

# Module 9: Random coefficient models

9.1	Introduction . . . . .	1
9.2	Example: Constructed data . . . . .	3
9.2.1	Simple regression analysis . . . . .	3
9.2.2	Fixed effects analysis . . . . .	4
9.2.3	Two step analysis . . . . .	5
9.2.4	Random coefficient analysis . . . . .	6
9.3	Example: Consumer preference mapping of carrots . . . . .	8
9.4	Random coefficient models in perspective . . . . .	14

## 9.1 Introduction

Random coefficient models emerge as natural mixed model extensions of simple linear regression models in a hierarchical (nested) data setup. In the standard situation, we are interested in the relationship between  $x$  and  $y$ . Assume we have observations  $(x_1, y_1), \dots, (x_n, y_n)$  for a subject. Then we would fit the linear regression model, given by

$$y_j = \alpha + \beta x_j + \epsilon_j$$

Assume next that such regression data are available on a number of subjects. Then a model that expresses different regression lines for each subject is expressed by:

$$y_{ij} = \alpha_i + \beta_i x_{ij} + \epsilon_{ij}$$

or using the more general notation:

$$y_i = \alpha(\text{subject}_i) + \beta(\text{subject}_i)x_i + \epsilon_i \tag{9.1}$$

This model has the same structure as the different slopes ANCOVA model of the previous module, only now the regression relationships are in focus. Assume finally that the interest lies in the average relationship across subjects. A commonly used “ad hoc” approach is to employ a two-step procedure:

1. Carry out a regression analysis for each subject.
2. Do subsequent calculations on the parameter estimates from these regression analyzes to obtain the average slope (and intercept) and their standard errors.

Since the latter treats the subjects as a random sample, it would be natural to incorporate this in the model, by assuming the subject effects (intercepts and slopes) to be random:

$$y_i = a(\text{subject}_i) + b(\text{subject}_i)x_i + \epsilon_i$$

where

$$a(k) \sim N(\alpha, \sigma_a^2), b(k) \sim N(\beta, \sigma_b^2), \epsilon_i \sim N(0, \sigma^2)$$

and where  $k = 1, \dots, K$  with  $K$  being the number of subjects. The parameters  $\alpha$  and  $\beta$  are the unknown population values for the intercept and slope. This is a mixed model, although a few additional considerations are required to identify the typical mixed model expression. The expected value is

$$E y_i = \alpha + \beta x_i$$

and the variance is

$$\text{Var} y_i = \sigma_a^2 + \sigma_b^2 x_i^2 + \sigma^2$$

So, an equivalent way of writing the model is the following where the fixed and the random part is split:

$$y_i = \alpha + \beta x_i + a(\text{subject}_i) + b(\text{subject}_i)x_i + \epsilon_i \quad (9.2)$$

where

$$a(k) \sim N(0, \sigma_a^2), b(k) \sim N(0, \sigma_b^2), \epsilon_i \sim N(0, \sigma^2) \quad (9.3)$$

Now the linear mixed model structure is apparent. Although we do not always explicitly state this, there is the additional assumption that the random effects  $a(k)$ ,  $b(k)$  and  $\epsilon_i$  are mutually independent. For randomly varying lines  $(a(k), b(k))$  in the same  $x$ -domain this may be an unreasonable assumption since the slope and intercept values may very well be related to each other. It is possible to extend the model to allow for such a correlation/covariance between the intercept and slope by assuming a bi-variate normal distribution for each set of line parameters:

$$(a(k), b(k)) \sim N\left(0, \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix}\right), \epsilon_i \sim N(0, \sigma^2) \quad (9.4)$$

The model given by (9.2) and (9.4) is the standard random coefficient mixed model.

