Факторный анализ
Смешанные модели

Начинай от низшего степени, чтобы дойти до высшего; другими словами: не чеши затылок, а чеши пятки.

Козьма Прутков

Модель факторного анализа

Модель факторного анализа основана на идее, в соответствии с которой структура связей между m анализируемыми переменными $X_1, \ldots, X_m$ может быть объяснена тем, что они зависят от меньшего числа других, непосредственно не измеряемых ("скрытых") факторов $F_1, \ldots, F_k$, $k < m$. Помимо общих факторов каждый из исходных признаков X_i зависит также и от некоторой специфической для него остаточной случайной компоненты ε_i.

Каноническая модель факторного анализа выглядит так:

$$X_i = q_{i1}F_1 + q_{i2}F_2 + \ldots + q_{ik}F_k + \varepsilon_i, \quad i = 1, \ldots, m,$$

Неизвестные константы q_{ij} называются *факторными нагрузками* (loadings).

Примером достаточно прозрачной интерпретации факторной модели может служить ее формулировка в терминах так называемых **интеллектуальных тестов**. При этом наблюдение $X_i^{(l)}$ выражает отклонение от некоторого среднего уровня оценки в баллах, полученной l-м индивидуумом на экзамене по i-му тесту. Естественно предположить, что в качестве ненаблюдаемых факторов $F_1, \ldots, F_k$, от которых будут зависеть оценки индивидуумов по всем m тестам, выступят такие факторы, как характеристика *общей одаренности индивидуума* (F_1), характеристики его *математических* (F_2), *технических* (F_3) или *гуманитарных* (F_4) способностей.

Принципиальное отличие модели факторного анализа от регессионной модели состоит в том, что переменные F_j не являются непосредственно наблюдаемыми, в то время как в регрессионном анализе значения признаков измеряются на статистически обследованных объектах.

Допущения и их следствия

$$X_i = q_{i1}F_1 + \ldots + q_{ik}F_k + \varepsilon_i, \quad i = 1, \ldots, m. \tag{1}$$

Допущения. Здесь предполагается, что общие факторы $F_1, \ldots, F_k$ — некоррелированые случайные величины, $\mathbf{M}F_j = 0$ и $\mathbf{D}F_j = 1$ для $j = 1, \ldots, k$; характерные факторы $\varepsilon_1, \ldots, \varepsilon_m$ также некоррелированы между собой, $\mathbf{M}\varepsilon_i = 0$ и $\mathbf{D}\varepsilon_i = v_i$ (неизвестные параметры); величины ε_i и F_j некоррелированы для всех i и j.

Получим несколько формул, вытекающих из предположений факторной модели. Во-первых, в силу некоррелированности всех F_j и ε_i дисперсия признака X_i может быть представлена в виде

$$\sigma_{ii} = \mathbf{D}X_i = q_{i1}^2\mathbf{D}F_1 + \ldots + q_{ik}^2\mathbf{D}F_k + \mathbf{D}\varepsilon_i = \sum_{j=1}^{k} q_{ij}^2 + v_i, \tag{2}$$

где $\sum q_{ij}^2$ известна под названием *общности* (communality). Она представляет собой часть дисперсии, обусловленную общими факторами, а v_i — часть дисперсии, обусловленная ошибкой.

Во-вторых, ковариация между X_s и X_t $(s, t = 1, \ldots, m)$ задается выражением

$$\sigma_{st} = \mathbf{cov}\,(X_s, X_t) = \mathbf{M}X_sX_t = \sum_{j=1}^{k} q_{sj}q_{tj}, \quad s \neq t. \tag{3}$$

Аналогично выводим, что

$$\mathbf{cov}\,(X_i, F_j) = \mathbf{cov}\,(q_{i1}F_1 + \ldots + q_{ik}F_k + \varepsilon_i, F_j) = q_{ij}\mathbf{M}F_j^2 = q_{ij}. \tag{4}$$

Таким образом, факторные нагрузки — суть *ковариации* между наблюдаемыми признаками и неизвестными факторами.

