Trendy v podnikání - Business Trends (2018), 8(1), 45-55.
https://doi.org/10.24132/jbt.2018.8.1.45_55

LINEAR MODELS BASED ON BUSINESS DATA FROM THE PHARMACEUTICAL INDUSTRY

David Říha[1], Michael Stros[2]

[1] David Říha, University of Economics, Prague, Czech Republic, david.riha@vse.cz
[2] Michael Stros, University of Applied Sciences and Arts of Southern Switzerland (SUPSI), Fernfachhochschule Schweiz, Brig, Switzerland, michael.stros@ffhs.ch

Abstract: This article describes the analysis of heterogeneous market data. For this purpose, the most relevant methodological aspects are discussed and analyses using a hierarchical linear model and multiple regression are presented. In the first step, the applied data set is presented, and the assumed hierarchical two-level structure is shown. The data are then prepared for the analysis. The data are checked for outliers, a multicollinearity check is conducted, a new variable introduced, missing values are replaced by estimated values, a transformation procedure is conducted in order to obtain normality, the data are aggregated for each hierarchical level and a sample size test is performed. The results of both methods are discussed. Finally, it is concluded that whereas the application of a hierarchical linear model appears to be one option, a multiple regression analysis can be employed instead if the quality of the data, especially the sample size, is not sufficient.

Keywords: Business data analysis, Heterogeneous data set, Hierarchical linear model, Multiple regression

JEL Classification: C18

INTRODUCTION

The analysis of heterogeneous data, i.e., data belonging to different clusters, can be a challenging task. In order to analyse relations between constructs with different aggregation levels, multilevel research settings are applied in social science, especially in the disciplines of education and medicine (see Browne et al., 2002; Leeuw and Kreft, 1986). However, not much empirical research in marketing has applied multilevel and hierarchical data so far (Pieters and Wedel, 2004). Despite the fact that multilevel analysis is commonly used in several scientific disciplines, it is not often applied in marketing research (see Jong et al., 2004; Pieters and Wedel, 2004). This statement is supported by MacKenzie (2001) who summarizes that in marketing, "researchers have tended to emphasize either a micro- or macro-level perspective without recognizing the interaction between the two". However, as stated by several researchers (Liao and Chuang, 2004; van Bruggen et al., 2002) in some cases, the use of a single-level analysis might not be appropriate. As a result, there is an increasing demand for the aggregation of multiple responses (see van Bruggen et al., 2002). In addition to this, according to Osborne (2002), the application of standard statistical approaches for analysing multilevel or naturally grouped data can lead to false conclusions.
The aim of this paper is to provide a practical guideline on how such data can properly be analysed. For this purpose, two methods will be discussed: the hierarchical linear model (HLM) and multiple regression

1. LITERATURE REVIEW

There is no single best way to analyse a multilevel structure. As stated by Harrell (2001), the individual steps that a researcher should take in building a model are dependent on the investigator's research questions,

on whether the analysis is explanatory or confirmatory and on whether the analytic emphasis is on parameter estimation, model fit or prediction.

One option that may be considered for analysing heterogeneous data is the hierarchical linear model (HLM), also called a random coefficient model (see Leeuw and Kreft, 1986; Longford, 1993). This methodology seems especially suitable because, as indicated by Kozlowski and Klein (2000), the nesting of micro- and macro-level phenomena is taken into account, as are macro-level effects that occur through interactions with micro-level elements (Kozlowski and Klein, 2000). Consequently, according to Goldstein (1995), the major advantage of the HLM is the possibility of linking multiple levels simultaneously in a single regression equation. However, according to most researchers (Hox and Maas, 2002; Wieseke et al., 2008), a minimum sample size per level and group is required in order to run an HLM. A rule of thumb recommends a minimum of 30 samples per group (Bell et al., 2008; Hox and Maas, 2002; Moineddin et al., 2007).

If the data do not fulfil this requirement, an HLM analysis cannot be applied. Consequently, a multiple regression analysis must be conducted instead.

A multiple regression analysis is defined by Hair et al. (1998, p20) as “a general statistical technique used to analyse the relationship between a single dependent variable and several independent variables”. In other words, multiple regression is only able to test hypotheses with respect to a single dependent variable. This means that a complete conceptual model cannot be tested all at once, and therefore multiple models must be examined instead. In this case, the application of regression analysis is viewed as the best strategy for heterogeneous data.