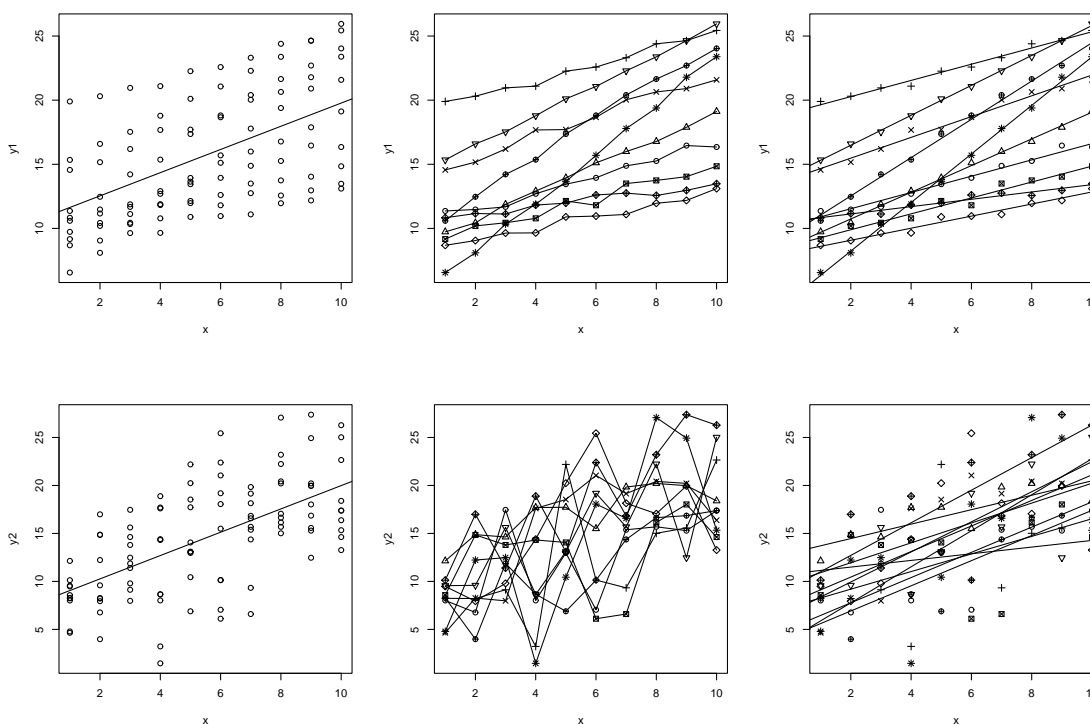


Figure 9.1: Constructed data: Top: data set 1, bottom: data set 2. Left: Raw scatter plot with simple regression line, middle: Individual patterns, right: individual lines

## 9.2 Example: Constructed data

To illustrate the basic principles we start with two constructed data sets of 100 observations of  $y$  for 10 different  $x$ -values, see figure 9.1. It reflects that a raw scatter plot of a data set can be hiding quite different structures, if the data is in fact hierarchical (repeated observations on each individual rather than exactly one observation for each individual).

### 9.2.1 Simple regression analysis

Had the data NOT been hierarchical, but in stead observations on 100 subjects, a simple regression analysis, corresponding to the model

$$y_i = \alpha + \beta x_i + \epsilon_i \quad (9.5)$$

where  $\epsilon_i \sim N(0, \sigma^2)$ ,  $i = 1, \dots, 100$  would be a reasonable approach. For comparison we state the results of such an analysis for the two data sets. The parameter estimates are:

Parameter	Data 1			Data 2		
	Estimate	SE	P-value	Estimate	SE	P-value
$\sigma^2$	15.9899			20.5229		
$\alpha$	10.7280	0.8638		7.8356	0.9786	
$\beta$	0.90461	0.1392	<0.0001	1.21519	0.1577	<0.0001

See figure 9.1(left) for the estimated lines.

## 9.2.2 Fixed effects analysis

If we had special interest in the 10 subjects, a fixed effects analysis corresponding to model (9.1) could be carried out. The F-tests and P-values from the Type 1 (successive) ANOVA tables become:

Source	DF	Data set 1		Data set 2	
		F	P-value	F	P-value
x	1	9220.98	<.0001	70.74	<.0001
subject	9	2091.49	<.0001	3.07	0.0033
x*subject	9	277.71	<.0001	1.02	0.4311

For data set 1 the slopes are clearly different whereas for data set 2 the slopes can be assumed equal, but the intercepts (subjects) are different. Although it is usually recommended to rerun the analysis without an insignificant interaction effect, the Type I table shows that the result of this will clearly be that the subject (intercept) effect is significant for data set 2, cf. the discussion of Type I/Type III tables in Module 3. So for data set 1, the (fixed effect) story is told by providing the 10 intercept and slope estimates and/or possibly as described for the different slopes ANCOVA model in the previous module. For data set 2, an equal slopes ANCOVA model can be used to summarize the results. The common slope and error variance estimates are:

$$\hat{\beta} = 1.2152, SE_{\hat{\beta}} = 0.1446, \hat{\sigma}^2 = 17.2582$$

The confidence band for the common slope, using the 89 error degrees of freedom becomes

$$1.2152 \pm t_{0.975}(89)0.1446$$

which, since  $t_{0.975}(89) = 1.987$ , gives

$$[0.9279, 1.5025]$$

The subjects could be described and compared as for the common slopes ANCOVA model of the previous module.