Вопросы факторного анализа

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{F} + \boldsymbol{\varepsilon}, \tag{5}$$

где $\boldsymbol{X} = (X_1, \ldots, X_m)^T$, $\boldsymbol{Q} = ||q_{ij}||_{m\times k}$ — матрица нагрузок, $\boldsymbol{F} = (F_1, \ldots, F_k)^T$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_m)^T$. С учетом формул (2)–(3) имеем следующее представление для ковариационной матрицы признаков:

$$m(m+1)/2 \longrightarrow \boldsymbol{\Sigma} = ||\sigma_{st}||_{m\times m} = \boldsymbol{Q}\boldsymbol{Q}^T + \boldsymbol{V}, \longleftarrow mk + m \tag{6}$$

где $\boldsymbol{V}$ — диагональная матрица размерности $m \times m$ с элементами v_i на главной диагонали.

Допустим, что известна только ковариационная матрица $\boldsymbol{\Sigma}$. При использовании факторной модели приходится последовательно анализировать и решать **следующие вопросы**.

1) *Существование.* При каких ограничениях на $\boldsymbol{\Sigma}$ найдутся такие матрицы $\boldsymbol{Q}$ и $\boldsymbol{V}$, что для них выполняется равенство (6).

2) *Единственность (идентифицируемость).* Если матрицы $\boldsymbol{Q}$ и $\boldsymbol{V}$ существуют, то при каких условиях на ковариационную матрицу $\boldsymbol{\Sigma}$ они определяются единственным образом.

3) *Оценка параметров модели* на основе выборочной ковариационной матрицы $\widehat{\boldsymbol{\Sigma}}$. Необходимо уметь оценивать элементы матриц $\boldsymbol{Q}$ и $\boldsymbol{V}$ на основе $\widehat{\boldsymbol{\Sigma}}$.

4) *Проверка гипотез*, в частности, гипотезы об истинном числе k общих факторов.

5) *Построение статистических оценок* для ненаблюдаемых общих факторов $F_1, \ldots, F_k$.

Условия существования модели факторного анализа известны. Что касается единственности, то следует обратить внимание на следующее: *факторы и факторные нагрузки определяются не однозначно, а с точностью до произвольного ортогонального преобразования.* Действительно, пусть $\boldsymbol{C}$ — ортогональная матрица размера $k \times k$, т. е. $\boldsymbol{C}\boldsymbol{C}^T = \boldsymbol{E}$. Тогда факторная модель (5) может быть переписана в виде

$$\boldsymbol{X} = \boldsymbol{Q}(\boldsymbol{C}\boldsymbol{C}^T)\boldsymbol{F} + \boldsymbol{\varepsilon} = (\boldsymbol{Q}\boldsymbol{C})(\boldsymbol{C}^T\boldsymbol{F}) + \boldsymbol{\varepsilon},$$

что можно рассматривать как модель с факторными нагрузками $\boldsymbol{Q}\boldsymbol{C}$ и факторами $\boldsymbol{C}^T\boldsymbol{F}$.

«Квартимакс», «варимакс», «промакс» — методы поворота осей

Как правило, исследователь, получив некоторое конкретное решение, отказывается от него и производит дополнительное ортогональное преобразование (поворот) факторов для того, чтобы содержательно интерпретировать новые факторы. При этом обычно обращают внимание только на наибольшие нагрузки и стремятся преобразовать факторы таким образом, чтобы большие нагрузки увеличились, а маленькие — уменьшились ("занулились"). Увеличивая различие между нагрузками, не нужно обращать внимание на их алгебраические знаки. Из-за того, что при работе с абсолютными величинами возникают некоторые математические трудности, удобнее максимизировать различие между *квадратами* факторных нагрузок.

Метод "варимакс", введенный Г. Кайзером в 1958 г., является модификацией квартимакс-метода, при которой акцент делается на упрощение *столбцов* матрицы факторных нагрузок. Согласно Кайзеру, простота отдельного фактора F_j определяется дисперсией квадратов его нагрузок, т. е. величиной

$$S_j^2 = \frac{1}{m}\sum_{i=1}^{m} q_{ij}^4 - \left(\frac{1}{m}\sum_{i=1}^{m} q_{ij}^2\right)^2.$$

Если эта дисперсия максимальна, то фактор наилучшим образом интерпретируем, поскольку при этом его нагрузки в основном близки к 0 или 1. Критерий наибольшей простоты полной матрицы $\boldsymbol{Q}$ сводится, следовательно, к максимизации

$$T^2 = \sum_{j=1}^{k} S_j^2.$$

В R по умолчанию применяется «варимакс».

