# Analysis of a Contingency Table Using MDI Estimation in Excel 3.0

***Fred Phillips***

## Overview

These notes describe tests of three hypotheses (“switching proportional to brand share”; “switching proportional to share of non-repeat volume”; and “Mercedes, Renault and Peugeot in the same segment”) using the 198D data on switching among automobile buyers in France. Also described is the fitting of a “gravity” or log-linear model to these data. A point of interest is that these hypotheses, some of which involve minimizing a nonlinear function of many variables, with constraints, was performed on a standard spreadsheet program, Microsoft Excel 3.0 for the Macintosh, using its built-in “Solver” facility.

According to the doctrine expounded by the Hendry Corporation (see Kalwani, 1980; see also Charnes, Cooper, Learner and Phillips, 1986), “Product alternatives are in direct competition if the switching to (and between) them is in proportion to their shares.” Symbolically, the switching $p_{ij}$ between two such brands is

$$p_{ij} = KS_iS_j. \tag{1}$$

This is a “clustering” concept. It differs from the premise behind e.g. perceptual maps, which determines perceived “distances” between brands. Under the latter premise, the K in equation (1) is replaced by the perceived proximity, or inverse distance in perceptual space, to yield a “gravity” model (see Huff, 1964; Phillips, 1991)

$$p_{ij} = (D_{ij})^{-1}u_iv_i. \tag{2}$$

In equation (2), the “potentials” $u_i$ and $v_j$ are not identical to the shares $S_i$ and $S_j$. In light of equation (2), equation (1) implies that brands “in direct competition” are indistinguishable, or at any rate a matter of indifference, to the consumer. Some applications require knowing the relationship between $p_{ij}$ and $D_{ij}$. Phillips, White and Haynes (1976) detail a procedure for computing this relationship. Other applications, particularly when $D_{ij}$ are not known, require clustering the brands into subsets for which equation (1) holds, up to the limits of sampling confidence, with a distinct value of K for each of the subsets. See Charnes, Cooper, Learner and Phillips (1986) for an information-theoretic algorithm for such a clustering.

## Hypotheses to be Tested

If the Hendry hypothesis is correct for the entire car switching matrix, i.e. if the set of brands is homogeneous, the only possible value for K that provides accounting balance (preserves shares) is 1/T = $(S_jS_j)$-1. This is known from contingency table theory, and algebraically astute readers will understand it is immaterial whether we work with the “units sold” data directly or whether we reduce these to shares of the total T. The first hypothesis is thus

$$p_{ij} = \sum_j S_j / T \tag{3}$$

Often, repeat buying is greater than equation (3) would imply. Furthermore, in brand *shifting* studies[1] (as opposed to *switching* analyses), repeat buying is directly observable where shifting

[1] Shifting Analysis is concerned with adjacent periods of time, during each of which several purchases may occur. Switching analysis addresses successive purchase occasions. See Phillips, 1991, for a complete exposition of the differences between these analytic approaches.

between brands is not. So other models (including SANDDABS - see Phillips, 1991) remove the observed repeat buying from the analysis and presume that in a directly competing set, shifting is proportional to share of non-repeat volume. Let $s_j$ denote the non-repeat volume for brand j. Then we have

$$s_j = S_j - p_{jj} \tag{4}$$

and

$$p_{ij} = Ks_i s_j \tag{5}$$

In the case of a product such as an automobile, with a long interpurchase cycle, differences between switching and shifting approaches are immaterial. We may use the car matrix for either kind of analysis.

The hypothesized quantities (3) and  can be computed from the data matrix. The hypothesis test can be based on Pearson's chi-square, Kullback entropy, or other such measures. Results appear below.

The third hypothesis to be entertained here, that the data can be represented by a gravity equation of type (2), must be computed by nonlinear programming. As no perceptual distance or "cost of switching" data are available for the present exercise, this computation is presented only as a test of the performance of "Solver".

Inspection of the data suggests that Mercedes, Renault and Peugeot (and, perhaps, others) inhabit the same market segment. A final computation, highlighting different features of "Solver," tests the hypothesis that these three brands share a switching constant.