### 9.2.3 Two step analysis

If the interest is NOT in the individual subjects but rather in the average line, then a natural ad hoc approach is simply to start by calculating the individual intercepts and slopes and then subsequently treat those as simple random samples and calculate average, variance and standard error to obtain confidence limits for the population average values. So e.g. for the slopes we have  $\hat{\beta}_1, \dots, \hat{\beta}_{10}$  and calculate the average

$$\bar{\beta} = \frac{1}{10} \sum_{i=1}^{10} \hat{\beta}_i,$$

the variance

$$s_{\hat{\beta}}^2 = \frac{1}{9} \sum_{i=1}^{10} (\hat{\beta}_i - \bar{\beta})^2$$

and the standard error

$$SE_{\hat{\beta}} = \frac{s_{\hat{\beta}}}{\sqrt{10}}$$

to obtain the 95% confidence interval: (using that  $t_{0.975}(9) = 2.26$ )

$$\bar{\beta} \pm 2.26 SE_{\hat{\beta}}$$

The variances for data set 1 are:

$$s_{\hat{\alpha}}^2 = 16.2779, \quad s_{\hat{\beta}}^2 = 0.2465$$

and for data set 2:

$$s_{\hat{\alpha}}^2 = 8.5663, \quad s_{\hat{\beta}}^2 = 0.2130$$

The results for the intercepts and slopes for the two data sets are given in the following table:

	Data set 1		Data set 2	
	$\alpha$	$\beta$	$\alpha$	$\beta$
Average	10.7279	0.9046	7.8356	1.2152
SE	1.2759	0.1570	0.9255	0.1460
Lower	7.8416	0.5495	5.7419	0.8850
Upper	13.6142	1.2597	9.9293	1.5454

Note that for data set 2, the standard error for the slope is almost identical to the standard error from the fixed effect equal slopes model from above. However, due to the smaller degrees of freedom, 9 instead of 89, the confidence band is somewhat larger here. This reflects the difference in interpretation: In the fixed effects analysis the  $\beta$

estimates the common slope for these specific 10 subjects. Here the estimate is of the population average slope (the population from which these 10 subjects were sampled). This distinction does not alter the estimate itself, but does change the statistical inference that is made.

Note, by the way, that for estimating the individual lines, it does not make a difference whether an overall different slopes model is used or 10 individual ("small") regression models separately.

Although not used, the observed correlation between the intercepts and slopes in each case can be found:

$$\text{corr}_1 = -0.382, \quad \text{corr}_2 = -0.655$$

## 9.2.4 Random coefficient analysis

The results of fitting the random coefficient model given by (9.2) and (9.4) to each data set, see **the SAS section** for details, is given in the following table:

	Data set 1		Data set 2	
	$\alpha$	$\beta$	$\alpha$	$\beta$
Average	10.7279	0.9046	7.8356	1.2152
SE	1.2759	0.1570	0.9255	0.1460
Lower	7.8416	0.5495	5.7419	0.8850
Upper	13.6142	1.2597	9.9293	1.5454

Note that this table is an exact copy of the result table for the two-step analysis above! The parameters of the variance part of the mixed model for data set 1 is estimated at:

$$\hat{\sigma}_a^2 = 16.2451, \quad \hat{\sigma}_b^2 = 0.2456, \quad \hat{\sigma}_{ab} = -0.7607, \quad \hat{\sigma}^2 = 0.0732$$

and for data set 2:

$$\hat{\sigma}_a^2 = 0.5292, \quad \hat{\sigma}_b^2 = 0.004285, \quad \hat{\sigma}_{ab} = 0.2636, \quad \hat{\sigma}^2 = 17.2224$$

Compare with the variances calculated in the two-step procedure: For data set 1, the random coefficient model estimates are slightly smaller, whereas for data set 2, they are considerably smaller. This makes good sense, as the variances in the two-step procedure also will include some additional variation due to the residual error variance (just like the mean squares in a standard hierarchical model). For data set 1, this residual error is estimated at a very small value (0.0732) whereas for data set 2 it is 17.2224. This illustrates how the random coefficient model provides the proper "story" about what is going on, and directly distinguishes between the two quite different situations exemplified here.