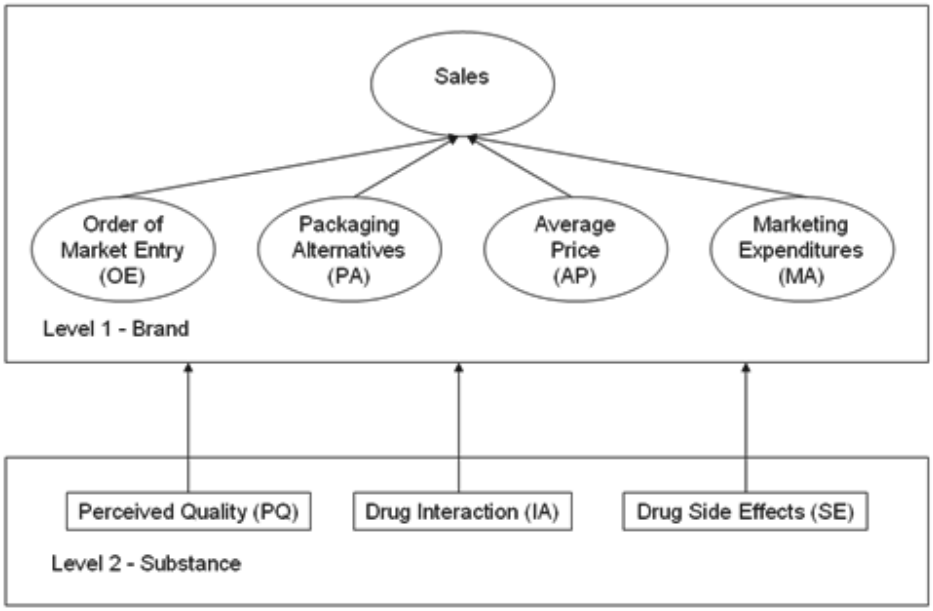
2. METHODOLOGY

In the following paragraph, the practical application of multilevel and multiple regression will be discussed, using the same data set in each case. First, the applied data set is presented.

As an illustrative example, the analysis is conducted using a heterogeneous data set. This is a data set that contains primary and secondary data from the Swiss pharmaceutical sector. The secondary data were collected by a market research company, via a network of associated doctors, pharmacists and wholesalers, by gathering data on the medical drugs sales transactions on a monthly basis. The data set covered a total of five prescription-drug classes, and contained sales information on 37 substances from 108 products (brands) in Switzerland for the period between 1995 and 2005. In order to conduct the analysis, the data were transformed into a specific format. For this purpose, relevant information on each medical product, indicating the drug class code (DC), substance code (SC), brand name code (BN), perceived quality (PQ), order of market entry (OE), number of packaging alternatives (PA), application range (AR), number of drug interactions (IA) and number of side-effects (SE) was collated on an Excel spreadsheet. In addition, the total detailing expenditure (DE) in Swiss francs, total mailing expenditure (ME), total advertising expenditure (AE), average daily drug dose (DDD), price (AP), average of product sales (AS) and beta of sales (BS) were also collected.

An initial data analysis revealed a hierarchical data structure. Therefore, a two-level structure, containing a brand (first) level and a substance (second) level, is suggested. The substance level includes perceived quality (PQ), drug interaction (IA) and drug side-effects (SE). These data refer only to a specific substance, and there is no dependency on a specific brand (multiple brands can use the same substance, e.g., paracetamol). The brand level, on the other hand, contains the order of market entry (OE), number of packaging alternatives (PA), average price (AP) and marketing expenditure (MA) as independent variables, whereas sales is a dependent variable, as shown in the following illustration (see Figure 1):

Fig. 1 Visualization of the data structure



Source: Own processing, 2018

2.1. Data Preparation and Assumption Test

In order to be able to conduct a data analysis, there are some requirements for the data. It is especially relevant that statistical independence of observations, normality and linear relationships between the dependent and independent variables are present, together with the equality of variance (homoscedasticity) (see Hair et al., 1998; Kleinbaum et al., 1998). Several diagnostic statistics and diagrams were produced to identify outliers and to analyse the violation of assumptions, multicollinearity and the power of the test (see Hair et al., 1998; Kleinbaum et al., 1998; Kaplan, 1995). The assumptions of normality, linearity and homoscedasticity were examined using graphical techniques (see Hair et al., 1998; Kleinbaum et al., 1998). However, for the data collection process, as previously described, the statistical independence of the observations can be assumed.

According to Kleinbaum et al. (1998), outliers could have a negative influence on the analysis outcome, since outliers may negatively influence the normal distribution. The deletion of outliers is controversial in the literature (see Barnett and Lewis, 1994), since the procedure might influence the results of the statistical analysis. However, depending on the reason for the existence of outliers, such as errors in answering questions or data imputation errors, deletion might be justified. Unfortunately, there is no generally applicable strategy for dealing with outliers (see West et al., 1995). A range of statistical methods available in SPSS can be applied in order to identify possible outliers. The indicated outliers can be justified. They have not been caused by a measurement or data handling error. However, the fact that outliers have been removed needs to be considered when statistical results are interpreted.