**Hypothesis 1** The observed data matrix appears in the appendix. We will call the observations $q_{ij}$. Table 1 reflects the hypothesized table, based on equation (3). Note that the Hendry theory does not claim that the repeat volume must also satisfy equation (3). However, if it does not, the estimated table does not preserve the known brand shares. The $p_{ij}$ entries in Table 1 are therefore calculated according to equation (3). The Kullback test statistic, $2\Sigma\Sigma p_{ij}\ln(p_{ij}/q_{ij})$, is 17197.3. The Pearson chi-square is 61014.39. These statistics are distributed as $\chi^2$ with $(15)^2$-1=224 degrees of freedom, leading in this case to the use of the approximately standard normal variate $\surd(\chi^2) - \surd(2(224)\text{-}1)$ as a test statistic. This figure has a value of 109 for the Kullback statistic, indicating rejection of the hypothesis represented by equation (1). Given that the Pearson statistic is often called a quadratic approximation to the Kullback statistic, it is interesting to see the extent to which they diverge when differences between observed and estimated values are extreme.

**Table 1: Switching proportional to market share**

| | Alfa | BMW | Cit | Fiat | Ford | GM | Lada | Mer | Peu | Ren | Rov | Saab | Seat | VW | Volv |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Alfa | 2 | 4 | 32 | 16 | 18 | 13 | 2 | 3 | 58 | 80 | 4 | 0 | 4 | 20 | 2 |
| BMW | 3 | 5 | 45 | 23 | 25 | 18 | 3 | 4 | 81 | 111 | 6 | 0 | 5 | 28 | 3 |
| Citroen | 26 | 46 | 405 | 204 | 231 | 161 | 30 | 39 | 735 | 1006 | 54 | 4 | 49 | 251 | 25 |
| Fiat | 9 | 16 | 141 | 71 | 81 | 56 | 10 | 14 | 256 | 351 | 19 | 2 | 17 | 87 | 9 |
| Ford | 10 | 18 | 158 | 80 | 90 | 63 | 12 | 15 | 288 | 393 | 21 | 2 | 19 | 98 | 10 |
| GM | 5 | 10 | 85 | 43 | 48 | 34 | 6 | 8 | 154 | 211 | 11 | 1 | 10 | 53 | 5 |
| Lada | 1 | 2 | 21 | 11 | 12 | 8 | 2 | 2 | 38 | 53 | 3 | 0 | 3 | 13 | 1 |
| Mercedes | 2 | 3 | 24 | 12 | 14 | 10 | 2 | 2 | 44 | 60 | 3 | 0 | 3 | 15 | 2 |
| Peugeot | 39 | 69 | 605 | 305 | 345 | 241 | 44 | 58 | 1098 | 1501 | 80 | 6 | 74 | 375 | 38 |
| Renault | 60 | 108 | 950 | 480 | 542 | 378 | 69 | 91 | 1724 | 2358 | 126 | 10 | 116 | 588 | 60 |
| Rover | 3 | 5 | 41 | 20 | 23 | 16 | 3 | 4 | 74 | 101 | 5 | 0 | 5 | 25 | 3 |
| Saab | 0 | 0 | 2 | 1 | 1 | 1 | 0 | 0 | 4 | 6 | 0 | 0 | 0 | 1 | 0 |
| Seat | 1 | 2 | 17 | 9 | 10 | 7 | 1 | 2 | 31 | 42 | 2 | 0 | 2 | 11 | 1 |
| VW | 11 | 19 | 171 | 86 | 97 | 68 | 12 | 16 | 310 | 424 | 23 | 2 | 21 | 106 | 11 |
| Volvo | 1 | 2 | 20 | 10 | 12 | 8 | 1 | 2 | 37 | 50 | 3 | 0 | 2 | 13 | 1 |

**Hypothesis 2** Table 2 shows the switching volumes only, estimated under the hypothesis of equation (5), i.e. "Switching proportional to share of non-repeat volumes."