The covariance parameter estimates can be used to calculate the estimate of the correlation between the intercept and the slope:

$$\hat{\rho}_{ab} = \frac{\hat{\sigma}_{ab}}{\hat{\sigma}_a \hat{\sigma}_b}.$$

For data set 1, this gives  $\hat{\rho}_{ab} = -0.381$  which is close to the observed correlation calculated in the two-step procedure. However, for data set 2 the estimated correlation becomes  $\hat{\rho}_{ab} = 5.54!!!$  This obviously makes no sense! We encounter a situation similar to the the "negative variance" problem discussed previously. The correlation may become meaningless when some of the variances are estimated very small, which is the case for the slope variance here. To put it differently, for data set 2 the model we have specified include components (in the variance) that is not actually present in the data. We already new this, since the equal slopes model was a reasonable description of this data. In the random coefficient framework the equal slopes model is expressed by

$$y_i = \alpha + \beta x_i + a(\text{subject}_i) + \epsilon_i \quad (9.6)$$

where

$$a(k) \sim N(0, \sigma_a^2), \epsilon_i \sim N(0, \sigma^2) \quad (9.7)$$

The adequacy of this model can be tested by a residual likelihood ratio test, cf. Module 5. For data set 2 we obtain

$$G = -2l_{REML,1} - (-2l_{REML,2}) = 1.3$$

which is non-significant using a  $\chi^2$  distribution with 2 degrees of freedom. For data set 1 the similar test becomes

$$G = -2l_{REML,1} - (-2l_{REML,2}) = 249.9$$

which is extremely significant.

For data set 2 the conclusions should be based on the equal slopes model given by (9.6) and (9.7), and we obtain the following:

	Data set 2	
	$\alpha$	$\beta$
Estimate	7.8356	1.2152
SE	1.0774	0.1446
Lower	5.6544	0.9278
Upper	10.0168	1.5026

We see a minor change in the confidence bands: believing in equal slopes increases the (estimated) precision (smaller confidence interval) for this slope, whereas the precision of the average intercept decreases.

### 9.3 Example: Consumer preference mapping of carrots

In a consumer study 103 consumers scored their preference of 12 danish carrot types on a scale from 1 to 7. The carrots were harvested in autumn 1996 and tested in march 1997. A number of background information variables were recorded for each consumer, **see the data description for details**. The aim of a so-called "external preference mapping" is to find the "sensory drivers" of the consumer preference behaviour and to investigate if these are different in different segments of the population. To do this, in addition to the consumer survey, the carrot products are evaluated by a trained panel of tasters, the sensory panel, with respect to a number of sensory (taste, odour and texture) properties. Since usually a high number of (correlated) properties(variables) are used, in this case 14, it is a common procedure to use a few, often 2, combined variables that contain as much of the information in the sensory variables as possible. This is achieved by extracting the first two principal components in a principal components analysis(PCA) on the product-by-property panel average data matrix. PCA is a commonly used multivariate technique to explore and/or decompose high dimensional data.

We call these two variables `sens1` and `sens2` and they are given by

$$\text{sens1}_i = \sum_{j=1}^{14} a_j v_j^i \text{ and } \text{sens2}_i = \sum_{j=1}^{14} b_j v_j^i$$

where  $v_1^i, \dots, v_{14}^i$  are the 14 average sensory scores for carrot product  $i$  and the coefficients  $a_j$  and  $b_j$  defining the two combined sensory variables are as depicted in figure 9.2. So `sens1` is a variable that (primarily) measures bitterness vs. nutty taste whereas `sens2` measures sweetness (and related properties). The actual "preference mapping" is carried out by first fitting regression models for the preference as a function of the sensory variables for each individual consumer using the 12 observations across the carrot products. Next, the individual regression coefficients are investigated, often in an explorative manner in which a scatter plot is used to look for a possible segmentation of consumers in these regression coefficients. In stead of looking for segmentation ("Cluster analysis") we investigate whether we see any differences with respect to the background variables in the data, e.g. the gender or homesize (number of persons in the household). Let  $y_i$  be the  $i$ th preference score. The natural model for this is a model that expresses randomly varying individual relations to the sensory variables, but with average (expected) values that may depend on the homesize.