Another issue that should be taken into account is multicollinearity. As stated by Kleinbaum et al. (1998), multicollinearity takes place when there is a significant correlation between independent variables in a model. Consequently, it is difficult to separate the effects of each independent variable, resulting in unstable statistical results (see also Cohen and Cohen, 1975; Kleinbaum et al., 1998). One approach employed to tackle this problem is the deletion of one of the collinear variables or the transformation of collinear variables (see Cohen and Cohen, 1975). A test for multicollinearity was performed, during which tolerance values and their variance inflation factors were examined. According to Kleinbaum et al. (1998), problematic multicollinearity is indicated by tolerance values below 0.1 and variance inflation factors above 30. Based on the above, multicollinearity

between advertising expenditure (AE) (tolerance = 0.138; variation inflation = 7.261), detailing expenditure (DE) (tolerance = 0.064; variation inflation = 15.591) and mailing expenditure (ME) (tolerance = 0.1; variation inflation = 10.018) was detected. There is also a fairly high correlation between these factors, as illustrated in the following table (see Table 1)..

Tab. 1: Marketing variable correlations

		DE	ME	AE
Detailing expenditure (DE)	Pearson Correlation Sig. (2-tailed)	1.000	.943 .000	.921 .000
Mailing expenditure (ME)	Pearson Correlation Sig. (2-tailed)	.943 .000	1.000	.883 .000
advertising expenditure (AE)	Pearson Correlation Sig. (2-tailed)	.921 .000	.883 .000	1.000

Source: Own processing, 2018

Therefore, these variables were combined by adding the values and then calculating the monthly average. This resulted in the single marketing variable of average marketing expenditure (AM). In the next step, the data were checked for outliers and missing data. Five outliers from the average price (AP) variable, and one outlier from the average sales (AS) and perceived quality (PQ) variables were removed. The drug interaction (IA) and side-effects (SE) variables contained three missing values (3.4%). In addition, there were five missing values (5.8%) for average marketing expenditure (AM). The indicated outliers can be justified. They were not caused by a measurement or data handling error; instead, they were missing due to the unavailability of data. These data can therefore be characterized as MAR (missing at random) values. This means that whatever causes the data to be missing does not depend upon the missing data themselves (Little and Rubin, 2002). Consequently, there are no restrictions given when replacing these data with an estimated value, as described in the following section.

The handling of missing values is quite challenging. SPSS offers single imputation approaches such as mean value and regression substitution. However, several authors such as Graham (2009), Howell (2007) and Schafer (1997) do not recommend the use of these methods because of their weaknesses (alteration of the correlation coefficient). Instead, the EM algorithm (multiple imputation) is recommended (Graham, 2009; Little and Rubin, 2002; Schafer, 1997). These researchers highlight the fact that multiple imputations are suitable because it has been shown that they produce unbiased parameter estimates, they are robust to departures from normality assumptions and they provide adequate results in cases of small sample size. For this purpose, the freely available software for multiple imputation NORM (see Pennsylvania State University homepage: sites.stat.psu.edu) was used in this study (see also Schafer, 1997). The missing values were replaced by estimates derived from the NORM routine. It should be added that the low number of missing values (below 5%) can be viewed as statistically insignificant (see Howell, 2007; Little and Rubin, 2002).

According to Osborne (2002), a serious violation of the assumption of normality can affect a result. Furthermore, it must be pointed out that, according to Micceri (1989), it is not unusual to find that data are not distributed normally in the fields of psychology and education. In this case, the literature suggests a data transformation procedure (see Backhaus et al., 2003; Hair et al., 1998; Hartwig and Dearing, 1979; Osborne, 2002). However, Kleinbaum et al. (1998, p117) stated that “only extreme departures from normality lead to spurious results”. Furthermore, in addition to individual univariate normality, multivariate normality should be assessed. Even when all individual univariate distributions are normal, it is not necessarily true that the multivariate distribution will be normal (Hair et al., 1998; Sharma, 1996).

In a second step, a transformation was performed in order to obtain normality. In the literature (Backhaus et al., 2003; Hair et al., 1998; Hartwig and Dearing, 1979; Osborne, 2002), three different transformation procedures have been suggested: (1) square root transformation, (2) logarithmic transformation and (3)

inverse transformation. It is suggested that the minimum amount of transformation should be applied, beginning with the square root transformation, in order to improve normality (Osborne, 2002). In this case, the logarithmic transformation using e as the base was regarded as appropriate because this function has shown the best results for improvements towards a normal distribution. It should be added that a higher base tends to pull in extreme values to a greater extent than a lower base (Cleveland, 1984). Transformation improves normality by reducing the distances between data points. However, Osborne (2002) states that all data points remain in the same relative order as they were prior to the transformation, which allows researchers to continue to interpret results in terms of increasing scores. The transformation resulted in a significant improvement in normality, as illustrated in the following table (see Table 2).