**Table 2: Switching proportional to share of non-repeat volume**

| | Alfa | BMW | Cit | Fiat | Ford | GM | Lada | Mer | Peu | Ren | Rov | Saab | Seat | VW | Volv |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Alfa | | 2.6 | 15.9 | 14.8 | 14.9 | 12.6 | 2.3 | 2.2 | 35.1 | 33.0 | 4.3 | 0.3 | 5.2 | 15.9 | 1.6 |
| BMW | 1.6 | | 19.3 | 18.0 | 18.2 | 15.3 | 2.8 | 2.6 | 42.7 | 40.1 | 5.2 | 0.4 | 6.3 | 19.4 | 2.0 |
| Citroen | 11.9 | 23.1 | | 133.0 | 134.2 | 113.2 | 20.4 | 19.5 | 315.0 | 296.5 | 38.5 | 3.0 | 46.4 | 143.2 | 14.6 |
| Fiat | 5.0 | 9.7 | 60.0 | | 56.6 | 47.7 | 8.6 | 8.2 | 132.7 | 124.9 | 16.2 | 1.3 | 19.5 | 60.3 | 6.2 |
| Ford | 4.8 | 9.2 | 57.0 | 53.2 | | 45.3 | 8.2 | 7.8 | 126.0 | 118.6 | 15.4 | 1.2 | 18.5 | 57.3 | 5.8 |
| GM | 2.7 | 5.1 | 31.6 | 29.5 | 29.8 | | 4.5 | 4.3 | 69.9 | 65.8 | 8.5 | 0.7 | 10.3 | 31.8 | 3.2 |
| Lada | 0.8 | 1.6 | 10.1 | 9.4 | 9.5 | 8.0 | | 1.4 | 22.3 | 21.0 | 2.7 | 0.2 | 3.3 | 10.1 | 1.0 |
| Mercedes | 0.5 | 0.9 | 5.8 | 5.4 | 5.4 | 4.6 | 0.8 | | 12.8 | 12.0 | 1.6 | 0.1 | 1.9 | 5.8 | 0.6 |
| Peugeot | 16.0 | 31.0 | 190.8 | 178.1 | 179.8 | 151.6 | 27.4 | 26.1 | | 397.1 | 51.6 | 4.0 | 62.1 | 191.8 | 19.6 |
| Renault | 23.0 | 44.5 | 274.0 | 255.8 | 258.2 | 217.7 | 39.3 | 37.5 | 606.0 | | 74.1 | 5.7 | 89.2 | 275.5 | 28.1 |
| Rover | 1.7 | 3.4 | 20.8 | 19.4 | 19.6 | 16.5 | 3.0 | 2.8 | 45.9 | 43.2 | | 0.4 | 6.8 | 20.9 | 2.1 |
| Saab | 0.1 | 0.1 | 0.9 | 0.8 | 0.8 | 0.7 | 0.1 | 0.1 | 1.9 | 1.8 | 0.2 | | 0.3 | 0.9 | 0.1 |
| Seat | 0.8 | 1.6 | 9.9 | 9.2 | 9.3 | 7.9 | 1.4 | 1.4 | 21.9 | 20.6 | 2.7 | 0.2 | | 9.9 | 1.0 |
| VW | 5.0 | 9.6 | 59.3 | 55.4 | 55.9 | 47.1 | 8.5 | 8.1 | 131.2 | 123.5 | 16.0 | 1.2 | 19.3 | | 6.1 |
| Volvo | 0.7 | 1.4 | 8.4 | 7.9 | 7.9 | 6.7 | 1.2 | 1.2 | 18.6 | 17.5 | 2.3 | 0.2 | 2.7 | 8.5 | |

The Kullback statistic is 1479.146 and the Pearson chi-square is 1603.466. Either of these yields a test statistic of about 19, unlikely to be realised from a normal variate with unit variance. The test of hypotheses 1 and 2 show the usually expected relationships not to arise from the dataset as a whole, and therefore the market must be segmented.

**Hypothesis 3** The fitted gravity model is identical to the first decimal place with the raw data. It was computed using the default options in "Solver" in several hundred iterations in an elapsed time of about 15 minutes on a Macintosh II. The mathematics were as follows. Charnes, Haynes, Phillips and White (1977) showed that maximizing the negative of the Kullback information number subject to the preservation of marginal totals (shares) $O_r$ and $D_s$,

$$\max \{-\sum\sum x_{rs} \ln(x_{rs}/ef_{rs})\} \ s.t.\ \sum_s x_{rs} = O_r;\quad \sum_r x_{rs} = D_s;\quad and\ all\ x_{rs} \geq 0 \tag{6}$$

is equivalent to minimizing this problem's extended geometric programming dual, namely

$$\min \{\sum\sum f_{rs} e^{u_r + v_s} - \sum_r O_r u_r - \sum_s D_s v_s\} \tag{7}$$

