# Predictions in mixed-effects models

Štefan Varga

Department of Mathematics
Faculty of Chemical and Food Technology STU,
Radlinského 9, 812 34 Bratislava
E-mail : stefan.varga@stuba.sk

**Abstract**. We consider models of indirect measurements with mixed effects and their applications in chemical and food technology. Predictions of response variables are obtained using the maximum likelihood method.

The classical regression model of indirect measurements is usually studied in the form

$$y = a_1 f_1(x) + a_2 f_2(x) + ... + a_m f_m(x) \qquad (1)$$

where $f_i(x)$ $(i = 1, 2, ..., m)$ are known functions of the input variable $x$ (predictor), $y$ is an output variable (response) and the vector of unknown parameters $\theta = (a_1, a_2, ..., a_m)^T$ is not random but fixed. On the other hand mixed-effects models are regression models with a random vector of unknown parameters. They are used to analyze grouped data, repeated measures data or data impacted some factor variables. In the model of direct measurements, using analysis of variance, we can examine an influence of the factor variables on the values (expected values) of the measured variable. In the case of the model of indirect measurements an influence of the factor variables on the dependence of the observed variable $Y$ on the predictor $x$ is studied. Therefore it is reasonable to consider the unknown parameters $a_1, a_2, ..., a_m$ as sums of fixed and random parts.

The simplest mixed-effects model of indirect measurements could be

$$y_{ij} = (a_1 + a_{i1}) f_1(x_j) + (a_2 + a_{i2}) f_2(x_j) + ... + (a_m + a_{im}) f_m(x_j) + e_{ij} \qquad (2)$$

where $y_{ij}$ is a value of the observed variable $Y$, the level of the factor is $i$ $(i = 1, 2, ..., k)$ and the value of the predictor is $x_j$ $(j = 1, 2, ..., n)$. The values $a_1, a_2, ..., a_m$ are fixed components of the unknown regression coefficients and $a_{i1}, a_{i2}, ..., a_{im}$ are the random effects in the coefficients associated with the $i_{th}$ level of the factor. It is assumed that the vectors $\theta_i = (a_{i1}, a_{i2}, ..., a_{im})^T$ are independent $(i = 1, 2, ..., k)$ and identically distributed with $k$-dimensional normal distribution $N_k(0, \sigma^2 H)$ and that the $e_{ij}$ are errors of the measurements $y_{ij}$ independent and identically distributed with $N(0, \sigma^2)$ distribution.

The aim is to estimate both the fixed parts and the random parts of the unknown regression coefficients, to test statistical significance of the vector of random components of the unknown regression coefficients and to predict observed variable $Y$ for different level of the factor (factors) or for different subjects. We will present solutions of these problems using the example from the field of food technology.

In the process of wine fermentation the dependence of the capacity of the alcohol on the time was studied. One sort of wine fermented in 15 wine barrels. Regulated fermentation was in 8 barrels and in 7 barrels was unregulated fermentation. At the beginnings of the fifth, tenth, twentieth, thirtieth and fiftieth days of the fermentation was measured the capacity of the alcohol in all barrels. There is knew that each barrel has own life in these processes and therefore the regression coefficients (we considered exponential dependence with negative exponent) could be little different from barrel to barrel. It means that we can expect a random effect in the regression coefficients associated with the barrel (subject). We have the factor variable in our model of indirect measurement too. This is type of fermentation (regulated = 1, unregulated = 0). The data file contains 75 rows and 5 columns. The problem is to predict the capacity of the alcohol in any time and in any barrel.

| Obser. numb. | Barrel *i* | Regul *r* | Time *x* | Alcohol *y* |
|---|---|---|---|---|
| 1 | 1 | 1 | 5 | 4,41 |
| 2 | 1 | 1 | 10 | 6,68 |
| 3 | 1 | 1 | 20 | 9,5 |
| 4 | 1 | 1 | 30 | 10,43 |
| 5 | 1 | 1 | 50 | 10,93 |
| 6 | 2 | 1 | 5 | 4,27 |
| 7 | 2 | 1 | 10 | 6,77 |
| 8 | 2 | 1 | 20 | 9,77 |
| 9 | 2 | 1 | 30 | 10,51 |
| 10 | 2 | 1 | 50 | 10,93 |
| ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... |
| 66 | 14 | 0 | 5 | 6,4 |
| 67 | 14 | 0 | 10 | 8,67 |
| 68 | 14 | 0 | 20 | 10,39 |
| 69 | 14 | 0 | 30 | 10,47 |
| 70 | 14 | 0 | 50 | 11,07 |
| 71 | 15 | 0 | 5 | 6,47 |
| 72 | 15 | 0 | 10 | 8,47 |
| 73 | 15 | 0 | 20 | 9,82 |
| 74 | 15 | 0 | 30 | 10,79 |
| 75 | 15 | 0 | 50 | 11,41 |