Tab. 1: Normality test results

Variable	Skewness	z-Score	Kurtosis	z-Score	Comments
Perceived Quality (PQ)	-1.076	-4.074	1.297	2.455	improvement was obtained
Order of market Entry (OE)	0.521	1.973	-0.602	-1.139	normally distributed
Packaging Alternatives (PA)	-0.058	-0.220	-0.644	-1.220	normally distributed
Drug Interaction (IA)	0.363	1.374	1.072	2.029	normally distributed
Drug Side-effects (SE)	0.344	1.302	-0.802	-1.518	normally distributed
Average Marketing Expenditure (AM)	-0.146	-0.553	-0.659	-1.247	normally distributed
Average Price (AP)	0.493	1.866	2.427	4.594	improvement was obtained
Average Sales (AS)	0.251	0.950	-0.205	-0.388	normally distributed

Source: Own processing, 2018

For a multilevel data structure, an analysis using multiple regression must be conducted for each single level separately, and therefore the data must be aggregated for the second level, as suggested by Hox (2010). For the data aggregation at the second level (substance), first-level (brand) data were taken and the average value for each substance was calculated. This resulted in a reduced data set (initially 86 data points) containing 26 data points at the second level. The data were then standardized using SPSS, resulting in an overall average of zero and a standard deviation and variance of one.

Regarding the sample size, the market data can be considered complete for the previously described five drug classes. Consequently, these five drug classes were defined as the overall population size (100%) in the current research (containing 108 brands and 37 substances). Taking into account the fact that an expected sampling frequency of 50% could be assumed [for samples providing the required precision levels, if unknown, a value of 50% is taken (Rovezzi, 2002)], calculation of the sample size revealed that for the brand level (confidence level 95%, confidence interval 5%) , at least 84 data points are required. [The determination of the sample size is described by Armitage et al. (2002). The sample size can also be determined using online calculation tools, e.g., www.macorr.com/sample-size-calculator.htm.] At the substance level, at least 26 data points are required (confidence level 95%, confidence interval 10%) (please refer also to the statistical literature, e.g., Backhaus et al., 2003; Lenth, 2001). It can therefore be concluded that, with respect to sample size, this data set provides a robust basis for statistical analysis.

2.2. Multilevel Data Analysis

In order to test the presented model (see Figure 1), a hierarchical linear model (HLM), also called a random coefficients model (see Leeuw and Kreft 1986; Longford 1993), was applied. This methodology is especially suitable, because, as stated by Kozlowski and Klein (2000), HLM explicitly takes the nesting of micro- and macro-level phenomena into account. In addition to this, HLM accounts for macro-level effects that occur

through interactions with micro-level elements (Kozlowski and Klein, 2000). Consequently, according to Goldstein (1995), the major advantage of the HLM is the possibility of linking multiple levels simultaneously in a single regression equation. Furthermore, according to Hofmann (1997), HLMs overcome the weakness of disaggregated and aggregated multilevel analysis approaches. For the present case, HLM seems to be the most appropriate method for the analysis of the data set. The following model was set up (see Figure 2):

Fig. 2 Linear multilevel regression equation

$$\begin{aligned}\text{Current\_Sales}_{Substance,\,Brand} &= \beta_0 \text{PQPerceivedQuality}_{Brand} + \beta_1 \text{OE}_{Brand} + \beta_2 \text{DEDetailingExpenditures}_{Brand} + \\ &\quad \beta_3 \text{MACMailingExpenditures}_{Brand} + \beta_4 \text{AEAdvertisingExpenditures}_{Brand} + \\ &\quad \beta_5 \text{APAveragePrice}_{Brand} + \beta_{6Brand} + \beta_7 \text{PKSnumberofpackingalternatives}_{Brand} + \beta_8 \text{DSE}_{Brand} \\ &\quad + e_{Substance,\,Brand} \\ \beta_{6Brand} &= \beta_6 + u_{6Brand}\end{aligned}$$

Source: Own processing, 2018

In the regression equation, the dependent variable is current sales, and the independent variables of perceived quality, order of market entry, detailing expenditures, mailing expenditures, advertising expenditures, average price, number of packaging alternatives and drug side-effects are weighted with factors β1 to β8. Furthermore, the intercept β6 and the first-level residual ε (substance) as well as μ, the second-level residual (brand) are introduced.

