# Customer satisfaction and PLS structural equation modeling. An application to automobile market

Valentina Stan and Gilbert Saporta

Conservatoire National des Arts et Métiers
292 Rue Saint Martin
F 75141 Paris Cedex 03
(e-mail: valentina_titu@yahoo.fr, saporta@cnam.fr)

**Abstract.** We present the principal concepts of structural equation modeling and a comparison between the two main approaches: PLS (Partial Least Square) and LISREL (Linear Structural Relationship). A structural model uses 2 types of models: the measurement model (outer model) and the structural model (inner model). An application to real life data on customer satisfaction is given.
**Keywords:** Structural equation modeling, Partial least square, PLS approach.

## 1 An introduction to structural equation modeling

### 1.1 General considerations

Let $p$ variables be observed upon $n$ units. The $p$ variables are partitioned in $J$ subsets or blocks of $k_j$ variables which are presumed to be pertinent for describing the phenomenon. Each of these blocks is designed to describe a theme of the general phenomenon. We shall designate these blocs by $X_j$ and we shall consider them as matrices with dimension $(n \times k_j)$. In structural models the observed variables are called manifest variables. The latent variables are not observable: they exist by the relations they have with the manifest variables. In the following we shall always suppose that each block is associated with only one latent variable (unidimensionality). Therefore we can identify the blocks by the same name as their latent variable. The latent variable corresponding to the $X_j$ block will be designated by $\xi_j$. A structural model needs 2 types of models: the measurement model (outer model) which connects the manifest variables to the latent variables and the structural model (inner model) which connects latent variables between them.

**1.1.1 The measurement model (outer model)** After having determined the blocks, we must specify the type of relationship between latent variables and manifest variables which correspond to block $X_j$. There are 3 ways: the reflective way, the formative way, the MIMIC way (Multiple effect Indicators for Multiple Causes).

**The reflective way** In this way, the manifest variables are considered like the "reflection" of their latent variables [Tenenhaus *et al.*, 2005]. This kind of situation exists for instance in models which analyse customer satisfaction of a particular kind of service: a set of questions about the image of the service which represents a latent variable in the model. Each manifest variable is related to its latent variable, as follows:

$$x_{jh} = \pi^0_{jh} + \pi_{jh}\xi_j + \epsilon_{jh} \quad \forall h = 1 \dots k_j$$

$\pi^0_{jh}$ = constant term; $\pi_{jh}$ = regression coefficient; $\epsilon_{jh}$ = residual term.

**The formative way** Here the latent variables represents the "reflection" of the manifest variables which belong to block $X_j$, and are thus a result of these [Tenenhaus and al., 2005]. In this type, the latent variable is a linear function of the manifest variables which generate it:

$$\xi_j = \Sigma^{k_j}_{h=1}\varpi_{jh}x_{jh+\delta_j}$$

;

$\varpi_{jh} \quad (h = 1 \dots h_j)$= multiple regression coefficients of $\xi_j$ on ; $\delta_j$ = residual term.

#### 1.1.2 The structural model (inner model)

Opposite to the measurement model, which deals with the relations between latent variables and their manifest, the structural model concerns the mode of estimation of latent variable between them. The relations between latent variables have the form:

$$\xi_j = \beta^0_j + \Sigma^J_{i=1,i\neq j}\beta_{ji}\xi_i + \zeta_j \quad \forall j = 1 \dots J \tag{1}$$

$\beta^0_j$ = constant term; $\beta_{ji}$ = regression coefficient; $\zeta_j$ = residual term.

Wold [Wold, 1966] formalized the concept of partial least squares. His algorithm consists in estimating the latent variables (outer estimate and inner estimate) and the structural equations by OLS (Ordinal Least Squares) multiple regression with an iterative process. The initial value of the coefficients being equal to $\pm 1$, according to the sign of the correlation between latent variables or between latent and manifest variables.

