Yield Analysis and Mixed Model

Eugene Demidenko[1] and Igor Mandel[2]
[1]Dartmouth College, 7927 Rubin, DHMC, Lebanon, NH 03756
eugened@dartmouth.edu
[2]Media Planning Group, 195 Broadway, New York, NY 10007
igor.mandel@mpg.com

Abstract

A concept of yield analysis, an approach for estimating an observation-specific prediction of the dependent variable, has been developed. A relationship with Hildretch-Houck regression with random coefficients is established. Two types of estimation are considered: (1) distribution-free that is based on variance least squares and weighted least squares and (2) maximum likelihood that uses normal assumption; not normally distributed regression coefficients are considered as well. After fixed effect coefficients are estimated, observation-specific predictions are computed as posterior means using the BLUP approach. The mixed model implies the yield model when the variance of random coefficients goes to infinity and the design matrix is orthogonal. Monte Carlo simulation results are presented proving models high efficiency. Case study results are shown, where the application of the proposed technique to a car advertising data demonstrates a high predictability while standard regression fails with weak correlations between cost of ad and car sales.

Keywords: random coefficients, Hildretch-Houck regression, mixed model, variance least squares, Monte Carlo simulation

1. Introduction

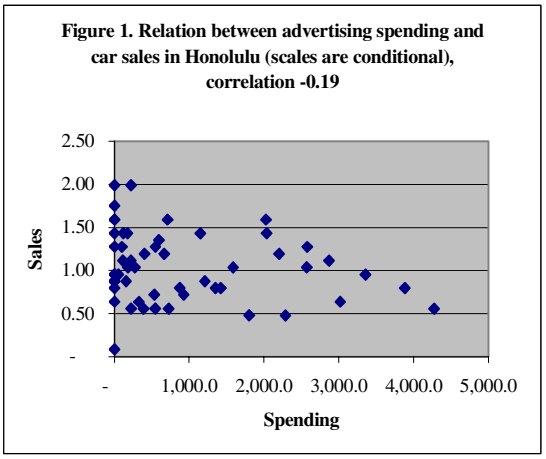
Linear regression with random coefficients is a developed statistical topic. Perhaps the earliest authors are (Hildreth and Houck 1968) and (Swamy 1971). A comprehensive account of available estimation methods was presented in (Raj 1975); the latest results for maximal likelihood estimation are obtained in (Zaman 2001). On the other hand, the concept of yield analysis, an approach where the main goal is to estimate individual coefficients of influence ("yields") on each object of the population, was presented (with heuristic estimation algorithm) in (Mandel 2003) and applied to times series problems in (Mandel and Hauser 2005 1,2). Here we merge the two approaches and stress that regression with random coefficients may be viewed as a mixed effects model where the size of the cluster is one. As such the theory of regression with random coefficients may be further advanced using recent developments in mixed models (Demidenko 2004). While previous works on random coefficients was concentrated on estimation of fixed effects, we focus on estimation of the random effects (yields) – the most interesting part from the application point of view, which also have a great importance in many if not all regression-like applications.

In the second part of this article we give a real life example, demonstrating a typical problem which may occur in analysis and cannot be solved with traditional means; in parts 3-5 all of the main results and proofs are given; part 6 describes results the of a Monte Carlo simulation; part 7 gives details of practical implementation and recommendations; the conclusion summarizes results and points out the prospective areas for use and development of the approach.

2. Motivating Example

Companies spend billions of dollars on advertising. Does advertising increase sales? We show the scatter-plot of sales versus cost of advertising of a large car manufacturing company in Fig. 1. The correlation coefficient –0.19 assumes a negative effect of advertising to sales (!); around 6% sales should be "explained" by that "advertising". Thus, according to the standard statistical test the answer is "Not really." Why do businessmen spend money in vain? It's possible, in principle, but not really plausible on a large scale - controversial pictures like presented on a Fig.1 are not unusual for many situations. What may be wrong?

Indeed, the affect of advertising on sales is a complicated process. Many things are going on, such as advertisement of the same type of product by competitors, bad economical situation and so on. The hypothesis that the rate at which ads affect sales is not constant over time and varies within a week or even a day looks not only possible, but, in fact, the only viable. And this is exactly what we try to capture and what may explain picture like that and many others.