Let us consider the factor structure of the setting. The basic setting is a randomized block experiment with 12 treatments (carrot products), the factor `prod`, and 103 blocks (consumers), the factor `cons`. Homesize (`size`) is a factor that partitions the consumers into two groups, those with homesize of 1 or 2, and those with a larger

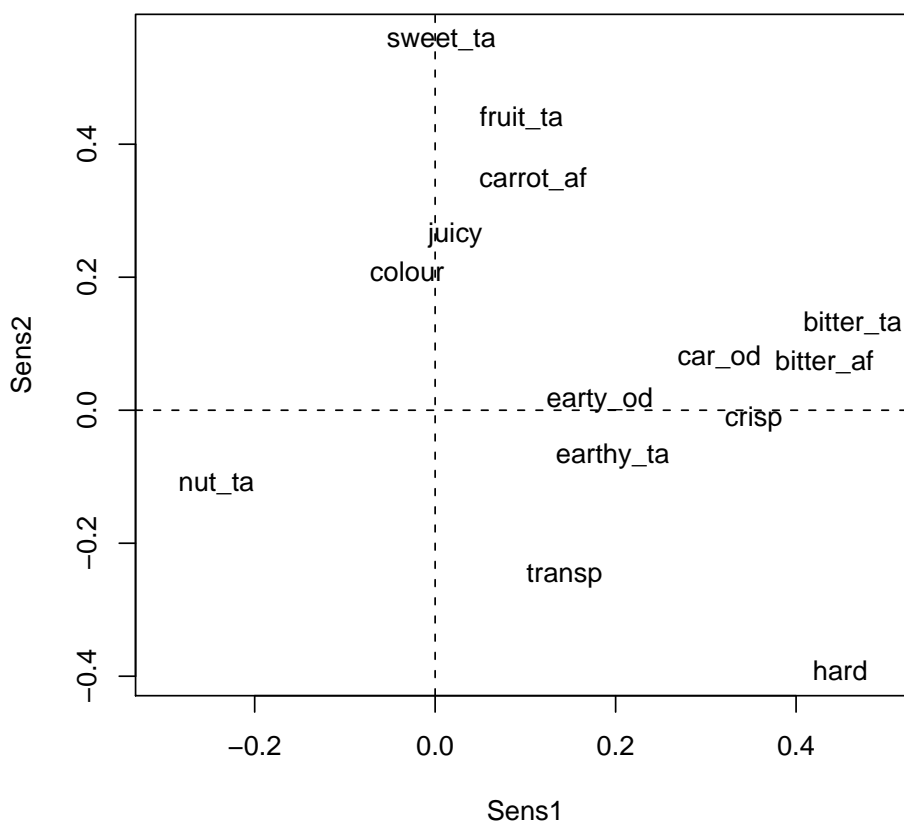


Figure 9.2: Loadings plot for PCA of sensory variables: Scatter plot of coefficients  $b_j$  versus  $a_j$ .

homesize. So the factor `cons` is nested within `size`, or equivalently `size` is coarser than `cons`. This basic structure is depicted in figure 9.3.

The linear effect of the sensory variables is a part of the `prod` effect, since these covariates "are on product level". So they are both coarser than the product effect. The sensory variables in the model will therefore explain some of the product differences. Including `prod` in the model as well will enable us to test whether the sensory variables can explain all the product differences. As we do not expect this to be the case, we adopt the point of view that the 12 carrot products is a random sample from the population of carrot products in Denmark, that is, the product effect is considered as a random effect. In other words, we consider the deviations in the product variation from what can be explained by the regression on the sensory variables, as random variation. Finally, the interactions between `homesize` and the sensory variables should enter the