### 1.2 A comparison between PLS and LISREL

We will follow here [Jöreskog and Wold, 1982], [Chin, 2000] and [Vinzi, 2003]. In PLS approach, there are less probabilistic hypotheses, data are modeled by a succession of simple or multiple regression and there is no identification problem. On the contrary in LISREL, the estimation is done by maximum likelihood, based on the hypothesis of multinormality and allows the modelisation of the variance-covariance matrix. However, identification problems and non-convergence of the algorithm are sometimes encountered. The differences between the estimations for a causal model using PLS and LISREL depends on the order in which the parameters of the model and latent variables are computed. For PLS the estimated latent variables are first computed by making them belong to the space spanned by their manifest variables. The model parameters are computed by using OLS multiple regression. With LISREL, one computes the model parameters by maximum likelihood and impose some constraints on latent variables. Consequently, the structural equations are more significant in LISREL than in PLS (the $R^2$ are larger) and the correlations between the manifest variables and their latent are larger in PLS. In LISREL approach, each latent variable is estimated by multiple regression, using all manifest variables. In PLS, latent variables are calculated as a linear combination of the associated manifest variables. PLS favours the outer model and LISREL the inner model. The table 1 summarizes criteria for choosing between PLS and LISREL.

| *Criteria* | *PLS* | *LISREL* |
|---|---|---|
| Objective | Prediction oriented | Oriented to parameters estimation |
| Approach | Variance based | Covariance based |
| Latent variables | Each latent variable is a linear combination of its own manifest | The latent variables are estimated using the whole set of manifest variables |
| Relationship between a latent variable and its manifest variables | Formative or reflective way | Reflective way only |
| Implications | Optimal for prediction accuracy | Optimal for parameter accuracy |
| Model complexity | Large complexity (e.g., 100 latent and 1000 manifest) | Small / moderate complexity (e.g., less than 100 manifest) |
| Sample size | Minimal recommendations range from 30 to 100 cases. | Minimal recommendations range fr m 200 to 800 |
| Theory requirements | Flexible | Strong assumptions |
| Missing data treatment | NIPALS algorithm | Maximum likelihood method |
| Identification | Under recursive models is always identified | Depends on the model; ideally need 4 or more manifest per latent to be over determined, 3 to be just identified |

**Table 1.** Criteria for choosing between PLS and LISREL.

## 2 Practical application

### 2.1 Satisfaction in automobile market

Taking into account that the PLS approach is less used than LISREL in marketing research, even though it is more advantageous than the latter, our objective was to introduce how PLS works and to show its' capacities. To reach this goal, we used data provided by the PSA Company (Peugeot Citroën) on customers' satisfaction. We used the experimental PLSX module of the SPAD software, which has been developed within the framework of the ESIS project about the construction of a tool to analyze European customer satisfaction.

### 2.2 The questionnaire

The data obtained by questionnaire (which is confidential) represents satisfaction scores (with the scale of 1 to 10) on about thirty services. 2922 customers participated. Manifest variables are the followings (table 2):

| Variable | | Variable | |
|---|---|---|---|
| General satisfaction | S01 | Radio - CD - rom | S17 |
| General quality | S02 | Heating - ventilation | S18 |
| Quality -price ratio | S03 | Boot capacity | S19 |
| Absence of small, irritating defects | S04 | Security | S20 |
| Absence of noise and vibrations | S05 | Braking | S21 |
| General state of the paintwork | S06 | Acceleration | S22 |
| Robustness of commands, buttons | S33 | Handling | S23 |
| Solidity and robustness | S08 | Suspension comfort | S24 |
| Lock, door and window mechanisms | S09 | Silence in rolling | S25 |
| Inside space and seat modularity | S34 | Maniability | S26 |
| Inside habitability | S11 | Direction | S27 |
| Dashboard: quality of materials and finishing | S12 | Gears | S28 |
| Insider: quality of mat. and finishing | S13 | Mechanic reliability | S29 |
| Front seat comfort | S14 | Oil consumption | S30 |
| Driving position | S15 | Mechanic's efficiency in solving problems | S31 |
| Visibility from driver's seat | S16 | Maintenance cost and repairs | S32 |

**Table 2.** Manifest variables.

Since we are interested in the relationships between variables, and not in their values, it was not necessary to rescale the answers, despite the fact that customers do not use the scale in the same way.