Телекоммуникационные услуги

	Factor		
	1	2	3
Long distance last month	.062	-.121	.854
Toll free last month	.726	.018	.191
Equipment last month	.067	.831	.049
Calling card last month	.348	-.012	.431
Wireless last month	.530	.637	.146
Multiple lines	-.025	.384	.438
Voice mail	.455	.539	.054
Paging service	.468	.566	.044
Internet	-.049	.722	-.045
Caller ID	.787	.056	.008
Call waiting	.779	.033	.054
Call forwarding	.768	.062	.048
3-way calling	.743	.050	.078
Electronic billing	-.107	.686	-.080

На основе данной матрицы факторных нагрузок можно дать рекомендации для более рационального предложения клиентам новых услуг. Например, клиенты, уже пользующиеся «Дополнительными услугами», скорее всего, будут склонны воспользоваться Беспроводной связью (Wireless last month), чем Интернетом (Internet).

Факторы можно интерпретировать так:

первый фактор — «Дополнительные услуги» (Extras),
второй фактор — «Технические средства» (Tech),
третий фактор — «Междугородняя связь» (Long distance).

Этот пример позаимствован из Help пакета SPSS.

Практическое задание 1

1) Преобразуйте файл Parodont.xls в текстовый формат, импортируйте его в RStudio с заголовками: поставьте переключатель Heading в положение «Yes»

2) Выберите из таблицы данных признаки с помощью команды

```
x=Parodont[c(11:15,19:22,24:31)]
```

(Это – индикаторы наличия 5 видов бактерий, 4 видов вирусов и 8 генетических признаков, из которых 4 предохраняют от развития пародонтита и 4 предрасполагают к его развитию.)

3) Удалите из таблицы все строки с пропусками: y=na.omit(x)

4) Спроецируйте на 2 первые главные оси с помощью команды biplot(prcomp(y)) и нажмите Zoom. Сколько наблюдается пучков?

5) Примените модель факторного анализа и выведите нагрузки:

```
r=factanal(covmat=cor(y), factors=5); r$loadings
```

Сколько факторов надо оставить? Как их интерпретировать? (Обратите внимание на большие факторные нагрузки.)

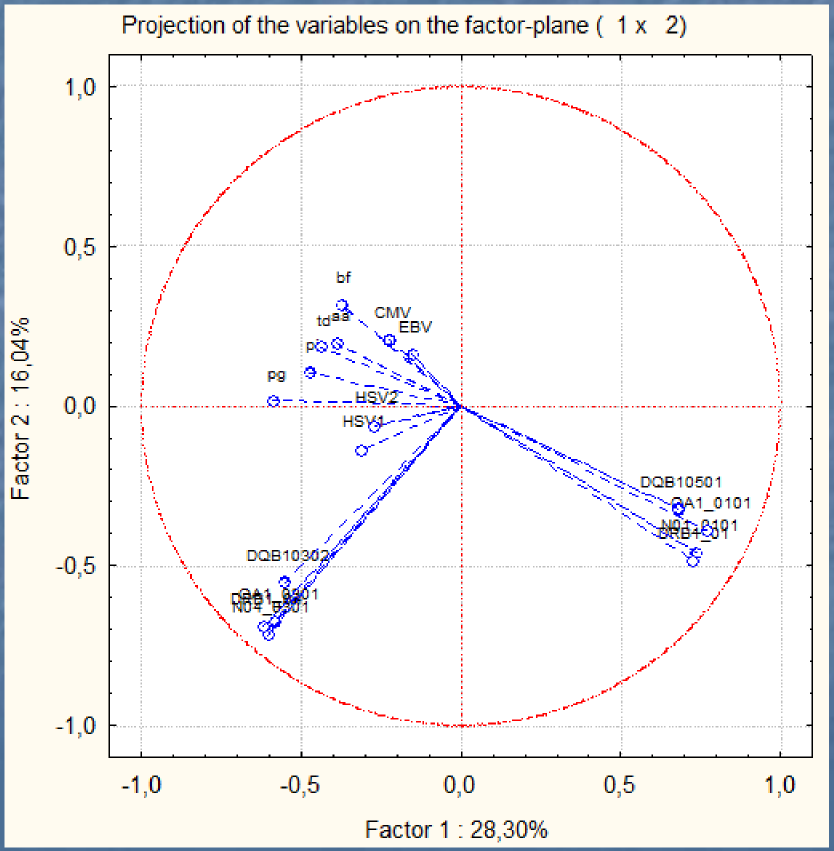
Диаграмма проекций признаков на две первые главные оси

Видим, что переменные распадаются на 3 «пучка»:

1) бактерии + вирусы;

2) генетические признаки, предохраняющие от заболевания пародонтитом (справа от оси ординат);

3) генетические признаки, предрасполагающие к заболеванию пародонтитом (слева внизу).