Traditional regression-like models fail in doing that. To account for uncontrollable factors random effects may be introduced that leads to regression with random coefficients, but with estimation of those on each object (or moment of time). Then the possible explanation will be, that advertising works, but it's efficiency changes over time in such a way, that bigger spending bring less per dollar spent (still positively affecting sales).

3. Regression Model with Random Coefficients

In this section we describe how yields can be studied via regression model with random coefficients, namely,

$$y_i = \beta_1 + b_{i1}x_{i1} + ... + b_{im}x_{im} + \varepsilon_i$$

where formally $x_{i0} = 1$.We can combine the intercept β_1 with regression error ε_i , so intercept may be treated random as well:

$$b_{i1} = \beta_1 + \delta_{i1}, ..., b_{im} = b_{im} + \delta_{im}$$

where δ_{ij} are independent and

$$E(\delta_{im}) = 0, E(\delta_{ij}^2) = \sigma_j^2$$

or

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_m x_{im} + u_i , \qquad (1)$$

where

$$E(u_i) = 0, E(u_i^2) = \sigma^2 + \sigma_1^2 x_{i1}^2 + ... + \sigma_m^2 x_{im}^2 .$$

Throughout the paper we denote $z_{ij} = x_{ij}^2$; when $\sigma_j^2 = const$, we have $y_i = \vec{b}_i' \vec{x}_i$, where $\vec{x}$ denotes vector, $\vec{X}$ - matrix.

Under normal assumption

$$y_i \sim N(\vec{\beta}\vec{x}_i, \sum_{j=0}^{m} \sigma_j^2 x_{ij}^2), i = 1,2,...n . \qquad (2)$$

Parameters σ_j^2 are called variance parameters.

If they are known we apply weighted least squares

$$\hat{\vec{\beta}} = (\vec{X}'\vec{S}^{-1}\vec{X})^{-1}\vec{X}'\vec{S}^{-1}\vec{y}, \qquad (3)$$

where $\vec{S}_{n\times m}$ is a matrix, where diagonal elements are $\sum_{j=0}^{m} \sigma_j^2 x_{ij}^2$ and the off-diagonal elements are zeroes.

4. Methods of Estimation

It is interesting to know the limit of the GLS estimator when one of the variances goes to infinity. Lets pick *k* from 1, 2, …,m and put σ_k^2 k to infinity so that other variances stay bounded, so that $\lim_{\sigma_k^2 \to \infty} (\sigma_j^2 / \sigma_k^2) = 0.$ Letting variances go to infinity we obtain $\tilde{\tilde{\beta}}_k = \lim_{\sigma_k^2 \to \infty} \hat{\tilde{\beta}} = (\vec{X}'\vec{Z}_k^{-1^{-1}}\vec{X})^{-1}\vec{X}'\vec{Z}_k^{-1^{-1}}\vec{y}$,

where $\vec{Z}_k = diag(x_{1k}^2, x_{2k}^2, ..., x_{nk\,i1}^2)$.

Two types of estimation methods are discussed in this section. The first type, which we call variance least squares, does not assume a distribution for random coefficients and therefore is distribution free. The second type is maximum likelihood (ML) estimation that assumes that coefficients are normally distributed. We discuss standard and restricted ML estimation.

4.1 Variance Least Squares

The ordinary least squares (OLS) estimator $\hat{\vec{\beta}}_{OLS} = (\vec{X}'\vec{X})^{-1}\vec{X}'\vec{y}$ is an unbiased and consistent estimator of $\vec{\beta}$ .Therefore, at least for large sample, we may expect that the OLS residual $\hat{u}_i = y_i - \vec{x}_i\hat{\vec{\beta}}_{OLS}$ is a satisfactory estimate of the error term u_i in regression (1). Consequently, $\hat{u}_i^2$ may be treated as an empirical estimate of variance u_i which is $\sum_{j=0}^{m} \sigma_j^2 x_{ij}^2$, the theoretical variance. Hence we can find an estimate for the variance parameters which minimizes the Euclidean distance between the empirical and theoretical variances as

$$\sum_{i=1}^{n} (\hat{u}_i^2 - \sum_{j=0}^{m} \sigma_j^2 x_{ij}^2)^2 . \qquad (4)$$