For this analysis, MLwiN software was used (see University of Bristol, Centre for Multilevel Modelling homepage: www.bristol.ac.uk/cmm/software/mlwin/) applying an OLS regression with a maximum likelihood and IGLS estimation control method. Following the recommended procedure for HLM analysis (Hofmann 1997; Raudenbush and Bryk, 2002) the variance within and between the groups in the dependent variable was examined. In the next step, it was assessed whether there was a significant variance in the intercepts and slopes across groups, in order to specify the best fitting random coefficient model (see also Konradt et al., 2009)

The interpretation of the first multilevel regression analysis, using the dependent variable of sales, showed some level of variance for employed variables (see Figure 3). There is a high significant positive relationship to perceived quality (PQ), a strong negative relationship to order of market entry (OE) and a (low variance) negative relation to detailing (DE), a positive to mailing (ME) and advertising expenditures (AE). However, the three weighting factors have a low value. Therefore, it can be concluded that the relationship is weak.

Fig. 1 Linear multilevel regression equation

$$\begin{aligned}\text{Current\_Sales}_{Substance,\,Brand} &= 103605.695(113373.305)\text{PQPerceivedQuality}_{Brand} + \text{-2553.590(9571.609)OE}_{Brand} + \\ &\quad \text{-0.252(0.078)DEDetailingExpenditures}_{Brand} + 1.003(1.117)\text{MACMailingExpenditures}_{Brand} + \\ &\quad 1.539(0.164)\text{AEAdvertisingExpenditures}_{Brand} + \text{-213.025(235.158)APAveragePrice}_{Brand} + \\ &\quad \beta_{6Brand} + 15399.511(36135.820)\text{PKSnumberofpackingalternatives}_{Brand} + \\ &\quad \text{-1021.462(1491.157)DSE}_{Brand} + e_{Substance,\,Brand} \\ \beta_{6Brand} &= \text{-318814.031(459369.406)} + u_{6Brand} \\ u_{6Brand} &\sim N(0, \sigma^2_{u6}) \quad \sigma^2_{u6} = 210931351552.000(32166727680.000) \\ e_{Substance,\,Brand} &\sim N(0, \sigma^2_e) \quad \sigma^2_e = 0.000(0.000) \\ \textit{-2*loglikelihood} &= 2486.490(86 \text{ of } 99 \text{ cases in use})\end{aligned}$$

Source: Own processing, 2018

According to most researchers (Hox and Maas, 2002; Wieseke et al., 2008), there is a minimum sample size per level and group which is necessary in order to run an HLM. A rule of thumb recommends a minimum of 30

samples per group (Bell et al., 2008; Hox and Maas, 2002; Moineddin et al., 2007). The applied data do not fulfil this requirement. Therefore, a robust and valid HLM analysis cannot be applied. Consequently, a multiple regression analysis must be conducted instead.

2.3. Multiple Regression Data Analysis

In this section, an analysis of the multilevel structure is performed using multiple linear regression. In order to test the previously presented hypotheses, a set of multiple regression equations is produced. Every equation is then examined for violation of the assumption.

In this section, a multiple regression model was created by taking the findings from the previously performed data structure analysis and the previously hypothesized factor relations into account. Furthermore, the slope of the sales [beta sales (BS)] should be investigated as an additional independent variable. This was also taken into account when creating the multiple regression models, as discussed later. Consequently, for each level [(A) average sales (AS) and (B) beta sales (BS) as dependent variables], two models were created.

A number of different model selection methods are described in the literature (see Kleinbaum et al., 1998). Independent variables are chosen by model selection methods such as forwards, backwards, stepwise and simultaneous entry methods (see also Hair et al., 1998; Kleinbaum et al., 1998). However, it has been noted that stepwise entries are potentially problematic and should only be used for entirely predictive rather than explanatory models (Hair et al., 1998; Cohen and Cohen, 1975). Consequently, taking into account the fact that the purpose was to test hypotheses and not to predict any dependent variables, simultaneous entry methods were applied.

In order to test these variables for multicollinearity , tolerance values
(all above 0.658 > 0.1) and variance inflation factors (all below 1.519 < 30.0) were calculated by entering them simultaneously into the regression equation (see Hair et al., 1998; Kleinbaum et al., 1998; Kaplan, 1995). The results did not display any obvious problems. The following first-level model using average sales (revenue) as a dependent variable was then investigated by applying the multiple linear regression function in SPSS (see Table 4):

$$AS_i = \beta_0 + \beta_1*(OE_i) + \beta_2*(AP_i) + \beta_3*(PA_i) + \beta_4*(AM_i) + \beta_5*(PQ_i) + \beta_6*(IA_i) + \beta_7*(SE_i) \quad (2)$$

In the regression equation, average sales (AS) is the dependent variable, and the independent variables order of market entry (OE), average price (AP), number of packaging alternatives (PA), average marketing expenditure (AM), perceived quality (PQ), drug interaction (IA) and drug side-effects (SE) were weighted with factors β_1 to β_7. Furthermore, the intercept β_0 was introduced.

