

Более сложные модели: обобщенная линейная модель (GLM) и иерархические регрессионные модели.

Более сложные регрессионные модели применяются в тех случаях, когда диагностика показала несостоятельность простых регрессионных моделей. Вот несколько наиболее распространенных случаев, когда применяются сложные регрессионные модели:

1. *Нелинейность модели*
2. *Ненормальное распределение остатков*
3. *Неодинаковое распределение остатков*
4. *Зависимость остатков*

В данной работе наибольшее внимание уделяется первому случаю, когда модель является по тем или иным характеристикам является нелинейной. Рассматриваются два случая:

- 1) Зависимая переменная связана с предикторами через некоторую нелинейную функцию, в этом случае применяется обобщенная линейная модель
- 2) Структура данных предполагает наличие групп наблюдений, в этом случае применяется иерархические регрессионные модели

Обобщенная линейная модель (Generalized Linear Model)

Обобщенная линейная модель – это модель вида:

$$E(Y) = g^{-1}(X\beta)$$

В этой модели зависимая переменная Y принадлежит экспоненциальному семейству, а монотонная дифференцируемая функция g называется функцией связи. Плотность распределения в экспоненциальном семействе определяется соотношением:

$$f(y; \theta, \varphi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\varphi)} + c(y, \varphi) \right\}$$

На основании соответствующих распределений различают следующие виды обобщенных линейных моделей:

Модель	Функция связи	Распределение зависимой переменной
Общая линейная модель	$g(\mu) = \mu$	Нормальное
Лог-линейная модель	$g(\mu) = \ln \mu$	Пуассоновское распределение
Логистическая модель	$g(\mu) = \ln \frac{\mu}{1 - \mu}$	Биноминальное распределение
Пробит-анализ	$g(\mu) = \Phi^{-1} \mu$	Биноминальное распределение
Анализ выживаемости	$g(\mu) = \mu^{-1}$	Гамма-распределение, экспоненциальное распределение

Обобщенные линейные модели в R

Для построения обобщенных линейных моделей в R используется функция `glm`, которая в общем виде строится следующим образом:

```
glm(formula, family...)
```

Иерархические регрессионные модели (Multilevel)

Иерархическая регрессия применяется в тех случаях, когда модель имеет межгрупповую вариацию. Ключ к пониманию иерархических регрессионных моделей заключается в понимании того, как принадлежность к определенной группе приводит к дополнительной вариации в модели. В иерархических моделях различают три члена общей дисперсии:

- Первый член τ_{00} в дисперсии образуется, когда принадлежность к определенной группе приводит к дополнительной вариации в модели – случай, когда группы различаются в среднем по предикторам.
- Второй член τ_{11} возникает, когда группы варьируются по наклону в линейной зависимости зависимой переменной от предиктора.
- Третий член σ – внутригрупповая дисперсия, она показывает, как реальное значение отличается от предсказываемого.

Шаги при построении иерархической регрессии

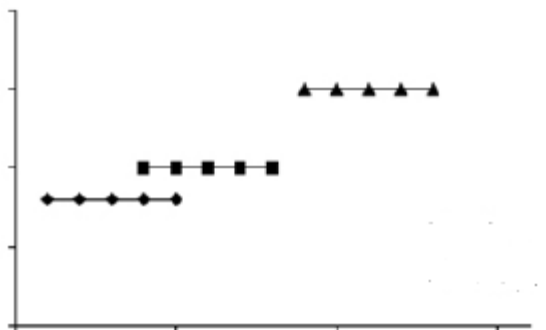
Шаг 1. Строится иерархическая регрессия с включением только вариации по свободному члену.

$$Y_{ij} = \beta_{0j} + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

В этой модели $Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}$ предполагается, что зависимая переменная является функцией общего свободного члена γ_{00} и двух ошибок: межгрупповой дисперсии u_{0j} и внутригрупповой дисперсии r_{ij} . В соответствии с этой моделью оцениваются два параметра – межгрупповая дисперсия τ_{00} и внутригрупповая – σ .

Ниже представлен графический пример структуры данных для этого случая:



Оценка ICC. Межклассовый коэффициент корреляции – $ICC = \frac{\tau_{00}}{\tau_{00} + \sigma}$. Данный коэффициент показывает, какая часть от общей дисперсии можно объяснить варьированием среднего в группах.

