Модели со смешанными эффектами

Представьте, что мы хотим оценить результативность некоторой (медицинской) операции на поциентах в зависимости от их данных и пусть у нас есть соответствующие данные, собранные из разных больниц. В таком случае, мы могли бы просто применить обычную линейную регрессию. Однако такой подход может оказаться слишком наивным если нам также известно какая операция прошла в какой больнице: этот факт может также оказывать влияние на исход операции, т.е. результаты операции внутри каждой больницы будут иметь более схожие исходы. Более того, если нам известно какой именно врач проводил какую операцию, то учитывать стоит и это.

Для такого рода иерархичеки организованных данных обычно уместно применять *модели со смешанными эффектами*: мы желаем учесть влияние ряда групируюбщих факторов, но при этом не хотим включать их в набор наших фиксированных эффектов (параметры нашей модели $\vec{\beta}$), относя их к случайным эффектам.

В общем случае такие модели, встречающиеся на практике, почти всегда можно описать в виде:

$$\vec{y} = X\vec{\beta} + U\vec{\xi} + \vec{\varepsilon},$$

Здесь, как и ранее, $\vec{y}$ и $\vec{\varepsilon}$ - вектора размера n, равного количеству наблюдений, $\vec{\beta}$ - вектор длины $p = k + 1$, а матрица X имеет размерность nxp. В новое слагаемое входят - матрица U (размерности nxs) и вектор случайных эффектов ξ (размерности s).

Случайные эффекты с одним уровнем

Рассмотрим случай, когда эффекты групируются по одному уровню. Из примера с операцией выше это может быть группировка по больнице, в которой производилась операция. Пусть у нас есть T больниц, для которых можем выписать T групп уравнений вида:

$$\vec{y}_t = X_t\vec{\beta} + U_t\vec{\xi}_t + \vec{\varepsilon}_t, \quad t = 1, \ldots, T,$$

где для каждого t задается набор из n_t уравнений (n_t - количество набродений в больнице t). Размерности величин, входящих в эту запись меняются соответствующим образом. Например здесь X_t - матрица размерности n_txp. От X_t и U_t ожидается, что они будут максимального ранга, с числом строк, не меньшим числа столбцов.

Альтернативный способ трактовать эту модель - лонгитюдное исследование, в котором мы в течении времени наблюдаем за одной и той же группой из T человек.

Мы будем делать следующие предположения касательно этой модели:

$$\vec{\varepsilon}_1, \ldots, \vec{\varepsilon}_T \overset{i.i.d.}{\sim} N_{n_t}(0, \sigma^2 I),$$

$$\vec{\xi}_1, \ldots, \vec{\xi}_T \overset{i.i.d.}{\sim} N_s(0, \Gamma).$$

Также мы предпологаем независимость этих последовательностей, так, чтобы выполнялось:

$$\mathbb{E}\vec{\varepsilon}_t\vec{\varepsilon}_{t'}^T = \mathbb{E}\vec{\varepsilon}_t\vec{\xi}_{t'}^T = \mathbb{E}\vec{\xi}_t\vec{\varepsilon}_{t'}^T = \mathbb{E}\vec{\xi}_t\vec{\xi}_{t'}^T = 0, \quad t \neq t'.$$

Также укажем, что эту можель можно переписать в виде:

$$\vec{y}_t = X_t\vec{\beta} + \tilde{\vec{\varepsilon}}_t,$$

где $\tilde{\vec{\varepsilon}}_t \sim N_{n_t}(0, \sigma^2 I + U_t \Gamma U_t^T)$.

Параметры σ^2 и Γ не предполагаются известными, но мы будем полагать, что $\Gamma \equiv \Gamma(\vec{\alpha})$ параметризуется некоторым вектором $\vec{\alpha}$.

Количество морских жифотных на пляже

В качестве практического примера возьмём рассмотренный в книжке *Мастицкого С.Э., Шитикова В.К. "Статистический анализ и визуализация данных с помощью R"* (опубликованной в декабре 2014 г. в электронном виде в рамках лицензии Creative Commons) о количестве различных видов, найденных на ряде пляжей. В нём будут использоваться данные, которые можно взять здесь.

Посмотрим на эти данные:

```
RIKZ <-read.table(file = "RIKZ.txt", header = TRUE, dec = ".")
head(RIKZ)
```

Здесь *Richness* описывает количество обнаруженных видов, *Exposure* - индекс степени морских возмущений, учитывающий сразу множество факторов и *NAP* - высота пляжа над средним уровнем прилива. *Beach* указывает с какого пляжа брали пробу (взято по 5 проб с каждого). Требуется оценить *Richness* по *NAP* и *Exposure*.