Apparently, (4) is the sum of squares in regression of squared residuals on squared explanatory variables with the least squares solution

$$\hat{\vec{\sigma}}^2_{VLS} = (\dot{\vec{X}}\dot{\vec{X}})^{-1}\dot{\vec{X}}\dot{\hat{\vec{u}}}, \quad (5)$$

where, following the notation of (Hildreth and Houck 1968) and (Raj 1975), we use a dot over the vector or matrix to indicate that the elements are squared. This estimator was derived by the first authors - we call it variance least squares (VLS) estimator because this name is self-explanatory (also this method of variance estimation was used in (Demidenko 2004) in a more general setting).

The VLS estimator is biased because the expected value of $\hat{u}_i^2$ is not exactly equal to $\sum_{j=0}^m \sigma_j^2 x_{ij}^2$.

To find the unbiased VLS estimator we observe that the covariance matrix of the OLS residual vector is $\mathrm{cov}(\hat{\vec{u}}) = \vec{P}\vec{S}\vec{P}$, where $\vec{P} = \vec{I} - \vec{H}$ and $\vec{H} = \vec{X}(\vec{X}'\vec{X})^{-1}\vec{X}'$ is the projection or hat matrix. Matrix $\vec{H}$ emerges in connection with outliers detection; the diagonal element of this matrix is the called leverage value (Huber, 1981). The sum of squared elements of the difference between empirical and theoretical covariance matrices takes the form $tr(\hat{\vec{u}}\hat{\vec{u}}'\vec{P}\vec{S}\vec{P})^2$, which is minimized over $\vec{\sigma}^2$. After some algebra we arrive at the unbiased VLS (UVLS) estimator

$$\hat{\vec{\sigma}}^2_{UVLS} = (\dot{\vec{X}}'\dot{\vec{P}}\dot{\vec{X}})^{-1}\dot{\vec{X}}\dot{\hat{\vec{u}}}. \quad (6)$$

One may iterate and after variance parameters are estimated plug them in (3) to obtain new residuals and continue until convergence. This type of estimation is called iterative VLS.

4.2 Maximum Likelihood

Now we assume that random coefficients are normally distributed so that the model is written as (2). The log-likelihood function, up to a constant term is given by

$$l(\vec{\beta},\vec{\sigma}^2) = -\frac{1}{2}\left[\frac{1}{\sum_{j=0}^m \sigma_j^2 x_{ij}^2}(y_i - \vec{\beta}\vec{x}_i)^2 + \ln(\sum_{j=0}^m \sigma_j^2 x_{ij}^2)\right] \quad (7)$$

Then if σ^2 is held fixed l is maximized at the weighted least squares with the weights $w_i = 1/\sum_{j=0}^m \sigma_j^2 x_{ij}^2$ or in matrix form (3). There is no closed form solution for minimization over $\vec{\sigma}^2$ when $\vec{\beta}$ is held fixed, so we need to do iterations. For this we compute the first and second derivatives of the log-likelihood function:

$$\frac{\partial l}{\partial \sigma_k^2} = -\frac{1}{2}\sum_{i=0}^n \left[-\frac{e_i^2 x_{ik}^2}{(\sum_{j=0}^m \sigma_j^2 x_{ij}^2)^2} + \frac{x_{ij}^2}{\sum_{j=0}^m \sigma_j^2 x_{ij}^2}\right] \quad (8)$$

and the Hessian with the (k,k') th element

$$\frac{\partial^2 l}{\partial \sigma_k^2 \partial \sigma_{k'}^2} = -\frac{1}{2}\sum_{i=1}^n \left[-\frac{2e_i^2 x_{ik}^2 x_{ik'}^2}{(\sum_{j=0}^m \sigma_j^2 x_{ij}^2)^3} - \frac{x_{ik}^2 x_{ik'}^2}{(\sum_{j=0}^m \sigma_j^2 x_{ij}^2)^2}\right].$$