Tab. 2: Results of the first-level multiple regression

Multiple R = 0.551 R^2 = 0.330 Adjusted R^2 = 0.241 F = 4.854 (Sig. 0.000; $F_{critical}$ = 2.129)			
Independent Variable	Beta	t	Sig.
Order of market Entry (OE)	-0.083	-0.798	0.427
Drug Interaction (IA)	0.092	0.932	0.354
Drug Side-effects (SE)	0.103	0.943	0.349
Perceived Quality (PQ)	0.075	0.746	0.458
Packaging Alternatives (PA)	0.114	1172	0.245
Average Price (AP)	0.210	1.804	0075
Average Marketing Expenditure (AM)	0.423	4.147	0.000

Source: Own processing, 2018

The results gave an adjusted R2 of 0.241. This means that 24.1% of the variance can be explained by the elements of the equation, and that the independent variables are 24.1% related to the dependent variable. The rather low number can be justified by considering the complex nature of the sales process (see Cohen and Cohen, 1975). It should be noted at this point that other studies within sociology, having conducted

regression analyses, have also obtained similar variance values (see McKee et al., 2001; Wild et al., 2004). The equation is significant (sig = 0.000) and the F-value (4.854; explained variance divided by unexplained variance) is above the calculated critical F-value (2.129). Support for hypothesis H6 could be found (beta = 0.114; sig = 0.075). In the case of H7, the results give strong support (beta = 0.423; sig = 0.000). This means that an increase in average marketing expenses (AM) will lead to higher sales (revenue). It can be seen that hypotheses H2 to H5 do not find support. In other words, side-effects (SE), drug interactions (IA), perceived quality (PQ) and packaging alternatives (PA) do not influence the prescribing decision. However, it was revealed by the descriptive data analysis that there is variation between actual sales (revenue) and order of market entry.

In the next step, a test for linearity and homoscedasticity was performed using residual plots. No clear patterns could be found, and so the assumption of linearity and homoscedasticity was retained. In order to detect the presence of autocorrelation [a relationship between values separated from each other by a given time lag (Bhargava et al., 1983)], a Durbin-Watson test was performed, giving a value of 1.979. According to the rule of thumb (see Gujarati, 2003), the Durbin-Watson value should not be below 1.0. Therefore, it can be assumed that no autocorrelation is present, and a valid statistical test can be performed.

2.4. Multiple Regression Analysis of the Second-Level Model

For the second-level (substance) multiple regression model, using aggregated data, the variables were tested for multicollinearity. Tolerance values (all above 0.895 > 0.1) and variance inflation factors (all below 1.117 < 30.0) were calculated by entering them simultaneously into the regression equation (see Hair et al., 1998; Kleinbaum et al., 1998; Kaplan, 1995). The results do not display any obvious problems. In the next step, the model containing only level two (substance)-related variables was investigated by applying the multiple linear regression function in SPSS (see Table 5).

Tab. 3: Results of the second-level multiple regression

Multiple R = 0.584 R^2 = 0.341 Adjusted R^2 = 0.255 F = 3.962 (Sig. 0.021; $F_{critical}$=2.544)			
Independent Variable	Beta	t	Sig.
Drug Interaction (IA)	-0.056	-0.316	0.755
Drug Side-effects (SE)	0.423	2.364	0.027
Perceived Quality (PQ)	0.368	2.158	0.042

Source: Own processing, 2018

The results produced an adjusted R2 of 0.255. This means that 25.5% of the variance can be explained by the elements of the equation and that the independent variables are 25.5% related to the dependent variable. As previously stated, the rather low number can be justified by considering the rather complex nature of the sales process (see Cohen and Cohen, 1975). The equation can be considered as being significant (0.021) and the F-value (3.962; explained variance divided by unexplained variance) is above the calculated critical F-value (2.544). The results do not display any obvious statistical problems. Again, the regression statistics given above are basically in support of the previously discussed results, and are in line with the theory (see Chapter 2). The analysis showed that drug side-effects (SE) (beta = 0.423; sig = 0.027) and perceived quality (PQ) (beta = 0.368; sig = 0.042) are significantly positively related to the sales (revenue) slope. On the other hand, no significant relations were found for drug interactions (IA).

A test for linearity and homoscedasticity was performed, using residual plots. No clear patterns could be found, so the assumption of linearity and homoscedasticity was retained. In order to detect the presence of autocorrelation, a Durbin-Watson test was performed, giving a value of 1.963. Therefore, it can be assumed that no autocorrelation is present, and a valid statistical test can be performed.