Table 1. Data file

Generally we consider exponential regression model nonlinear in parameter but in the paper the parameter inside the exponent is known for us. The studied model is in the form

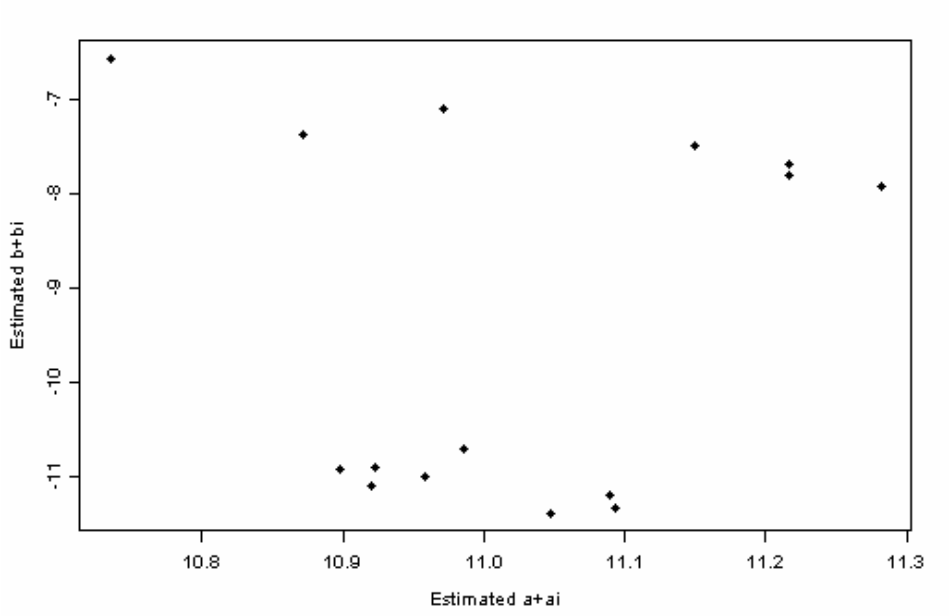
$$y_{ij} = (a + a_i) + (b + b_i)\, e^{-x_j/10} + e_{ij} \tag{3}$$

where $y_{ij}$ is the measured capacity of the alcohol in $i_{th}$ barrel $(i = 1, 2, ..., 15)$ in the time $x_j$ $(j = 1, 2, ..., 75)$, $a$, $b$ are the fixed parts of the unknown regression coefficients, $a_i, b_i$ are the random effects in the coefficients associated with the $i_{th}$ barrel and $e_{ij}$ is the random error of the measurement $y_{ij}$. Maximum likelihood method was used for estimations of the unknown coefficients in the model on assumption that $\theta_i = (a_i, b_i)^T$ are mutually independent $(i = 1, 2, ..., 15)$ and identically distributed with 15-dimensional normal distribution $N_{15}(0, \sigma^2 H)$ and errors $e_{ij}$ are independent too and distributed with normal distribution $N(0, \sigma^2)$. It is important to mention that the considered model (3) do not contain the factor variable „Regul“. This deficit will be repaired in the more general model (4).

The estimators (Table 2) of the fixed parts $a$, $b$ and the random parts $a_i$, $b_i$ ($i$ = 1, 2, ..., 15) in the model (3) shoved that the model with random effects is required in the presented example. The test of significance of the random effects confirmed this claim at the 0.999 confidence level. We can see it too in Picture 1 where the coefficients $a + a_i$ , $b + b_i$ ($i$ = 1, 2, ..., 15) are very different.

| **Estimators of Fixed and Random effects in the model (3)** | | | |
|---|---|---|---|
| $a$ | $a_i$ | $b$ | $b_i$ |
| 11.023455 | | -9.371134 | |
| | -0.03547066 | | -1.649722 |
| | -0.03564433 | | -1.657835 |
| | -0.03495688 | | -1.625816 |
| | -0.03893820 | | -1.811021 |
| | -0.02840248 | | -1.320988 |
| | -0.03310302 | | -1.539638 |
| | -0.03654262 | | -1.699563 |
| | -0.03909638 | | -1.818344 |
| | 0.04220953 | | 1.963241 |
| | 0.04028722 | | 1.873689 |
| | 0.03247228 | | 1.510325 |
| | 0.04216152 | | 1.960880 |
| | 0.04250486 | | 1.976914 |
| | 0.04289997 | | 1.995243 |
| | 0.03961916 | | 1.842635 |

Table 2. Estimators of the fixed and random parts of the coefficients in the model (3)