Optimal primal and dual variables are related by the gravity equation

$$x_{rs} = \sigma_r \tau_s \ / \ (f_{rs})^{-1} \tag{8}$$

where $\sigma_r = e^{u_r}$ and $\tau_s = e^{v_s}$. The dual (7) was solved on the spreadsheet optimizer (with, of course, $q_{ij} = ef_{rs}$), providing convenience in the input (no constraints had to be specified for the Solver), and better speed of solution than would be the case with a constrained problem, as we will see below. This test showed that the Solver can resolve the dual of an MDI problem of this size with some efficiency, accuracy and convenience. Of course, by reproducing the input data as a gravity equation, we have revealed nothing about the automobile market; such a computation with e.g. perceptual map data would be, on the contrary, a revealing summary of the data. As Brockett, Charnes and Cooper (1978) have shown that functionals of the type (6) *with arbitrary constraints* also have unconstrained convex duals, the above procedure offers considerable utility for statisticians.

The nonlinear programming capability in Solver is based on Lasdon's GRG code. Having linear and nonlinear constrained optimization available in a PC spreadsheet environment is a stunning advance. That Solver is included with every copy of Microsoft Excel 3.0 (millions have been sold) cannot but have an impact on the teaching and practice of statistics and management science. The fact that calculation is slow compared to dedicated mathematical programming or equation processing programs is offset by the convenience and accessibility of setting up a problem using standard spreadsheet rules, and being able to analyze the results further on the spreadsheet without importing and exporting data.

**Hypothesis 4** Here we proceed to segmentation directly from equation (1), without the detailed Hendry characterization of the constant K that underlies e.g. the analysis in Charnes, Cooper, Phillips and Learner (1984). The analysis is thus more ad hoc, but still illustrative. If $p_{ij} = KS_iS_j$ , then saying that brands i and j share the same switching constant (are in the same segment) is rather like saying that two points chosen at random in Euclidean space are collinear. Equation (1) is hardly a basis for saying that i and j compete directly. It is only when a third brand k enters the picture in such a manner that $p_{jk} = KS_jS_k$ holds with the same value of K that such inferences are meaningful. Substituting from these two equations gives us

$$p_{ij} / p_{jk} = S_i / S_k \ \ and \ \ p_{jk} / p_{ki} = S_j / S_i \tag{9}$$

if i, j and k are part of a homogeneous segment. Thus, in order to form an illustrative hypothesis about segments within the data of Table 1, brand triples were examined for values of the difference

$$\left| p_{ij}/ p_{jk} - S_i/S_k \right| \tag{10}$$

The value of (10) for the triple "Mercedes, Peugeot and Renault" was 0.00065. This was the minimum value for all triples examined. The values ranged up to 6,000 for other brand triples.

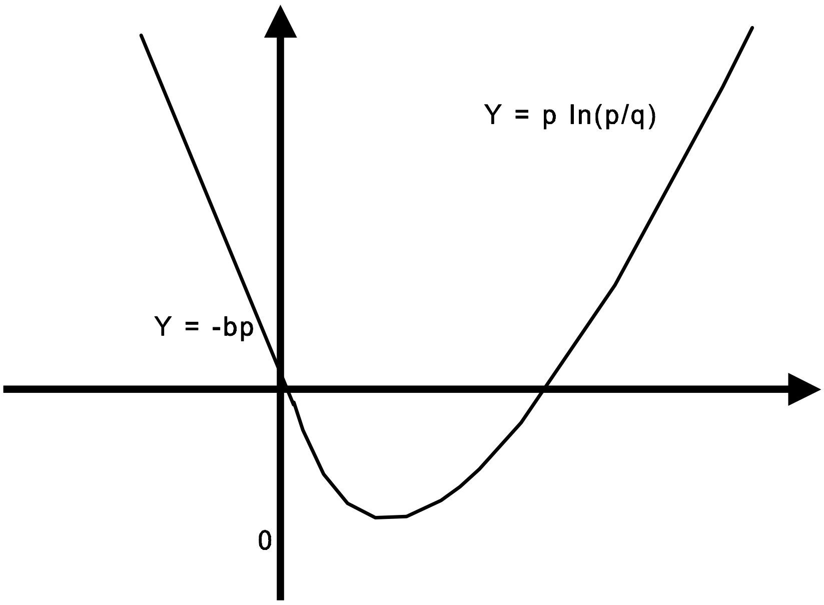
Equations (11) represent a nonlinear minimization that tests the hypothesis "Mercedes, Peugeot and Renault inhabit the same market segment."