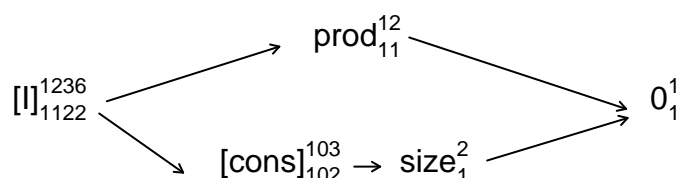


Figure 9.3: The factor structure diagram for the carrots data

model as fixed effects, allowing for different average slopes for the two homesizes, leading to the model given by

$$y_i = \alpha(\text{size}_i) + \beta_1(\text{size}_i) \cdot \text{sens1}_i + \beta_2(\text{size}_i) \cdot \text{sens2}_i + a(\text{cons}_i) + b_1(\text{cons}_i) \cdot \text{sens1}_i + b_2(\text{cons}_i) \cdot \text{sens2}_i + d(\text{prod}_i) + \epsilon_i \quad (9.8)$$

where

$$a(k) \sim N(0, \sigma_a^2), b_1(k) \sim N(0, \sigma_{b_1}^2), b_2(k) \sim N(0, \sigma_{b_2}^2), k = 1, \dots, 103. \quad (9.9)$$

and

$$d(\text{prod}_i) \sim N(0, \sigma_P^2), \epsilon_i \sim N(0, \sigma^2) \quad (9.10)$$

To finish the specification of a general random coefficient model, we need the assumption of the possibility of correlations between the random coefficients:

$$(a(k), b_1(k), b_2(k)) \sim N\left(0, \begin{pmatrix} \sigma_a^2 & \sigma_{ab_1} & \sigma_{ab_2} \\ \sigma_{ab_1} & \sigma_{b_1}^2 & \sigma_{b_1 b_2} \\ \sigma_{ab_2} & \sigma_{b_1 b_2} & \sigma_{b_2}^2 \end{pmatrix}\right) \quad (9.11)$$

However, it didn't succeed to fit this model to the data by SAS (in a reasonable amount of time). These things occur in practical data analysis. Maybe the computing power/memory availability of the computer at hand is exceeded. More likely in

this case is that we have asked the computer to do an impossible job: there may be too little information in the data to estimate all the parameters of the model – this is also known as identifiability problems. The only (practical) solution is to go for a simpler model and try to figure out, maybe step by step what structures may be found in the data. So we fit the model given by equations (9.9) and (9.10) only. The  $-2l_{REML}$  value and the variance parameter estimates are given by:

Model 1	
$-2l_{REML}= 3750.7$	
$\sigma_a^2$	0.1972
$\sigma_{b_1}^2$	0
$\sigma_{b_2}^2$	0.002963
$\sigma_P^2$	0.03354
$\sigma^2$	1.0400

Before studying the fixed effects, the variance part of the model is investigated further. The zero estimate of  $\sigma_{b_1}^2$  indicates that there is no variation between the `sens1` coefficients, so we start by removing `sens1` from the random effects (this does not change anything since this is what SAS does anyway). We then test the hypothesis that  $\sigma_{b_2}^2 = 0$  by fitting a model without the random varying `sens2` coefficients:

Model 2A	
Hypothesis: $\sigma_{b_2}^2 = 0$	
$-2l_{REML}=3758.0$	
Test versus model 1	$G = 7.3(0.006895)$
$\sigma_a^2$	0.1937
$\sigma_P^2$	0.03313
$\sigma^2$	1.0834

The  $G$ -statistic is the difference between the two  $-2l_{REML}$ -values and the P-value is found from the  $\chi^2$ -distribution with 1 degree of freedom. This means that we cannot discard the random varying `sens2` coefficients. We may also test whether there are any significant product differences not explained by the linear relationships with the sensory variables:

Model 2B	
Hypothesis: $\sigma_P^2 = 0$	
$-2l_{REML}=3766.8$	
Test versus model 1	$G = 3.9(0.0483)$
$\sigma_a^2$	0.1953
$\sigma_{b_2}^2$	0.002780
$\sigma^2$	1.0699

So we cannot discard the product variance component either. Finally, we try to expand the model to allow for a correlation between the random intercepts (consumer main effects) and the `sens2` coefficients:

$$(a(k), b_2(k)) \sim N(0, \begin{pmatrix} \sigma_a^2 & \sigma_{ab_2} \\ \sigma_{ab_2} & \sigma_{b_2}^2 \end{pmatrix}) \quad (9.12)$$

In fact this model now fits easily to the data giving the results:

Model 0	
-2l <sub>REML</sub> =3750.0	
Test of model 1 versus model 0	G = 0.7(0.4028)
$\sigma_a^2$	0.1972
$\sigma_{b_2}^2$	0.002963
$\sigma_{ab_2}$	0.004302
$\sigma_P^2$	0.03356
$\sigma^2$	1.0400

However, the starting model 1 is not rejected when tested against the extended model 0 (the correlation between the `sens2` coefficients and the intercepts can be assumed zero) and this starting model is the one, we use for statements about the fixed effects: (Type III ANOVA table)

Effect	Num DF	Den DF	F Value	P value
Homesize	1	101	5.20	0.0247
Sens1	1	9.03	0.49	0.5014
Sens2	1	11	17.19	0.0016
Sens1*Homesize	1	1016	0.18	0.6730
Sens2*Homesize	1	101	1.04	0.3111

Removing the non-significant Sens1\*Homesize effect and rerunning the analysis gives:

Effect	Num DF	Den DF	F Value	P value
Homesize	1	101	5.20	0.0247
Sens1	1	8.99	0.52	0.4883
Sens2	1	11	17.18	0.0016
Sens2*Homesize	1	101	1.04	0.3110

and rerunning without the non-significant Sens2\*Homesize effect:

Effect	Num DF	Den DF	F Value	P value
Homesize	1	101	5.20	0.0248
Sens1	1	8.99	0.52	0.4881
Sens2	1	10.9	16.74	0.0018

and finally without the Sens1 effect:

Effect	Num DF	Den DF	F Value	P value
Homesize	1	101	5.20	0.0247
Sens2	1	12.2	17.49	0.0012

The final model for these data is therefore given by:

$$y_i = \alpha(\text{size}_i) + \beta_2 \cdot \text{sens2}_i + a(\text{cons}_i) + b_2(\text{cons}_i) \cdot \text{sens2}_i + d(\text{prod}_i) + \epsilon_i \quad (9.13)$$

where

$$a(k) \sim N(0, \sigma_a^2), \quad b_2(k) \sim N(0, \sigma_{b_2}^2), \quad k = 1, \dots, 103. \quad (9.14)$$

and

$$d(\text{prod}_i) \sim N(0, \sigma_P^2), \quad \epsilon_i \sim N(0, \sigma^2) \quad (9.15)$$

The estimates of the variances are listed in the following table:

$\sigma_a^2$	0.1973
$\sigma_{b_2}^2$	0.002972
$\sigma_P^2$	0.03148
$\sigma^2$	1.0392

The conclusions regarding the relation between the preference and the sensory variables are that no significant relation was found to sens1, but indeed so for sens2. The relation does not depend on the homesize and is estimated at:(with 95% confidence interval)

$$\hat{\beta}_2 = 0.0706, \quad [0.0339, 0.107]$$

So two products with a difference of 10 in the 2nd sensory dimension (this is the span in the data set) are expected to differ in average preference with between 0.34 and 1.1. Sweet products are preferred to non-sweet products, cf. figure 9.2 above. The expected values for the two homesizes (for an average product) and their differences are estimated at:

$$\begin{aligned} \hat{\alpha}(1) + \hat{\beta}_2 \cdot \overline{\text{sens2}} &= 4.9069, \quad [4.7306, 5.0832] \\ \hat{\alpha}(2) + \hat{\beta}_2 \cdot \overline{\text{sens2}} &= 4.6661, \quad [4.4776, 4.8546] \\ \hat{\alpha}(1) - \hat{\alpha}(2) &= 0.2408, \quad [0.03126, 0.4504] \end{aligned}$$

So homes with more persons tend to have a slightly lower preference in general for such carrot products.

## 9.4 Random coefficient models in perspective

Although the factor structure diagrams with all the features of finding expected mean squares and degrees of freedom are only strictly valid for balanced designs and models with no quantitative covariates, they may still be useful as a more informal structure visualization tool for these non-standard situations.

The setting with hierarchical regression data is really an example of what also could be characterized as repeated measures data. A common situation is that repeated measurements on a subject (animal, plant, sample) are taken over time then also known as longitudinal data. So apart from appearing as natural extensions of fixed regression models, the random coefficient models are one option for analyzing repeated measures data. The simple models can be extended to polynomial models to cope with non-linear structures in the data. Also additional residual correlation structures can be incorporated. In Modules 11 and 12 a thorough treatment of repeated measures data is given with a number of different methods – simple as well as more complex approaches.