## 3 The analysis

### 3.1 Blocks building

We first had to partition the manifest variables (MV) into homogenous blocks, each one being explicitly associated with only one latent variable. After many trials and with the help of experts, we considered the following division of the 32 variables into 6 blocks (table 3):

### 3.2 The causality scheme

The measurement model has been established in the previous paragraph.

**3.2.1 The structural model (inner model)** Supposing that the themes reflect correctly the characteristics of the satisfaction, we must then propose relations between these themes, so as to explain the latent variable "general satisfaction". In the figure 1 we can visualize the structural model which shows the relations between the latent variables:

| Block | Label | MV | Block | Label | MV | Block | Label | MV |
|---|---|---|---|---|---|---|---|---|
| General satisfaction | Gsat | S01 | Internal comfort | Comf | S11 | Driving quality | Drivq | S20 |
| | | S02 | | | S12 | | | S21 |
| | | S03 | | | S13 | | | S22 |
| Construct quality | Conq | S04 | | | S14 | | | S23 |
| | | S05 | | | S15 | | | S24 |
| | | S06 | | | S16 | | | S25 |
| | | S28 | | | S17 | | | S26 |
| | | S29 | | | S18 | | | S27 |
| Solidity | Soli | S08 | | | S19 | Costs | Costs | S30 |
| | | S09 | | | S34 | | | S31 |
| | | S33 | | | | | | S32 |

**Table 3.** The 6 blocks of manifest variables.

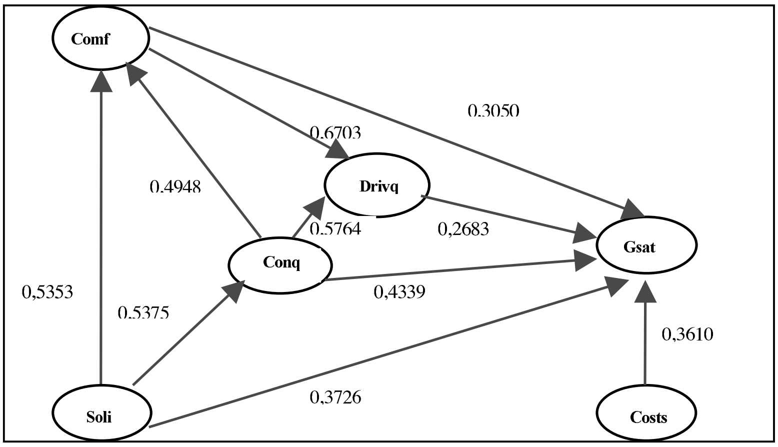


**Fig. 1.** The causality scheme with correlations values between latent variables.

### 3.3 Results and interpretations

We see that the variable "construction quality" is the most important variable for the "general satisfaction" (the correlation coefficient is 0,4339) and the less important is the "driving quality" (the correlation coefficient is 0,2683). Consequently, in order to increase the general satisfaction of the client, the producer should concentrate firstly on the "construction quality" and then on the "solidity", "costs", "internal comfort" and "driving quality". Let us now interpret the results in detail.