We take the expectation of the Hessian to compute the Fisher information matrix. After some algebra we obtain that the information matrix for $\vec{\sigma}^2$ is given by

$$\vec{\Im}_{\vec{\sigma}^2} = \frac{1}{2}\left[\sum_{i=0}^n \frac{x_{ik}^2 x_{ik'}^2}{(\sum_{j=0}^m \sigma_j^2 x_{ij}^2)^2}\right], k,k' = 1,2,..,m\text{H}$$

ence, the Fisher scoring algorithm for the variance parameters takes the form

$$\vec{\sigma}^2_{s+1} = \vec{\sigma}^2_s - (\vec{\Im}_{\vec{\sigma}^2})^{-1}\frac{\partial l}{\partial \vec{\sigma}^2}, s = 0,1,...$$

Note that the information matrix for $\vec{\beta}$ and $\vec{\sigma}^2$ is block-diagonal which means that (i) the log-likelihood may be maximized alternating between $\vec{\beta}$ and $\vec{\sigma}^2$ maximization, (ii) the MLE for coefficients and variance parameters are independent, (iii) any consistent estimator of $\vec{\sigma}^2$ leads to an efficient estimator of $\vec{\beta}$ in the estimated weighted least squares.

The inverse of information matrix can serve as an estimate of covariance matrix for variance parameters and appropriate hypothesis testing.

4.3 Restricted Maximum Likelihood

It is well known that in a small sample the maximum likelihood estimation of variance parameters is biased. In particular, if only intercept is random (classic regression) the MLE, $\hat{\sigma}^2 = RSS/n$ is negatively biased. To obtain an

unbiased estimator $\hat{\sigma}^2 = RSS/(n-m)$ the restricted MLE should be used. Following the line of derivation for linear mixed model (Demidenko 2004, p. 58) we obtain that the restricted log-likelihood differs from the usual log-likelihood (7) by term

$$T = -\frac{1}{2}\ln\left|\vec{X}'\vec{S}\vec{X}\right| = -\frac{1}{2}\ln\left|\sum_{i=1}^{n} \vec{x}_i\vec{x}_i'\sum_{j=0}^{m}\sigma_j^2 x_{ij}^2\right| =$$

$$= -\frac{1}{2}\ln\left|\sum_{j=0}^{m} E_j\sigma_j^2\right|,$$

where the $n\times n$ matrices $\vec{E}_j$ are defined as

$$\vec{E}_j = \sum_{i=1}^{n} x_{ij}^2\vec{x}_i\vec{x}_i', \; j = 1,2,...,m.$$

To compute the RMLE by Fisher scoring we need the first and second derivatives which are

$$\frac{\partial T}{\partial\sigma_k^2} = -\frac{1}{2}tr(\vec{E}_k\vec{E}^{-1}),$$

$$\frac{\partial^2 T}{\partial\sigma_k^2\partial\sigma_{k'}^2} = -\frac{1}{2}tr(\vec{E}_k\vec{E}^{-1}\vec{E}_{k'}\vec{E}^{-1}), \quad (9)$$

where $\vec{E} = \sum_{j=0}^{m} \vec{E}_j\sigma_j^2$. Then the first and second derivatives of the standard log-likelihood function, *l* are augmented by (8) and Fisher scoring algorithm applies.

5. Hypothesis Testing

We can estimate the covariance matrix as

$$\mathrm{cov}(\hat{\vec{\sigma}}^2{}_{VLS}) = \hat{\tau}^2(\dot{\vec{X}}'\dot{\vec{X}})^{-1},$$

where $\hat{\tau}^2 = RSS/(n-m)$ and RSS is the residual sum of squares, the minimum value of (4). For example, we reject $H_0: \sigma_j^2 = 0$ if $(\hat{\vec{\sigma}}^2{}_{VLS})_j > \hat{\tau}\ (\dot{\vec{X}}'\dot{\vec{X}})^{-1}_{jj} q_{1-\alpha}$, where $q_{1-\alpha}$ is the $(1-\alpha)$ th quintile of the *t*-distribution with $n-m$ degrees of freedom. Note we use a one-tail hypothesis test. Analogously, we can apply UVLS to compute the covariance matrix using $(\dot{\vec{X}}'\dot{\vec{P}}\dot{\vec{X}})^{-1}$. As a word of caution, these tests are not exact because $\hat{u}_i^2$ is not normally distributed.