Выделение 4-х новых факторов на основе таблицы факторных нагрузок

Видим, что теперь переменные распадаются на 4 группы:

1) генетические признаки, предохраняющие от заболевания пародонтитом;

2) генетические признаки, предрасполагающие к заболеванию пародонтитом;

3) бактерии (кроме бактерий вида aa);

4) вирусы (за исключением вирусов вида EBV).

Variable	Factor Loadings (Varimax raw) (Parodont) Extraction: Principal components (Marked loadings are > ,600000)			
	Factor 1	Factor 2	Factor 3	Factor 4
pi	0,120201	0,159262	0,732210	0,051382
bf	0,181207	-0,051313	0,783057	-0,159014
td	0,133232	0,071927	0,727181	0,072229
aa	0,265649	0,017092	0,273615	0,378293
pg	0,177668	0,283419	0,647456	0,225868
HSV1	0,049732	0,224702	0,046229	0,604444
HSV2	0,051440	0,121820	-0,013057	0,768865
CMV	0,168890	-0,113842	0,049766	0,671598
EBV	-0,040901	-0,078625	0,489988	0,131980
N01_0101	-0,929408	-0,066612	-0,141636	-0,009054
N04_0301	0,059010	0,931060	0,068002	0,076949
DRB1_01	-0,913296	-0,035074	-0,160428	-0,029478
DRB1_04	0,105246	0,923909	0,061051	0,024344
QA1_0101	-0,926826	-0,137299	-0,078362	-0,085077
QA1_0301	0,096631	0,894555	0,044462	-0,011998
DQB10302	0,147392	0,772247	0,012812	0,061633
DQB10501	-0,860543	-0,142283	0,033074	-0,075972
Expl.Var	3,545557	3,360100	2,480564	1,676651
Prp.Totl	0,208562	0,197653	0,145916	0,098627

Исконные и заимствованные английские суффиксы

Words such as *goodness* and *sharpness* can be analyzed as consisting of a stem, *good*, *sharp*, and an affix, the suffix *-ness*. Some affixes are used in many words, *-ness* is an example. Other affixes occur only in a limited number of words, for instance, the *-th* in *warmth* and *strength*. The extent to which affixes are used and available for the creation of new words is referred to as the productivity of the affix. Baayen (1994) addressed the question of the extent to which the productivity of an affix is codetermined by stylistic factors. Do different kinds of texts favor the use of different kinds of affixes?

The data set `affixProductivity` lists, for 44 texts with varying authors and genres, a productivity index for 27 derivational affixes. The 44 texts represent four different text types: religious texts (e.g. the *Book of Mormon*, coded `B`), books written for children (e.g. *Alice's Adventures in Wonderland*, coded `C`), literary texts (e.g. novels by Austen, Conrad, James, coded `L`), and other texts (including officialese from the US government accounting office), coded `O`. The classification codes are given in the column labeled `Registers`:

```
        ian    ful      y   ness   able     ly Registers
Mormon    0 0.1887 0.5660 2.0755 0.0000 2.2642         B
Austen    0 1.2891 1.5654 1.6575 1.0129 6.2615         L
Carroll   0 0.2717 1.0870 0.2717 0.4076 6.3859         C
Gao       0 0.3306 1.9835 0.8264 0.8264 4.4628         O
```