Picture 1. Estimators of the coefficients $a+a_i$, $b+b_i$

Table 3 contains the measured values of the capacity of the alcohol (observed values), the values fitted by the model with fixed regression coefficients (population fit) and the values fitted by the model with random effects (cluster fit). It is evident the difference in the two mentioned columns.

| **Observed and fitted values in the model (3)** | | | |
|---|---|---|---|
| Obs. num. | Observed values | Population fit | Cluster fit |
| 1 | 4.41 | 5.339575 | 4.303497 |
| 2 | 6.68 | 7.576008 | 6.933638 |
| 3 | 9.50 | 9.755210 | 9.496474 |
| 4 | 10.43 | 10.556894 | 10.439289 |
| 5 | 10.93 | 10.960313 | 10.913727 |
| 6 | 4.27 | 5.339575 | 4.298403 |
| 7 | 6.77 | 7.576008 | 6.930480 |
| 8 | 9.77 | 9.755210 | 9.495202 |
| 9 | 10.51 | 10.556894 | 10.438711 |
| 10 | 10.93 | 10.960313 | 10.913498 |
| ... | ... | ... | ... |
| 66 | 6.40 | 5.339575 | 6.592651 |
| 67 | 8.67 | 7.576008 | 8.352917 |
| 68 | 10.39 | 9.755210 | 10.068137 |
| 69 | 10.47 | 10.556894 | 10.699131 |
| 70 | 11.07 | 10.960313 | 11.016657 |
| 71 | 6.47 | 5.339575 | 6.496809 |
| 72 | 8.47 | 7.576008 | 8.293494 |
| 73 | 9.82 | 9.755210 | 10.044203 |
| 74 | 10.79 | 10.556894 | 10.688253 |
| 75 | 11.41 | 10.960313 | 11.012348 |

Table 3. Observed and fitted values in the model (3)

As we already mentioned, the model (3) did not contain the factor variable „Regul“ which sorts the observations in dependence on the fermentation (regulated fermentation = 1 unregulated fermentation = 0). The model that contained the variable „Regul“ we considered in the form

$$y_{ij} = (a + a_0 * \text{regul} + a_i) + (b + b_0 * \text{Regul} + b_i)\, e^{-x_j/10} + e_{ij}$$

The coefficient $a_0$ was not statistically significant and therefore the resultant model for our situation is

$$y_{ij} = (a + a_i) + (b + b_0 * \text{Regul} + b_i)\, e^{-x_j/10} + e_{ij} \qquad (4)$$

The analysis of variance confirmed statistically significant difference (p-value = 5,0384e-012) in the models (3) a (4). The estimators of the unknown coefficients of the model (4) are in the table 4.

| **Estimators of Fixed and Random effects in the model (4)** | | | | |
|---|---|---|---|---|
| $a$ | $a_i$ | $b$ | $b_i$ | $b_0$ |
| 11.023455 | | -7.341406 | | -3.805740 |
| | -1.061144e-007 | | 4.095338e-009 | |
| | 8.531391e-008 | | -3.496774e-009 | |
| | -1.383528e-007 | | 5.525367e-009 | |
| | -1.037273e-007 | | 2.917588e-009 | |
| | 1.951139e-007 | | -5.556958e-009 | |
| | 1.728419e-007 | | -6.149079e-009 | |
| | -2.329156e-007 | | 8.753335e-009 | |
| | -2.949271e-007 | | 1.039699e-008 | |
| | -3.490391e-007 | | 1.436148e-008 | |
| | 3.792574e-007 | | -1.491732e-008 | |
| | -5.081369e-007 | | 1.758294e-008 | |
| | 3.524593e-007 | | -1.327629e-008 | |
| | 7.867251e-009 | | 3.999526e-010 | |
| | 2.818044e-007 | | -1.026333e-008 | |
| | 2.585551e-007 | | -1.037324e-008 | |

Table 4. Estimators of the fixed and random parts of the coefficients in the model (4)

**Conclusion.** The aim of the contribution was to show certain possibilities of utilizations of mixed-effects models with booth, qualitative and quantitative predictors in the field of chemical and food technology. For reasonability of the paper we used only more simple models and only minimum graphical and numerical outputs from the software S-plus.

**Literature:**


[1] Härdle, W. : Smoothing Techniques with Implementation in S. New York, Springer - Verlag 1981.

[2] Krishnaiah, P.R., Sen, P.K. :Handbook of statistics 4, Nonparametric methods. Amsterdam, North Holland 1984.

[3] Venables, W.N., Ripley, B.D. : Modern Applied Statistics with S-plus. New York, Springer - Verlag 1994.

[4] Varga, Š. : Uncertainty in regression models. Proceedings of the Scientific Colloquium PRASTAN, Kočovce 2001, 154 – 160.