The test on randomness of regression coefficient is essential. Let *j* be any. We want to test that the *j*th coefficient is fixed (not random), in other words, $H_0: \sigma_j^2 = 0$. We use score test. We observe that H_0 would be reasonable to reject if $\partial l/\partial\sigma_j^2 > 0$, because under H_0 we have $\partial l/\partial\sigma_j^2 = 0$, where the derivative is given by (8).

6. Estimation of Random Coefficients

After estimates for fixed effects coefficients $\vec{\beta}$ and variance parameters $\vec{\sigma}^2$ are obtained we may estimate random coefficients $\vec{b}_i$. In essence estimation of $\vec{b}_i$ is equivalent to estimation of random effects in a mixed effects model. Also, it should be noted that *estimation* is understood not in a classic sense but as a prediction. There are two equivalent approaches to estimate $\vec{b}_i$ assuming that variance parameters are known.

First, we can estimate $\vec{b}_i$ as conditional or *posterior mean* following the Bayesian paradigm. Indeed, the regression with random coefficients may be viewed as conditional model for $y_i|\vec{b}_i$. We may reverse this to obtain the posterior distribution $\vec{b}_i|y_i$, particularly the posterior mean, $E(\vec{b}_i|y_i)$. To obtain $E(\vec{b}_i|y_i)$ we observe that $(y_i, \vec{b}_i)$ have a *(1+m)* multivariate normal distribution with mean and covariance matrix

$$\vec{\mu} = \begin{bmatrix} \vec{\beta}'\vec{x}_i \\ \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_m \end{bmatrix}, \; \vec{\Omega} = \begin{bmatrix} \sum_{j=0}^{m}\sigma_j^2 z_{ij}...\sigma_1^2 x_{i1}...\sigma_m^2 x_{im} \\ \sigma_1^2 x_{i1}.....0.....0 \\ \sigma_1^2 x_{i2}.....\sigma_1^2.....0 \\ .....0.....0 \\ .....0.....0 \\ \\ \sigma_m^2 x_{im}.....0.....\sigma_m^2 \end{bmatrix}$$

Using standard formula of multivariate distribution we obtain

$$b_{ij} = E(b_{ij}|y_i) = \beta_j + \frac{\sigma_j^2 x_{ij}}{\sum_{j=0}^{m}\sigma_j^2 z_{ij}}(y_{ij} - \vec{\beta}\vec{x}_i) =$$

$$= \beta_j + \frac{\sigma_j^2 x_{ij}}{\sum_{j=0}^{m}\sigma_j^2 z_{ij}} e_i \quad (10)$$

Second, we can use *penalized least squares* (Demidenko 2004, p. 15):

$$\frac{(y_i - \vec{b}_i' \vec{x}_i)^2}{\sum_{j=0}^{m} \sigma_j^2 z_{ij}} + \sum_{j=0}^{m} \frac{(b_{ij} - \beta_i)^2}{\sigma_j^2} \Rightarrow \min \quad (11)$$

Differentiating with respect b_{ij} we obtain

$$-2\frac{e_i x_i}{\sum_{j=0}^{m} \sigma_j^2 z_{ij}} - 2\frac{(b_{ij} - \beta_j)}{\sigma_j^2} = 0,$$

which gives estimate (10) for b_{ij}.

It is worth to note that two equal ways of estimating of coefficients (10) and (11) show once again the relativity of intensive disputes between "frequentists" and 'Bayesians", since two approaches very often give just the same results depending on of minimization functional, see details in (Demidenko 2004).

6.1 Monte Carlo Simulation

A simulation was performed to compare generated yields and those estimated by the algorithm. Each data set had the same size (240 data points), with one or three factors, and was generated by a formula (1), where *X* was taken as a uniformly distributed variable and then kept fixed during experiments. Yields were estimated by (10) with variance estimation by (5). The effects of five parameters were analyzed.

1. Error was normally distributed random variable with different variance and zero mean. Error's level was controlled as a ratio of its variance to the standard deviation of *Y* without an error. It took levels 0, 10%, 20%, and 40%.

2. Correlation between factors was set on zero and high (0.60-0.70) levels.

3. Correlation between factor and its yields was either zero or significant (about 0.5-0.6) for one factor and zero for two others.

4. Generated yields (*G*) level of variation was set on low (5%), medium (22%-25%), and high (100%) levels for all factors.