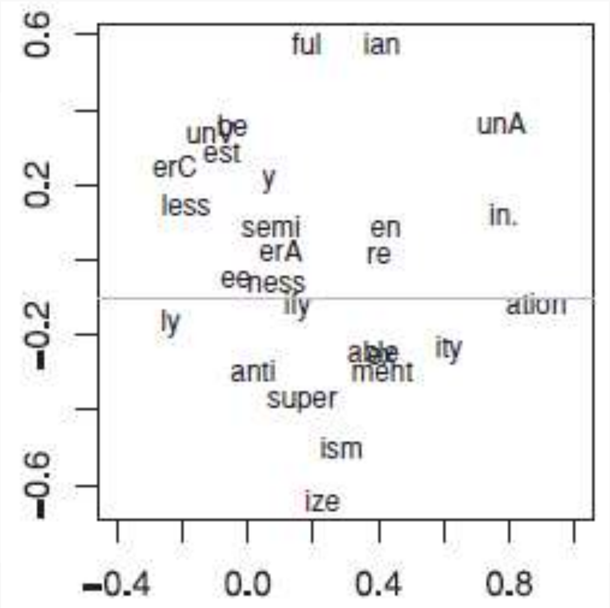
Практическое задание 2

1) Убедитесь, что подключён пакет languageR
2) Примените факторный анализ с поворотом «промакс»:
f=factanal(affixProductivity[,1:27], factors=3, rotation = "promax")
3) Запомните факторные нагрузки: q=loadings(f)
4) Постройте проекции признаков (суффиксов) на два главных фактора: plot(q, type="n", xlim=c(-0.4, 1)); text(q, rownames(q))
5) Проведите разделительную горизонтальную серую прямую:
abline(h=-0.1, col="darkgrey")
и нажмите Zoom

Most non-native affixes are located below the horizontal grey line; most native affixes are found above this line.

Native affixes (e.g. *-ness*, *-less*, *-er*)

Non-native affixes (e.g. *-ation*, *super-*, *anti-*)



Смешанные модели

Consider a study addressing the consequences of adding white noise to the comprehension of words presented auditorily over headphones to a group of subjects, using auditory lexical decision latencies as a measure of speed of lexical access. In such a study, the presence or absence of white noise would be the treatment factor, with two levels (noise versus no noise). In addition, we would need identifiers for the individual words (items), and identifiers for the individual participants (or subjects) in the experiment. The item and subject factors, however, differ from the treatment factor in that we would normally only regard the treatment factor as REPEATABLE.

A factor is repeatable, if the set of possible levels for that factor is fixed, and if, moreover, each of these levels can be repeated. In our example, the treatment factor is repeatable, because we can take any new acoustic signal and either add or not add a fixed amount of white noise. We would not normally regard the identifiers of items or subjects as repeatable. Items and subjects are sampled randomly from populations of words and participants, and replicating the experiment would involve selecting other words and other participants.

The statistical literature therefore makes a crucial distinction between factors with repeatable levels, for which we use FIXED-EFFECTS terms, and factors with levels randomly sampled from a much larger population, for which we use RANDOM-EFFECTS terms. MIXED-EFFECTS MODELS, or more simply, MIXED MODELS, are models which incorporate both fixed and random effects.

Пример из книги Pinheiro J., Bates D. «Mixed-Effects Models in S and S-PLUS»

1. Двухфакторная смешанная модель: с. 12
2. Тестирование инвалидных кресел: с. 13
3. Матричное представление модели: с. 14
4. Вызов модели в языке R (функция lme из nlme): с. 15
5. Значимость и оценки контрастов: с. 16
6. Доверительные интервалы для контрастов и стандартных отклонений: с. 20

Полезно также изучить с. 4-12 этой книги

Рандомизированная двухфакторная модель эксперимента по тестированию инвалидных кресел

A *randomized block design* is a type of experiment in which there are two classification factors: an *experimental* factor for which we use fixed effects and a *blocking* factor for which we use random effects.

The data shown in Figure 1.5 and available as the object `ergoStool` in the nlme library are from an ergometrics experiment that has a randomized block design. The experimenters recorded the effort required by each of nine different subjects to arise from each of four types of stools. We want to compare these four particular types of stools so we use fixed effects for the `Type` factor. The nine different subjects represent a sample from the population about which we wish to make inferences so we use random effects to model the `Subject` factor.