Значимость τ_{00} . В процессе анализа данных на данном этапе может встать вопрос, насколько значимо межгрупповая дисперсия отличается от нуля. Для ответа на этот вопрос строится модель без включения варьирования по свободному члену. После этого через дисперсионный анализ ANOVA сравниваются модели между собой. Если обнаруживается значимое различие между моделями, это означает, что τ_{00} отличается от нуля, в противном случае различие признается незначимым.

Шаг 2. Как правило, в модели зависимая переменная связана с предикторами более сложным образом. Рассмотрим для примера следующую модель:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

В данной модели уравнения означают следующее:

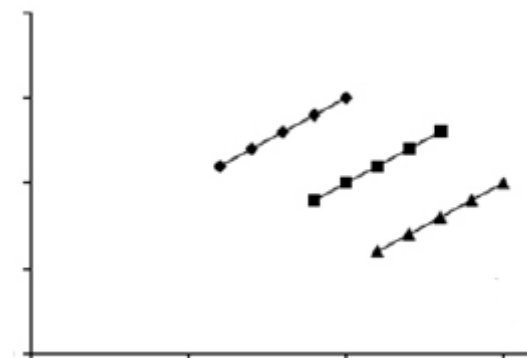
- 1) Первое уравнение моделирует зависимость переменной Y_{ij} от свободного члена β_{0j} - среднего в группе j , предиктора X_{ij} , входящего с некоторым коэффициентом β_{1j} (также определяемого принадлежностью к конкретной группе j) в модель, плюс ошибка r_{ij} - внутригрупповая вариация.
- 2) Второе уравнение показывает, что свободный член β_{0j} есть функция от некоторой константы γ_{00} , группового предиктора Z_j , входящего с некоторым коэффициентом γ_{01} и межгрупповой вариации u_{0j}
- 3) Третье уравнение показывает, что коэффициентом β_{1j} – линейный наклон предиктора X_{ij} есть константа γ_{10}

Если переписать эти уравнения в одно, получим:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + u_{0j} + r_{ij}$$

Мы видим, что данное уравнение отличается от простого регрессионного наличием межгрупповой вариации u_{0j} .

Ниже представлен графический пример структуры данных для этого случая:



Шаг 3. На третьем шаге усложним модель, введя вариацию по линейному наклону.

Рассмотрим следующую модель в качестве примера:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \beta_{2j}T_{ij} + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

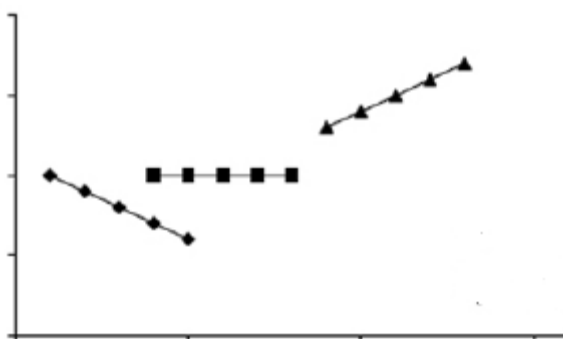
$$\beta_{2j} = \gamma_{20} + u_{2j}$$

Последнее уравнение означает, что наклон в линейной зависимости между зависимой переменной Y_{ij} и предиктором T_{ij} варьируется от группы к группе.

Перепишем эти уравнения в одно:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{20}T_{ij} + \gamma_{01}Z_j + u_{0j} + u_{2j}T_{ij} + r_{ij}$$

Ниже представлен графический пример структуры данных для этого случая:



Значимость τ_{12} . Для проверки значимости межгрупповой вариации по наклону по второму предиктору строится модель без включения вариации по этому предиктору и с включением, после чего модели сравниваются через дисперсионный анализ ANOVA. Если обнаруживается значимое различие между моделями, это означает, что τ_{12} отличается от нуля, в противном случае различие признается незначимым.

Иерархическая регрессия в R

Для построения иерархического анализа в статистической среде R существуют специальные модули. Рассмотрим модульный пакет multilevel.

Для загрузки этого пакета к среде используется функция:

```
library(multilevel)
```

Ниже приведены коды для построения иерархической регрессии на каждом шаге:

- 1) `Model1 = lme(Y~1, random =~1|group)`
- 2) `Model2 = lme(Y~X+Z, random =~1|group)`
- 3) `Model3 = lme(Y~X+Z+T, random =~T|group)`