5. Different yields variations were obtained by keeping first factor's yields with high variation, while others are with medium.

For each parameters combination 1000 random data sets were generated, where fixed values of factors were assigned to random values of yields. If variance estimation in (5) becomes negative, then the yields were not calculated, what gave about 600-750 yields sets for each setting. Part of the results for three factors is presented in Table 1 in a form of average values for all runs.

The quality of the yields approximation was measured by two statistics: root mean square error (RMSE, difference between actual and estimated yield values) was calculated as a ratio of RMSE and generated yields mean; correlation between generated and estimated yields. The quality of the entire model is described by determination coefficient R^2. The main findings are summarized below.

1. Determination after yield analysis everywhere is higher than after regression, which stresses the fact that the low original correlation may hide the actual presence of dependence. For example, at row 3, where no error is added to Y, the regression shows a weak dependence (24%), whereas yield analysis gives a 98% determination. It corresponds with a fact that outcome, indeed, depends entirely on those factors weighted by respective yields. Examples like this illustrate the possibility of two different scenarios in reality, what usually are inseparable by statistical methods: low determination may be explained by high levels of random noise ("environmental immeasurable influence") or by high varying yields. Depending on the model, interpretation of the results could be completely different (rows 3,8).

2. Correlation between generated and derived yields has moderate values, lowering from 0.5 to 0.3 while error rises from zero to 40%, regardless on yields variation (rows 1-3). It is hard to expect a much higher correlation, keeping in mind the complexity of the problem (for one factor model it ranges, however, from 0.7 to 0.9). But a fact of positive correlation is very important and tells one that some improvement vs. regression estimation is always guaranteed.

3. Yields variation does not practically affect correlation between generated and estimated yields only if all factors have yields with the same variation level. It's not true when equality does not hold: a factor with highest variation is recovered much better – compare 0.6 vs. 0.05-0.23 in rows 4,9,14, regardless on level of error in data (shadowed rows). Its worth to note that in general a recovering of yields with high variation is more important than with low one, since the latter will be captured by usual regression and does not significantly affect a model quality.

Table 1. Results of simulation, 1,000 runs for each setting

	Error	Factor-factor correlation	Yields-factor correlation	Generated yields variation, %	RMSE, %			Correlation between generated G and estimated b yields			Regression R^2, %	Yields R^2, %
					G 1	G2	G 3	G 1	G2	G 3		
1	No	No	No	5	4	4	4	0.55	0.52	0.53	99	100
2	No	No	No	25	21	22	22	0.56	0.53	0.53	82	99
3	No	No	No	100	88	89	91	0.56	0.53	0.51	24	98
4	No	No	No	100	204	82	120	0.63	0.06	0.06	44	76
5	10%	No	No	25	21	21	22	0.54	0.51	0.51	82	99
6	20%	No	No	25	22	23	23	0.45	0.43	0.41	77	97
7	40%	No	No	25	24	24	24	0.33	0.33	0.32	69	90
8	10%	Yes	No	100	124	131	93	0.13	0.25	0.70	23	97
9	20%	Yes	No	100	119	85	61	0.60	0.05	0.23	65	93
10	No	Yes	No	25	30	30	21	0.14	0.26	0.69	82	99
11	No	Yes	No	25	32	35	23	0.11	0.27	0.68	85	99
12	No	No	Yes	25	28	22	22	0.48	0.53	0.52	86	99
13	20%	No	Yes	25	30	24	23	0.39	0.42	0.44	81	97
14	20%	No	Yes	100	103	111	59	0.65	0.13	0.14	60	85

(*) Shadow means that high variation was set only for the first factor

4. While leaving correlations almost untouched, yields variation definitely decreases the level of approximation: RMSE is raised from 4% to around 80% with rise of variation (rows 1-3), and goes even to 200% for the first factor with high variation, while correlation of its yields with real ones is high (row 4). I.e., the higher correlation, the lower approximation, those two quality statistics work typically in opposite directions. Indirectly, it is reflected in a fact, that variations of yield estimates are roughly in 2 times smaller than variation of real yields (not shown in a table). All that means, that yields are better to use as directional range values, yet discovering new aspects of a model.