Результаты эргометрического эксперимента

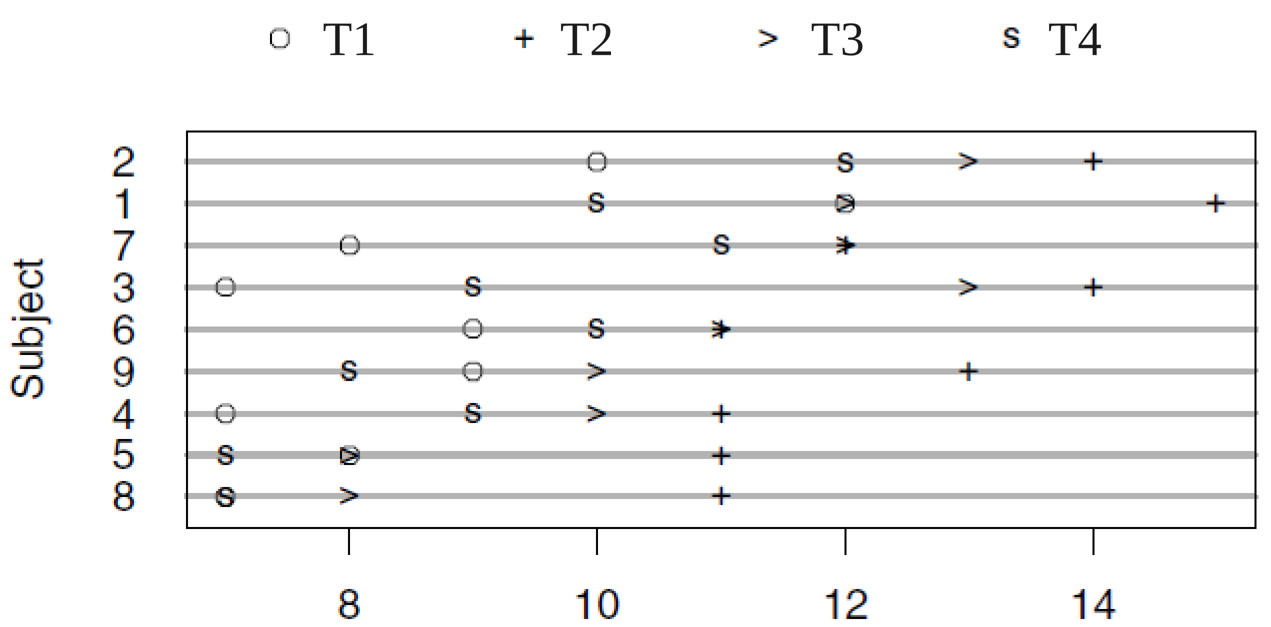


FIGURE 1.5. Effort required to arise (Borg scale)

From Figure 1.5 it appears that there are systematic differences between stool types on this measurement. For example, the `T2` stool type required the greatest effort from each subject while the `T1` stool type was consistently one of the low effort types.

Матричное представление модели

A model with fixed effects β_j for the `Type` factor and random effects b_i for the `Subject` factor could be written

$$y_{ij} = \beta_j + b_i + \epsilon_{ij}, \quad i = 1, \dots, 9, \quad j = 1, \dots, 4,$$
$$b_i \sim \mathcal{N}(0, \sigma_b^2), \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \tag{1.6}$$

Вместо девяти параметров, если бы b_i считались фиксированными, появился единственный параметр σ_b

or, equivalently,

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i b_i + \boldsymbol{\epsilon}_i, \quad i = 1, \dots, 9,$$
$$b_i \sim \mathcal{N}(0, \sigma_b^2), \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}),$$

where, for $i = 1, \dots, 9$,

$$\boldsymbol{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ y_{i4} \end{bmatrix}, \quad \boldsymbol{X}_i = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{Z}_i = \boldsymbol{1} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{\epsilon}_i = \begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \epsilon_{i3} \\ \epsilon_{i4} \end{bmatrix}.$$

Вызов смешанной модели в языке R

Надо подключить пакет nlme

```
> fm1Stool <- lme(effort ~ Type, data = ergoStool, random = ~ 1 | Subject)
> anova(fm1Stool)
```

```
            numDF denDF  F-value p-value
(Intercept)     1    24 455.0075  <.0001
Type            3    24  22.3556  <.0001
```

Значимость различий во всех уровнях Type

```
> summary(fm1Stool)
```

```
Random effects:
 Formula: ~1 | Subject
        (Intercept) Residual
StdDev:    1.332465 1.100295

Fixed effects: effort ~ Type
               Value Std.Error DF   t-value p-value
(Intercept) 8.555556 0.5760123 24 14.853079  0.0000
TypeT2      3.888889 0.5186838 24  7.497610  0.0000
TypeT3      2.222222 0.5186838 24  4.284348  0.0003
TypeT4      0.666667 0.5186838 24  1.285304  0.2110
```