RESULTS AND DISCUSSION

A pharmaceutical market data set was provided by a market research company via 12 Excel files containing various different formats that initially had to be combined. The data set contained 10 research-relevant variables. In the next step, the data set was cleaned. For this purpose, missing values were estimated, a check for outliers was performed and descriptive statistical properties such as arithmetic mean, variance, standard deviation, skewness and kurtosis of the data set were calculated. Unfortunately, these tests revealed that, in most cases, no normal distribution was present.

The data were further explored, using descriptive statistics. This procedure determined that order of entry does not seem to occur. Furthermore, it was recognized that different sales slopes (beta) occurred. Interestingly, it seems that this aspect has not been covered so far in marketing-related research, although it is widely used in price-demand theory in economics. Consequently, it was decided to include beta sales (revenue) as a dependent variable in the research. An analysis of the data structure revealed a multilevel arrangement, containing a brand (first) level and a substance (second) level. In order to be able to proceed with further analysis, these data had to be reorganized. For this purpose, the means of the required variables per product (brand) and beta sales (revenue) were calculated, producing a data set containing 86 data points. For the analysis of this multilevel data structure, it was intended that a hierarchical linear model (HLM) should be used. However, the data set did not fulfil the minimum requirement of 30 samples per group, as specified in the literature (Bell et al., 2008; Hox and Maas, 2002; Moineddin et al., 2007). An HLM test run also highlighted the instability of the results. It was therefore decided that a multiple regression analysis would be conducted instead.

The data were then prepared for analysis. A test for multicollinearity was performed, revealing a multicollinearity problem between three marketing variables. As a result, these variables were combined (all marketing expenses were added together) into one new marketing variable. A check for outliers and missing values was performed. It appeared that outliers were present and contained some extreme values. Although they were justifiable, it appeared that these products represented exceptions on the market. Since no generalization regarding these products could be made, these outliers were removed. The missing values were then replaced by estimates derived from a multiple imputation (EM) algorithm. In order to give statistically robust results, normally distributed data are required. In the present case, a logarithmic transformation had to be conducted in order to obtain a normal distribution. As data analysis using multiple regression must be conducted for each single level separately, the data had to be aggregated (see Hox, 2010) for the second level (substance), resulting in a data set of 26 data points. However, for the second level (substance), only the relevant variables were included. Calculation of the sample size indicated that robust results could be derived from the analysis.

Finally, the analysis was performed, calculating both levels (brand and substance) using both dependent variables (average and beta sales). All models were tested successfully for their statistical robustness in the first instance. For average sales (revenue), the results showed a strong positive relation with marketing expenditure (MA) (beta = 0.752; sig = 0.000) on the first (brand) level, and with side-effects (SE) (beta = 0.423; sig = 0.027) and perceived quality (PQ) (beta = 0.368; sig = 0.042) on the second (substance) level. For beta sales (revenue), the results indicated a strong positive relation with perceived quality (PQ) (beta = 0.463; sig = 0.000) as well as the order of market entry (OE) (beta = 0.218; sig = 0.054) on the first (brand) level, and with side-effects (SE) (beta = 0.316; sig = 0.028) and perceived quality (PQ) (beta = 0.666; sig = 0.000) together with a negative interaction with drug interactions (IA) (beta = -0.276; sig = 0.051) on the second (substance) level.

In this paper, two methods for the analysis of market data were presented. Although the application of a hierarchical linear model appears to be one option, a multiple regression analysis can also be employed as an alternative, when the quality of the data, especially the sample size, is not sufficient. The applied data

set did not meet the sample size criterion given by the hierarchical linear model (HLM), as a result, this analysis has lead to different result.