5. The rows 12-14 represent the results for yields, correlated with factor values, i.e. may reflect different patterns in data, like dynamic trends, nonlinearity, and so on. As seen, the level of recovery is almost the same as in situation without correlations, as in rows 6,7, 9 and alike. It means, that those hidden patterns in data may be revealed in an automatic way, without special hypothesizing. Indeed, in real data we found several times, that yields are correlated with a factor, what adds a lot to the process understanding. For example, it was discovered that big volume of advertising typically positively correlated with its outcomes per unit (Mandel and Hauser 2005 1,2).

The short summary of different considered factors could be found in Table 2, which shows the correlations between average values of those factors and average values of two dependant variables. The correlations are obtained from the table, analogous to Table 1, where instead of “No” and “Yes” are the real average values of simulation parameters. Of course, it is just a rough justification of the process, but it still gives sufficient understanding of what is going on. The mutual correlations between those five factors are quite low, which allowed calculating the regression. The determination coefficients are provided in the last row of the Table 2.

As one may see, the biggest negative impact to correlation between generated and estimated yields is correlation between factors, the biggest positive effect – relative variation (see comment 3 above). The first fact is an indirect reflection of the multicolinearity problem – unfortunately, yield analysis does not provide a remedy for that by itself (though it helps a lot, because it shifts the problem from correlation of factors to correlation of yields and other effects). The second one gives a hint, that if there are some preliminary ideas about the

variance of yields, it helps to estimate a reliability of yields obtained – the best ones would be those with high relative variance. But since estimates of variance are already obtained (see section 4), it could be used to answer that question directly.

Table 2. Correlations of factors with characteristics of quality of yields approximation

Factors	Correlation between generated and estimated yields	RMSE
Error	(0.03)	(0.24)
Correlation between factors	(0.74)	0.27
Yields variation	0.04	0.90
Correlation between factor and yields	0.03	(0.18)
Relative high variation of factor's yields	0.54	0.55
Regression determination,%	92%	91%

Characteristically, the relative yields variance spoils RMSE with the same strengths as it improves the correlation. It means, that the "optimal" solution is possible only for more or less equally variable yields. The level of RMSE is strongly affected by yield absolute variation (correlation 0.9), in sharp contrast to that for correlation (0.04). Together with the previous conclusion, this fact allows to estimate the level of reliability of yields estimation after calculation of yields variances. In general, listed parameters of data generation explain more than 90% of quality statistics variation. Since three of the most important (correlation between factors, yields absolute and relative variation) are to be known after calculations, those empirical findings from simulation together with direct estimation of individual yields allow to say how close are the estimates to real yields.

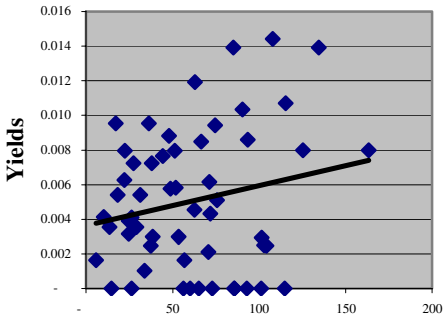
7. Practical Implementations

On materials of one large car manufacture the yield analysis was performed, some results of which are presented below. On Fig. 2 the yields of advertising are shown in dynamics over time (months); the scales are conventional.

It shows, that the efficiency of an ad is very different, with a slight negative trend. The next step could be, presumably, to understand why – for example, to see what exact content of an ad was every time, what time of the day was it aired, and so on – i.e., make a special model, where yields play a role of *dependent* variable.

In Fig. 2 the same yields are shown in relation with amount of money spent for advertising every month. The same type of conclusions may be drawn from here.