Константа — среднее значение уровня для Type1

Значимость различий между уровнем Type1 и другими уровнями

Оценки контрастов с Type1

Доверительные интервалы для контрастов и стандартных отклонений

> intervals(fm1Stool)

```
Approximate 95% confidence intervals

 Fixed effects:
                 lower      est.    upper
(Intercept)  7.3667247 8.5555556 9.744386
TypeT2       2.8183781 3.8888889 4.959400
TypeT3       1.1517114 2.2222222 3.292733
TypeT4      -0.4038442 0.6666667 1.737177
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: Subject
                   lower     est.    upper
sd((Intercept)) 0.749509 1.332465 2.368835

 Within-group standard error:
    lower      est.     upper
0.8292494 1.1002946 1.4599324
```

Смотри формулу (1.6) на слайде 34

Интервал содержит ноль

Границы для параметра σ_b:

Довольно широкий интервал

Границы для параметра σ:

Зависимость времени лексического решения от некоторых факторов

Recall that data set `lexdec` provides visual lexical decision latencies elicited from 21 subjects for a set of 79 words: 44 nouns for animals, and 35 nouns for plants (fruits and vegetables). An experimental design in which we have multiple subjects responding to multiple items is referred to as a repeated measures design. For each word (item), we have 21 repeated measures (one measure from each subject). At the same time, we have 79 repeated measures for each subject (one for each item). Subject and item are random-effects factors; fixed-effects factors that are of interest include whether the subject was a native speaker of English, and whether the word referred to an animal or a plant, as well as lexical covariates such as frequency and length.

Практическое задание 3

1) Удалите «выбросы» из таблицы данных:

```
lex=lexdec[lexdec$RT<7,]
```

2) Отберите строки с правильными лексическими решениями:

```
d=lex[lex$Correct == "correct",]
```

3) Изучите графически влияние номера попытки Trial на логарифм времени принятия лексического решения RT (возможны как адаптация, так и утомление):

```
xylowess.fnc(RT ~ Trial | Subject, data = d, ylab = "log RT")
```

Нажмите Zoom

4) Примените смешанную модель с фиксированными факторами Trial, Sex, NativeLanguage и случайными эффектами Subject, Word %in% Subject:

```
m=lme(RT ~ Trial + Sex + NativeLanguage, data = d, random = ~1 | Subject / Word)
```

5) Выведите отчёт:

```
summary(m)
```

Влияние каких факторов на RT является значимым?

Отчёт для практического задания

```
Linear mixed-effects model fit by REML
 Data: d
        AIC       BIC   logLik
  -1092.015 -1054.579 553.0075

Random effects:
 Formula: ~1 | Subject
        (Intercept)
StdDev:   0.1176422

 Formula: ~1 | Word %in% Subject
        (Intercept)   Residual
StdDev:   0.1525376 0.06051728

Fixed effects: RT ~ Trial + Sex + NativeLanguage
                        Value  Std.Error   DF   t-value p-value
(Intercept)          6.305102 0.04267470 1535 147.74802  0.0000
Trial               -0.000167 0.00008812 1535  -1.89568  0.0582
SexM                 0.066463 0.05635093   18   1.17946  0.2536
NativeLanguageOther  0.142453 0.05369331   18   2.65308  0.0162
```

$$AIC = -2\log \text{Lik} + 2n_{par},$$
$$BIC = -2\log \text{Lik} + n_{par}\log(N)$$

Главное в теме

- В отличие от модели главных компонент (см. тему 7), где основной целью является уменьшения размерности пространства признаков, модель факторного анализа используется, прежде всего, для выявления и интерпретации новых признаков (факторов)
- Вращения осей (например, варимакс) предназначены для приведения матрицы факторных нагрузок к виду с большим числом элементов, близких к 0 или 1
- Смешанные модели предназначены для описания влияния как фиксированных факторов (аналогичных факторам их однофакторной или двухфакторной моделях из темы 4), так и случайных (как правило — мешающих) факторов, обычно выражающих существенную неоднородность наблюдений


Никто не обнимет необъятного!
Козьма Прутков