**3.3.1 The measurement model** After convergence of the PLS algorithm, one obtains the final weights which allow us to link the manifest variables with the latent variables:

$$
\begin{aligned}
Gsat &= 0,2188 \times S01 + 0,5746 \times S02 + 0,4850 \times S03 \\
Soli &= 0,4682 \times S08 + 0,4242 \times S09 + 0,4151 \times S33 \\
Conq &= 0,2103 \times S04 + 0,2730 \times S05 + 0,3396 \times S06 + 0,3930 \times S28 \\
&\quad +0,3787 \times S29 \\
Drivq &= 0,1962 \times S20 + 0,1595 \times S21 + 0,1415 \times S22 + 0,1615 \times S23 \\
&\quad +0,1775 \times S24 + 0,1658 \times S25 + 0,1728 \times S26 + 0,1805 \times S27 \\
Comf &= 0,1492 \times S11 + 0,1795 \times S12 + 0,1756 \times S13 + 0,1542 \times S14 \\
&\quad +0,1667 \times S15 + 0,1424 \times S16 + 0,1282 \times S17 + 0,1457 \times S18 \\
&\quad +0,1092 \times S19 + 0,1513 \times S34 \\
Costs &= 0,2396 \times S30 + 0,5707 \times S31 + 0,5042 \times S32
\end{aligned}
$$

Table 4 presents only correlations larger than the mean of the absolute values (0,3723):

We observe that all latent variables are well correlated with their own manifest. So, the manifest variables “describe” their latent appropriately and the blocks are therefore validated. We see also that the largest correlation (0,8692) is between “general satisfaction” and their manifest “quality in general”.

The $R^2$ coefficients between connected latent variables are:

$$
\begin{aligned}
R^2(Conq; Soli) &= 0,2889 \\
R^2(Comf; (Soli, Conq)) &= 0,3468 \\
R^2(Drivq; (Conq, Comf)) &= 0,5286 \\
R^2(Gsat; (Soli, Conq, Comf, Drivq, Costs)) &= 0,2516
\end{aligned}
$$

In this table the most interesting relation concerns the “general satisfaction”. For this variable, the $R^2$ coefficient generated by the other latent variables is 25%, and we consider that as satisfactory because there are 2922 individuals. The correlations between the latent variables are given below in table 5.

We can see that to improve “internal comfort”, the producer should concentrate on “solidity” (correlation coefficient = 0,5353) and on the “construction quality” (0,4948). The producer?s efforts for improving “construction quality” also greatly affect the variable “leading quality” (0,5764). In order to obtain a good “construction quality” the producer could concentrate on “solidity” (0,5375).

We also observe an important correlation between “solidity” and “driving quality”. We have chosen not to establish a relation between these two because this relation does not in any way influence the model. Given the causality scheme the determination of “general satisfaction” is a complex procedure in which almost all the latent variables are directly involved. The equation is as follows:

$$
\begin{aligned}
Gsat = {} & 0,2721 \times Conq + 0,1678 \times Soli + 0,198 \times Costs \\
& +0,082 \times Comf + 0,095 \times Drivq
\end{aligned}
$$

| Variable | Solidity | Construct quality | Internal comfort | Driving quality | Costs | General satisfaction |
|---|---|---|---|---|---|---|
| S08 | *0,7988* | 0,4272 | 0,4568 | 0,4539 | | |
| S09 | *0,7492* | 0,3951 | 0,4040 | | | |
| S33 | *0,7425* | 0,4093 | | | | |
| S04 | | *0,5456* | | | | |
| S05 | | *0,5847* | | | | |
| S06 | 0,4187 | *0,5897* | | | | |
| S28 | | *0,6666* | 0,4336 | 0,5423 | | |
| S29 | | *0,6954* | | 0,4608 | | |
| S11 | | | *0,6965* | 0,4405 | | |
| S12 | 0,4173 | 0,3987 | *0,7276* | 0,4684 | | |
| S13 | 0,4089 | 0,3882 | *0,7317* | 0,4441 | | |
| S14 | | | *0,7382* | 0,4829 | | |
| S15 | | | *0,7662* | 0,5398 | | |
| S16 | | | *0,6278* | 0,4427 | | |
| S17 | | | *0,5055* | 0,3784 | | |
| S18 | | | *0,5741* | 0,4419 | | |
| S19 | | | *0,5379* | 0,3746 | | |
| S34 | 0,3956 | | *0,6535* | 0,4048 | | |
| S20 | 0,4556 | 0,4276 | 0,6083 | *0,7303* | | |
| S21 | 0,3796 | 0,4079 | 0,4520 | *0,6878* | | |
| S22 | | | 0,4385 | *0,6866* | | |
| S23 | | 0,3929 | 0,4903 | *0,7914* | | |
| S24 | | 0,4267 | 0,5029 | *0,7600* | | |
| S25 | | 0,4397 | 0,4051 | *0,6424* | | |
| S26 | | 0,4575 | 0,5104 | *0,7980* | | |
| S27 | 0,3749 | 0,4826 | 0,5155 | *0,7913* | | |
| S30 | | | | 0,3880 | *0,4450* | |
| S31 | | | | | *0,8374* | |
| S32 | | | | | *0,8242* | |
| S01 | | | | | | *0,6399* |
| S02 | | 0,4234 | | | | *0,8692* |
| S03 | | | | | | *0,7433* |