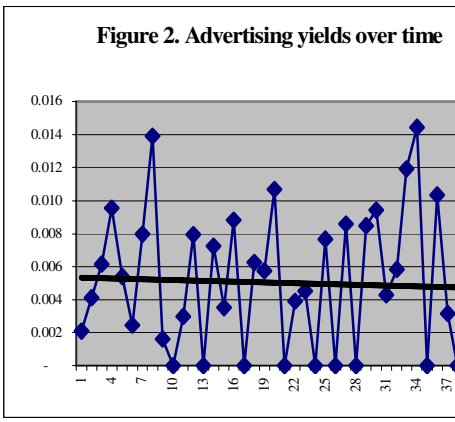
Figure 3. Advertising and its efficiency



In spite of slight positive correlation, in many instances a lot of money was spent in vain, even with zero result. Together with dynamic data like shown on Fig.2 this type of analysis provides managers with new very effective tool to operate business. Several general recommendations may be made about yields implementation from statistical and business points of view.

a) The level of ***data approximation*** in yield analysis is always higher than in traditional

Figure 2. Advertising yields over time



regression-like models; therefore it allows one to create plausible models in a much broader range of real life situations, including those with low correlations between the outcome and factors.

b) ***The prediction of the outcome*** for new data points, which is straightforward in

regression, is less obvious here since yields are not constant over space or time. The best way of doing this is to assign new factor values to new yields, based on forecasted trends, like shown in Fig.2.

c) If all yields are positive it is possible to calculate direct ***contributions*** of a factor to outcome for each data point what is very important (a problem (1) should be modified respectively by adding constraints).

d) Yields add a new dimension to data analysis. Particularly, one may do ***cluster analysis*** in yield space and find zones of high and low yields for many variables, which will differ from those based on the original values of the factors (see Fig. 3). So far the entire concept of ***homogeneity*** was based on clusters in the space of *X* or *Y* or *X* and *Y* together. But yields conceptually change it by introducing variables with new meaning. One may assume that yields for the given data set are not homogeneous and reformulate (1). On the same note, new models using yields as dependent variables may be built.

e) Yield analysis may be used for ***dynamic data*****.** This is a special topic, associated with concept of prolonged efficiency, which is partially covered in (Mandel and Hauser, 2005, 1,2).

8. Conclusion

For the first time an exact estimations of the random coefficients in regression model are obtained by several methods and its relation with yield analysis concept is shown. A precision and reliability of random coefficients estimates are checked through the simulation experiment with combinations of most influential parameters. Recommendations about relative importance of those factors are given; particularly, it is found that level of yields recovery strongly depends on yields variation and factors correlation. The importance of yield analysis was demonstrated on a level of practical implementation. The importance of yield analysis as a generic statistical tool was emphasized. It allows making such new things as to determine individual contributions of factors, differentiate zones of high and low influence of all factors, make clustering in new space, create new type of dynamic models and more. Two facts make yield analysis a universal tool for many practical applications. First, the model with random coefficients by its nature is much more natural than model only with fixed effects. Second, a data set for classical mixed model should contain several subsets of clustered data, which is not always available, whereas for yield analysis the structure of data set is absolutely typical (ordinary tables "object–variables" or "time-variables"), i.e. it may be used wherever regression and other traditional methods may. It allows hoping for wide use of the proposed approach. However, not all theoretical properties of yields are investigated yet, and advanced algorithms with better yields estimates are to come.

References

Demidenko, E. (2004). Mixed Models: Theory and Applications. *Hoboken, NJ: Wiley.*

Hildreth, C. and Houck, J.P. (1968). Some estimators for linear model with random coefficients. Journal of the American Statistical Association 63, 584-595.

Huber, P.J. (1981). Robust Statistics. *New York: Wiley.*

Raj, B. (1975). Linear regression with random coefficients: the finite sample and convergence
properties. Journal of the American Statistical Association 70, 127-137.

Swamy, P.A.V.B. (1971). Statistical Inference in Random Coefficient Regression Models. *New York: Springer-Verlag.*

Zaman A. (2001), Maximum likelihood estimates for the Hildreth–Houck random coefficients model, *Econometrics Journal,* volume **5**, pp. 1–26.

Mandel I. (2003) Multicollinearity problem in marketing studies. *Joint Statistical Meeting. Abstracts. Section Statistics in Marketing. San Francisco, p.14*

Mandel I., Hauser D. (2005,1) Object Varying Coefficients in Stochastic Environments:

Yield Analysis to Model Marketing Efficiency. *The International Symposium on Stochastic Models in Reliability, Safety, Security and Logistics, Beer Sheva, Israel, pp.243-246.*

Mandel I., Hauser D. (2005,2) Yield Analysis and Return on Investment Estimation. *Joint Statistical Meeting Proceedings, Minneapolis, this issue.*