Наивный подход состоит в применении обыной линейной регрессии:

```
fit <- lm(Richness ~ NAP + Exposure, data = RIKZ)
summary(fit)
```

Однако он будет некорретным по уже описанным выше причинам.

Модель со случайным свободным членом

Такая модель будет использовать местоположение пляжа как случайных фактор, однако предполагать, что весь эффект его влияния сконцентрирован в сдвиге свободного члена:

$$\vec{y}_t = X_t\vec{\beta} + \vec{1}\xi_t + \vec{\varepsilon}_t, \quad t = 1, \ldots, T,$$

$$\begin{bmatrix} y_{t1} \\ y_{t2} \\ y_{t3} \\ y_{t4} \\ y_{t5} \end{bmatrix} = \begin{bmatrix} 1 & NAP_{t1} \\ 1 & NAP_{t2} \\ 1 & NAP_{t3} \\ 1 & NAP_{t4} \\ 1 & NAP_{t5} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \xi_t + \begin{bmatrix} \varepsilon_{t1} \\ \varepsilon_{t2} \\ \varepsilon_{t3} \\ \varepsilon_{t4} \\ \varepsilon_{t5} \end{bmatrix}$$

Здесь $\Gamma(\vec{\alpha}) = \alpha^2$ (α - дисперсия случайного свободного члена) и матрица ковариации ошибок y_t

$$\begin{bmatrix} \sigma^2 + \alpha^2 & \alpha^2 & \alpha^2 & \alpha^2 & \alpha^2 \\ \alpha^2 & \sigma^2 + \alpha^2 & \alpha^2 & \alpha^2 & \alpha^2 \\ \alpha^2 & \alpha^2 & \sigma^2 + \alpha^2 & \alpha^2 & \alpha^2 \\ \alpha^2 & \alpha^2 & \alpha^2 & \sigma^2 + \alpha^2 & \alpha^2 \\ \alpha^2 & \alpha^2 & \alpha^2 & \alpha^2 & \sigma^2 + \alpha^2 \end{bmatrix}$$

Для расчета этой модели воспользуемся функцией `lme()` из пакета **nlme**, которая, в отличие от функции `lm()`, позволяет задать случайные эффекты. В частности, параметр `~1|Beach` определяет случайный свободный член модели (справа от знака '|' задается предикаторная переменная):

```
library(nlme)
RIKZ$fBeach <-factor(RIKZ$Beach)
fit.1 <- lme(Richness ~ NAP, random = ~1 | fBeach, data=RIKZ)
summary(fit.1)
```

Нарисуем график, отражающий смысл этой модели:

```
# Получи оценки наших данных
F0 <-fitted(fit.1, level=0)
F1 <-fitted(fit.1, level=1)
# Упорядочим по `NAP`
I <- order(RIKZ$NAP)
NAPs <- sort(RIKZ$NAP)
# Нарисуем график функции `NAP -> F[0]` для нулевого уровня
# (не указан пляж)
plot(NAPs, F0[I], lwd = 2, col="blue", type="l", ylim = c(0, 22), xlab="NAP", ylab="Число видов")
for (i in 1:9){
  x1 <- RIKZ$NAP[RIKZ$Beach==i]
  y1 <- F1[RIKZ$Beach==i]
  K <- order(x1)
  lines(sort(x1),y1[K])}
```

Жирная синяя линия определяет центральную тенденцию зависимости числа видов от высоты над уровнем прибоя. К свободному члену этой зависимости добавлены (сучетом знака!) случайные эффекты местоположения каждого пляжа.