$$\begin{aligned} &\text{Minimize} \sum\sum p_{ij} \ln (p_{ij} / q_{ij}) \text{ subject to} \\ &p_{\text{Peu,Merc}} = (S_{\text{Peu}}/S_{\text{Ren}})p_{\text{Merc,Ren}} \\ &p_{\text{Merc,Peu}} = (S_{\text{Peu}}/S_{\text{Ren}})\, p_{\text{Ren,Merc}} \\ &p_{\text{Peu,Ren}} = (S_{\text{Ren}}/S_{\text{Merc}})\, p_{\text{Peu,Merc}} \end{aligned} \tag{11}$$

with all $p_{ij} \geq 0$ and $\Sigma\Sigma p_{ij} = T$.

The test statistic, twice the minimum value of the objective function (11), is distributed $\chi^2$ with three degrees of freedom under the null hypothesis (See Charnes, Cooper, Phillips, and Learner, 1984). Its value of 305.61 indicates the hypothesis should be rejected at any reasonable significance level. This is somewhat surprising, but might be explained in terms of the use of market shares S in (11) rather than shares of non-repeat volume s; the fact that not all triples of the form (10) were examined prior to choosing Mercedes, Peugeot and Renault as likely candidates; and/or the imperfect symmetry of the matrix of observations. The optimization (11) was solved directly, although it too has an unconstrained convex dual. Solving on the primal side brought to light some interesting features of the Solver. First, the algorithm does not confine its search to the feasible region. Specifically, it will move to regions in which one or more $p_{ij} \leq 0$. Upon attempting to compute $\ln(p_{ij})$ in order to evaluate the objective function, the procedure will halt and issue an error message. This undocumented tendency may be dealt with by placing an Excel formula in the objective function cell that says, in effect, "If $p_{ij} > 0$, then $\ln(p_{ij})$; else, $-bp_{ij}$" Figure 1 illustrates the effect of this strategy, which will drive the search back into the feasible region. If b is chosen very large, it may be unnecessary to specify explicitly that all $p_{ij}$ must be nonnegative. (In fact, the specification of constraints in Solver is less convenient than in earlier spreadsheet optimizers like “BestAnswer.”) Of course, any defined, decreasing function of $p_{ij}$ may be used to the left of the origin.[2]

**Figure 1: Circumventing a bug in solver**



Attempting to solve (11) while varying the entire p matrix resulted in an error message, “too many adjustables.” The solver manual does not specify a maximum number of variables, nor even a maximum sum of variables and constraints. The maximum number may be presumed to depend on the number of constraints, the degree of nonlinearity of the functional and the constraints, and (given the intrusiveness of Excel 3.0 into the operating system files) the available memory in the PC. Luckily, because of symmetry, it was possible to solve (11) by varying only the upper diagonal of the p matrix, and in doing so no error messages appeared.

## Conclusion

Although its size limitations are bound to be irksome when run on computers with limited memory (Excel 3.0 itself is a notorious memory eater), Solver offers enormous flexibility for examining statistical problems in marketing. The spreadsheet environment supports data preparation and cleaning, and calculations of all relevant functions of outputs (e.g. reverse transformations, when the original data have been transformed prior to analysis) simultaneously with the statistical analysis. Further, by enabling the analyst to specify the form of minimand functions at will, statistical analysis can become simpler and yield better insights. For example, a likelihood function can be maximized directly if maximizing likelihood is deemed to be of primary importance; there is no need to use a standard analysis (or prove that a surrogate analysis yields a max-likelihood estimate) and then compute likelihood as a post-analysis. These notes have offered a few examples of how this flexibility may be exploited to analyze an automobile brand switching data set. They report some tests of the Solver and bring to light some undocumented features of the program, while exploring how some of the marketing literature’s

[2] Professor Lasdon has informed me that a new version is in preparation that will search only in the feasible region.

most prevalent concepts of brand switching and shifting may be translated into hypotheses that can be tested using the first widely available and user-friendly nonlinear programming system.