we have to keep in mind that on the aggregated level QAS does not reach the same quality, especially for the 'WHEN' sub model. But this was only model fit. We also have to do a validation.
For the validation model, coefficients are estimated with the input (household panel data) of year T, the model results are predicted for year $T+1$ (using model coefficients of year T and marketing mix of year $T+1$) and the model results are compared with real household panel data for year $T+1$. To predict the model results – individual purchase acts – the three household panel sub models 'WHEN', 'WHERE' and 'WHAT' are multiplied (W,W,W). In the case of the combined model 'W&W', sub models 'WHEN' and 'W&W' have to be multiplied (W&W,W). For the validitation calculation, the sub models are only applied, but not estimated. For this reason, the comparison of time for the different methods is not meaningful. As we have to calculate all combinations of households, weeks, retailers and brands, the matrices are much larger than for model estimation. Therefore the measures are in general lower than for model fit. Table 6 reports the results of the validation.

The individual R^2s are too small to interprete. Aggregated R^2s favor clearly the numerical optimization. On the other hand, individual MAE is always better for QAS and the aggregated measure is quite similar for both methods. Average probability for hits differs: It is best for fabric softener numerical optimization, best for chocolate bars $\text{QAS}_{W,W,W}$ and best for softdrinks $\text{QAS}_{W\&W,W}$. So each model and each dataset leads to

Table 6: Validation results for numerical optimization (num. opt.) vs. QAS

datasets	criteria	dim. of database[a]	models num. opt.	models QAS (W,W,W)	models QAS (W&W,W)
Fabric Softener	R^2	h,w,r,b	0.013	0.000	0.000
		w,r,b	0.646	0.390	0.403
	MAE	h,w,r,b	0.023	0.016	0.015
		w,r,b	0.008	0.007	0.007
	av.hit.prob	h,w,r,b	0.044	0.028	0.025
Chocolate Bars	R^2	h,w,r,b	0.003	0.000	0.000
		w,r,b	0.831	0.474	0.000
	MAE	h,w,r,b	0.022	0.020	0.017
		w,r,b	0.006	0.007	0.008
	av.hit.prob	h,w,r,b	0.053	0.122	0.096
Soft-drinks	R^2	h,w,r,b	0.048	0.000	0.000
		w,r,b	0.919	0.788	0.588
	MAE	h,w,r,b	0.044	0.042	0.038
		w,r,b	0.012	0.015	0.015
	av.hit.prob	h,w,r,b	0.140	0.199	0.207

[a] h: household; w: week; r: retailer; b: brand

a different solution. One may wonder why the probability is so low, but for each household, week, retailer and brand, there are a lot of options all together. The probability to beat (of the zero-model) is 0.023 for fabric softener, 0.017 for chocolate bars and 0.044 for softdrinks. Regarding also the amount of difference between the single solutions, av.hit.prob favors $\text{QAS}_{W,W,W}$.

Again – as for model fit – the summary of all measures shows comparable quality on the individual level. But on the aggregated level, the level where decisions are made in practice, QAS could not reach the level of numerical optimization.

Why do individual comparable results differ so much on an aggregated level? One reason could be: The R^2s, the strongest of the selected measures as it judges all differences in square, always led to numerical optimization. So far there is no contradiction. Especially av.hit.prob often favored QAS. But this measure only regards chosen alternatives, therefore the non-chosen ones could make the difference in aggregation.

6 Summary

The starting point was that numerical optimization for logit models is time consuming, dependent on initial values and unfortunately there is no guarantee for finding the global maximum. With Lipovetsky's quasi-analytical method we managed to save estimation time substantially. But as in numerical optimization (sign) restrictions and constraints are easy to include, this is not the standard in linear regression. (Sign) restrictions are solved through calculation of Gibson's and Johnson's measure, which is additionally optimized by Lipovetsky and Conklin (2015). To overcome the constraints we stack additional information with the original dataset and calculate GJ.β^* coefficients on the augmented data. To model household panel data for small categories, we additionally use a slightly modified Bayesian adjustment for the intercept suggested by King and Zeng (2001).

Finally we applied the combination of these methods to three household panel datasets: One small detergents sub category (fabric softener) and two larger categories from food (chocolate bars) and drinks (softdrinks). We compared the model fit on several dimensions and found out that the R^2 was always better for numerical optimization. The individual MAE and the average probability for hits was mostly better for the new methods. In summary the quality of the coefficients and the individual results seem to be equal for the numerical and for the quasi-analytical solution. The aggregated results favored the numerical method. But of course the QAS is much faster to calculate. This is a big advantage, especially for the large categories.

Additionally, we tested a new sub model for the existing household panel model. Instead of two distinct models for retailer choice and brand choice we tried to model both choices in one model. From a model fit perspective this worked quite well, although the number of coefficients to estimate was almost the sum of the coefficients of the two single sub models.

We also compared the validation results for the numerical and the QAS coefficients. Although again R^2 was always better for numerical optimization, individual MAE and average probability for hits favor QAS. In practice, decisions are made on an aggregated level, this led to numerical optimization as winner for all three datasets.

The strengths of the QAS approach are first of all

speed. This has enormous effect on research time and of course also for clients projects. The second strength is the capability with multicollinearity because of the GJ.β^* coefficients. And third, for some patterns in real data the aggregation to a contingency table really has advantages. The weakness of QAS is so far the quality of results in a complex model as GfK-BrandSimulator®. This leads to some research questions: How can we further improve which contingency table to select as this is the only non-analytical step in the whole procedure? Are there already patterns visible when we compare resulting coefficients directly? How can we visualize aggregated results to identify quality gaps. Also, the decision on whether to stick to the two seperate sub models for retailer and brand or to use the combined sub model in the future cannot be answered based on the existing results. A further research direction: Is there an option to apply QAS on truncated poisson models? Then the whole household panel model could be solved quasi-analytically.

Further applications of QAS are possible in a wide field of market research methods, especially preference modelling, forecasting and classification tasks.

References

Efron, B. (1978). Regression and ANOVA with Zero-One Data: Measures of Residual Variation. *Journal of the American Statistical Association 73(361), 113-121.*

Fahrmeir, L., Hamerle, A., and Tutz, G. (1996). *Multivariate statistische Verfahren, 2. Auflage.* DeGruyter.

Fair, R. C. (1984). *Specification, estimation, and analysis of macroeconometric models.* Harvard University Press.

Gibson, W. A. (1962). Orthogonal predictors: A possible resolution of the Hoffman-Ward controversy. *Psychol. Rep. 11, 32-34.*

Greene, W. H. (2012). *Econometric Analysis, Seventh Edition.* Upper Saddle River, NJ.

Johnson, R. M. (1966). The minimal transformation to orthonormality. *Psychometrika 31, 61-66.*

King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis 9(2), 137-163.*

Lipovetsky, S. (2014). Analytical closed-form solution for binary logit regression by categorical predictors. *Journal of Applied Statistics 42(1), 37-49.*

Lipovetsky, S. and Conklin, W. M. (2015). Predictor relative importance and matching regression parameters. *Journal of Applied Statistics 42(5), 1-15.*

Rossi, P. E. and Allenby, G. M. (1993). A Bayesian approach to estimating household parameters. *Journal of Marketing Research, 171-182.*

Train, K. E. (2009). *Discrete choice methods with simulation, Second Edition.* Cambridge university press.

Wildner, R. and Scherübl, B. (2006). Model-Assisted Analysis, Simulation and Forecasting with Consumer Panel Data. *Yearbook of Marketing & Consumer Research 2006(4), 5-29.*