**Table 4.** Correlations between manifest and latent variables.

| Variable | Soli | Conq | Comf | Drivq | Costs | Gsat |
|---|---|---|---|---|---|---|
| Soli | 1 | | | | | |
| Conq | 0,5375 | 1 | | | | |
| Comf | 0,5353 | 0,4948 | 1 | | | |
| Drivq | 0,4949 | 0,5764 | 0,6703 | 1 | | |
| Costs | 0,3116 | 0,4331 | 0,3144 | 0,3461 | 1 | |
| Gsat | 0,3726 | 0,4339 | 0,305 | 0,2683 | 0,361 | 1 |

**Table 5.** Correlations between latent variables.

The negative coefficient for "driving quality" can be explained by the fact that this variable increases with "construction quality" and the regression coefficient between "construction quality" and "general satisfaction" is 0,2721.

This multiplication coefficient is without doubt corrected by the negative coefficient on the "driving quality".

## 4 Conclusions

Firstly it must be underlined that this study did not follow the logical sequence of steps of the PLS approach: the construction of a model by experts, the construction of a questionnaire using this model, and the collection of customer data using this questionnaire. In our case, we have inverted the process: we have tried to build a model using data that had already been collected with a questionnaire. This fact has obviously effects on the final results which cannot be precisely measured. A hierarchy of the influence of the latent variables on general satisfaction can be established using the structural model: I. Construction quality; II. Solidity; III. Costs; IV. Internal comfort; V. Driving quality. The results obtained for general satisfaction are satisfactory: $R^2 = 25\%$ which is a good result for a large sample of almost 3000 respondents.

## References

[Chin, 2000]W. W. Chin. *Partial Least Square for researchers: a overview and presentation of recent advances using the PLS approach*, http://discnt.cba.uh.edu/chin/indx.html, 2000

[Jöreskog and Wold, 1982]K. G. Jöreskog, H. Wold. *The ML and PLS techniques for modeling with latent variables: historical and competitive aspects.* In K. G. Jöreskog and H. Wold, editors, *Systems under indirect observation*, Part 1, North-Holland, Amsterdam, pages 263–270, 1982

[Tenenhaus *et al.*, 2005]M. Tenenhaus, V. E. Vinzi, Y-M. Chatelin, C. Lauro. *PLS path modeling.* Computational Statistics & Data Analysis, volume 48, issue 1, pages 159–205, January 2005

[Vinzi, 2003]V. E Vinzi. *The PLS Approach to Path Modeling.* IASC-IFCS Summer School, Lisbon, 2003

[Wold, 1966]H. Wold. *Estimation of principal components and related models by iterative least squares*, in Krishnaiah, P.R., editor, Multivariate Analysis, Academic Press, New York, pages 391–420, 1966