REFERENCES

Armitage, P. et al. (2002). *Statistical Methods in Medical Research*. Malden, Blackwell Publishing.
Backhaus, K. et al. (2003). *Multivariate Analysemethoden*. Springer-Verlag, Berlin.
Barnett, V. and Lewis, T. (1994). *Outliers in statistical data*, 3rd edn, John Wiley & Sons, Chichester.
Bell, B. A., et al. (2008). Cluster Size Multilevel Models: The Impact of Sparse Data Structures on Point and Interval Estimates in Two-Level Models. *JSM*, 1122 - 1129.
Bhargava, A. et al. (1982). Serial Correlation and the Fixed Effects Model. *Review of Economic Studies*, 49, 533-549.
Cleveland, W. S. (1984). Graphical methods for data presentation: Full scale breaks, dot charts, and multibased logging. *The American Statistician*, 38, 270-280.
Cohen, J. and Cohen, P. (1975). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, Lawrence Elbaum and Associates, Hillsdale.
Furlong, N. E. et al. (2000) *Research Methods and Statistics - An Integrated Approach*. Harcourt College Publishers, Orlando.
Goldstein, H. (1995). *Multilevel statistical models*, Edward Arnold, London.
Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology*, 60, 549-576.
Gujarati, D. N. (2003). *Basic econometrics*. McGraw-Hill, Boston.
Hair, J. F. et al. (1998). *Multivariate Data Analysis*, Prentice-Hall International, London.
Harrell, F. E. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*, Springer, New York.
Hartwig, F. & Dearing, B. E. (1979). *Exploratory Data Analysis*, Sage, Newberry Park.
Hofmann, D. A. (1997). An Overview of the Logic and Rationale of Hierarchical Linear Models. *Journal of Management*, 23, 6, 723-744.
Howell, D. (2007). *The analysis of missing data*, Sage, London.
Hox, C. and Maas, J. M. (2002). Sample Sizes for Multilevel Modeling. In *Social Science Methodology in the Millennium: Proceedings of the Fifth International Conference on Logic and Methodology*, J. H. Jörg Blasius, Edith de Leeuw and Peter Schmidt, Leske & Budrich Verlag, Opladen.
Hox, J. (2010). *Multilevel Analysis, Techniques and Applications*. Routledge, Taylor & Francis Group, New York and Hove.
Jong, A. et al. (2004). Antecedents and Consequences of the Service Climate in Boundary-Spanning Self-Managing Service Teams. *Journal of Marketing*, 68, 2, 18-35.
Kaplan, D. (1995). *Statistical Power in Structural Equation Modeling*. Sage, Thousand Oaks.
Kleinbaum, D. G. et al. (1998). *Applied Regression Analysis and Other Multivariate Methods*, Duxbury, Pacific Grove.
Konradt, U. et al. (2009). Self-leadership in organizational teams: A multilevel analysis of moderators and mediators. *European Journal of Work and Organizational Psychology*, 18, 3, 322-346.
Kozlowski, S. W. J. and Klein, K. J. (2000). A Multilevel Approach to Theory and Research, in Organizations: Contextual, Temporal, and Emergent Processes. In *Multilevel Theory, Research and Methods in Organizations*, Foundations Extensions, and New Directions, W.J. Kozlowski, ed., Jossey-Bass, San Francisco.
Leeuw, J. and Kreft, I. (1986). Random Coefficient Models. *Journal of Educational Statistics*, 11, 55-85.

Lenth, R. V. (2001). Some Practical Guidelines for Effective Sample Size Determination. *The American Statistician*, 55, 187-193.
Liao, H. and Chuang, A. (2004). A Multilevel Investigation of Factors Influencing Employee Service Performance and Customer Outcomes. *Academy of Management Journal*, 47, 1, 41-58.
Little, R. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd edn, Wiley, New York.
Longford, N. (1993). *Random coefficient models*. Clarendon Press, Oxford.
MacKenzie, S. B. (2001). Opportunities for Improving Consumer Research Through Latent Variable Structural Equation Modeling. *Journal of Consumer Research*, 28, 1, 159-166.
McKee, A. J. et al. (2001). The Graduate Record Examination and undergraduate grade point average: Predicting graduate grade point averages in a Criminal Justice graduate program. *Journal of Criminal Justice Education*, 12, 311-317.
Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105.
Moineddin, R. et al. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology*, 7, 1-10.
Osborne, J. W. (2002). Notes on the use of data transformation. Practical assessment. *Research and Evaluation,* 8.
Pieters, R. and Wedel, M. (2004). Attention Capture and Transfer in Advertising: Brand, Pictorial, and Text-Size Effects. *Journal of Marketing*, 68, 2, 36-50.
Rasbash, J. et al. (2000). *A User's Guide to MLwiN. Centre for Multilevel Modeling*, Institute of Education, University of London.
Raudenbush, S. W. & Bryk A.S. (2002). *Hierarchical linear models*. Thousand Oaks.
Rovezzi, C. S. and Carroll, D. J. (2002). *Statistics Made Simple for School Leaders*, Rowman & Littlefield, Lanham.
Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*, Chapman & Hall, New York.
Sharma, S. (1996). *Applied Multivariate Techniques*, John Wiley & Sons, New York.
Van Bruggen, G. H. et al. (2002). Informants in Organizational Marketing Research: Why Use Multiple Informants and How to Aggregate Responses. *Journal of Marketing Research*, 39, 4, 469-478.
West, S. G. et al. (1995). *Structural Equation Models with Non-Normal Variables: Problems and Variables,* Sage, Thousand Oaks, CA.
Wieseke, J. et al. (2008). Multilevel Analysis in Marketing Research: Differentiating Analytical Outcomes. *Journal of Marketing Theory and Practice*, 16, 321-339.
Wild, M. R. et al. (2004). Can psychological factors help us to determine adherence to CPAP? A prospective study. *European Respiratory Journal*, 24, 461-465.