Модель со случайным свободным членом и коэффициентом угла наклона

В этой модели предположим, что от пляжа также зависит характер зависимости между *Richness* и *NAP*, т.е. будем кроме свободного члена корректировать угол наклона. Модель будет иметь следующий вид:

$$\begin{bmatrix} y_{t1} \\ y_{t2} \\ y_{t3} \\ y_{t4} \\ y_{t5} \end{bmatrix} = \begin{bmatrix} 1 & NAP_{t1} \\ 1 & NAP_{t2} \\ 1 & NAP_{t3} \\ 1 & NAP_{t4} \\ 1 & NAP_{t5} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} 1 & NAP_{t1} \\ 1 & NAP_{t2} \\ 1 & NAP_{t3} \\ 1 & NAP_{t4} \\ 1 & NAP_{t5} \end{bmatrix} \begin{bmatrix} \xi_{t0} \\ \xi_{t1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{t1} \\ \varepsilon_{t2} \\ \varepsilon_{t3} \\ \varepsilon_{t4} \\ \varepsilon_{t5} \end{bmatrix}.$$

Здесь

$$\Gamma(\vec{\alpha}) = \begin{bmatrix} \alpha_1^2 & \alpha_{12} \\ \alpha_{12} & \alpha_2^2 \end{bmatrix},$$

где α_1^2 и $\alpha_2 62$ - дисперсии случайного свободного члена и случайного наклона, а α_{12} - их ковариация.

Тогда матрица ковариации:

$$\sigma^2 I + \alpha_1^2 \vec{1}\vec{1}^T + \alpha_{12} \left(\begin{bmatrix} NAP_{t1} \\ NAP_{t2} \\ NAP_{t3} \\ NAP_{t4} \\ NAP_{t5} \end{bmatrix} \vec{1}^T + \vec{1} \begin{bmatrix} NAP_{t1} \\ NAP_{t2} \\ NAP_{t3} \\ NAP_{t4} \\ NAP_{t5} \end{bmatrix}^T \right) + \alpha_2^2 \begin{bmatrix} NAP_{t1} \\ NAP_{t2} \\ NAP_{t3} \\ NAP_{t4} \\ NAP_{t5} \end{bmatrix} \begin{bmatrix} NAP_{t1} \\ NAP_{t2} \\ NAP_{t3} \\ NAP_{t4} \\ NAP_{t5} \end{bmatrix}^T .$$

]

Для использования новой модели нам понадобится изменить параметр `random` на `~1 + NAP | fBeach`:

```
fit.2 <-lme(Richness ~ NAP, data = RIKZ, random = ~1 + NAP | fBeach)
summary(fit.2)
```

И нарисуем для неё аналогичные графики:

```
# Получи оценки наших данных
F0 <-fitted(fit.2, level=0)
F1 <-fitted(fit.2, level=1)
# Упорядочим по `NAP`
I <- order(RIKZ$NAP)
NAPs <- sort(RIKZ$NAP)
# Нарисуем график функции `NAP -> F[0]` для нулевого уровня
# (не указан пляж)
plot(NAPs, F0[I], lwd = 2, col="blue", type="l", ylim = c(0, 22), xlab="NAP", ylab="Число видов")
for (i in 1:9){
  x1 <- RIKZ$NAP[RIKZ$Beach==i]
  y1 <- F1[RIKZ$Beach==i]
  K <- order(x1)
  lines(sort(x1),y1[K])}
```

Смешанные модели

Приведем `Exposure` принимает только три разных значения и имеет смысл рассматривать его как качественную переменную. При этом, так как только один пляж имеет `Exposure == 8`, приведем сведем все данные к двум факторам `<= 10` и `> 10`:

```
RIKZ$fExp <- RIKZ$Exposure
RIKZ$fExp[RIKZ$fExp==8] <- 10
RIKZ$fExp <- factor(RIKZ$fExp, levels = c(10,11))
```

Теперь рассмотрим модель, использующую также и эту новую переменную:

```
fit.3 <-lme(Richness ~ 1 + NAP * fExp, random = ~1 + NAP | fBeach, data = RIKZ)
summary(fit.3)
```

Как всегда, здесь мы имеем некоторую (линейную) зависимость количества обнаруженных видов *Richness* с высотой пляжа *NAP* и интенсивностью приливных возмущений *fExp*. Это наши фиксированные факторы. Часть погрешностей объясняется вариацией между местообитаниями, для чего мы объявляем случайным эффектом *fBeach* и оценивает его вклад в виде стандартного отклонения нормального распределения.

Является ли найденная модель оптимальной? Можно попробовать расчитать ещё две модели: в первой исключаем член `NAP:fExp` , оказавшийся незначительным:

```
fit.4 <-lme(Richness ~ 1 + NAP + fExp, random = ~1 + NAP | fBeach, data = RIKZ)
summary(fit.4)
```

Во второй оставляем только свободный случайный член:

```
fit.5 <-lme(Richness ~ 1 + NAP + fExp, random = ~1 | fBeach, data = RIKZ)
summary(fit.5)
```

К сожалению, модели с разными случайными эффектами имеют различные функции правдоподобия и не могут быть сравнены согласно использующим правдоподобие критериям типа AIC.

Оценки максимального правдоподобия для моделей со смешанными эффектами

Укажем функцию правдоподобия для случая, когда наша модель имеетследующий вид:

$$\vec{y}_t = X_t\vec{\beta} + \tilde{\vec{\varepsilon}}_t,$$

где $\tilde{\vec{\varepsilon}}_t \sim N_{n_t}(0, \sigma^2 I + U_t \Gamma U_t^T)$.

Обозначив $V_t = V_t(\sigma^2, \vec{\alpha}) = \sigma^2 I + U_t\Gamma(\vec{\alpha})U_t^T = \sigma^2(I + U_t\tilde{\Gamma}(\vec{\alpha})U_t^T) = \sigma^2\tilde{V}_t(\vec{\alpha}), \tilde{\Gamma} = \frac{1}{\sigma^2}\Gamma$ можно выписать функцию правдоподобия для нормального распределения:

$$F(\beta, \sigma^2, \vec{\alpha}) = \prod_{t=1}^{T}(2\pi)^{-n_t/2}|V_t(\sigma^2, \vec{\alpha})|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\vec{y}_t - X_t\vec{\beta})^T V_t(\sigma^2, \vec{\alpha})^{-1}(\vec{y}_t - X_t\vec{\beta})\right\}.$$

Заменив β на его оценку максимального правдоподобия

$$\hat{\beta} = \hat{\beta}(\sigma^2, \vec{\alpha}) = \left(\sum_{t=1}^{T} X_t^T V_t(\sigma^2, \vec{\alpha})^{-1} X_t \right)^{-1} \sum_{t=1}^{T} X_t^T V_t(\sigma^2, \vec{\alpha})^{-1} y_t,$$

эту функцию правдоподобия можно привести к логарифмической функции правдоподобия, зависящей лишь от параметров (σ^2, α):

$$\ln L(\hat{\beta}(\sigma^2, \vec{\alpha}), \sigma^2, \vec{\alpha}) = \sum \left\{ -\frac{n_t}{2} \ln(2\pi) - \frac{1}{2} |V_t(\sigma^2, \vec{\alpha})| - \frac{1}{2} \hat{\varepsilon}_t V_t(\sigma^2, \vec{\alpha})^{-1} \hat{\varepsilon}_t \right\},$$

где $\hat{\varepsilon}_t = y_t - X_t \hat{\beta}$.

Для более общей модели вида

$$\vec{y} = X\vec{\beta} + U\vec{\xi} + \vec{\varepsilon}$$

с $\vec{\varepsilon} \sim N_t(0, V), \vec{\xi} \sim N_s(0, \Gamma),$
или, что тоже самое,

$$\vec{y} = X\vec{\beta} + \tilde{\vec{\varepsilon}}$$

с $\tilde{\vec{\varepsilon}} \sim N_T(0, W), W = V + U\Gamma U^T$
логарифчическую функцию правдоподобия можно выписать в виде:

$$L(\hat{\vec{\beta}}(\vec{\alpha}), \vec{\alpha}) = -\frac{1}{2} \left(\ln |W(\alpha)| + \hat{\vec{\varepsilon}}^T W(\alpha)^{-1} \hat{\vec{\varepsilon}} \right),$$

где $\hat{\vec{\beta}} = (X^T W(\alpha)^{-1} X)^{-1} X^T W(\alpha)^{-1} \vec{y}$
и $\hat{\vec{\varepsilon}} = \vec{y} - X\hat{\vec{\beta}}$.

REML (Restricted Maximum Likelihood Estimation)

Метод максимального правдоподобия приводит к смещенным оценкам вариационных параметров (σ^2, $\vec{\alpha}$). Чтобы этого избежать используют метод *Restricted Maximum Likelihood Estimation*, строящий свои оценки на методе максимального правдоподобия, применённого не ко всем данным, а к некоторому их преобразованию. Можно показать, что в таком случае логарифмическая функция правдоподобия будет иметь вид:

$$L(\hat{\vec{\beta}}(\vec{\alpha}), \vec{\alpha}) = -\frac{1}{2}\left(\ln|W(\alpha)| + \ln|X^T W(\alpha)^{-1} X| + \hat{\vec{\varepsilon}}^T W(\alpha)^{-1} \hat{\vec{\varepsilon}}\right).$$