МОДЕЛИ МНОГОМЕРНЫХ ПРОЦЕССОВ С ГЕТЕРОГЕННОСТЬЮ

3.1 Модель коррелированной уязвимости

Описание рассмотренного ранее процесса выхода из строя элементов в неоднородной популяции требует обобщения на случай наблюдения нескольких характеристик. Наиболее интересным в является изучение поведения не независимых, как это обычно делается, элементов, а элементов, свойства которых статистически связаны. Примером такого обобщения в биометрике является изучение продолжительности жизни и состояния здоровья родителей и детей, братьев и сестёр. В последнее время в связи с появлением результатов крупных популяционных исследований особый интерес представляет изучение пар монозиготных и дизиготных близнецов с целью выявить роль генов и условий окружающей среды в изменении состояния здоровья и рисков, опасных для жизни человека [23, 24].

Рассмотрим модель, в которой индивидуальные уязвимости двух человек коррелированны [25, 26]. Будем считать, что для каждого человека справедлива модель пропорционального риска с гамма распределённой уязвимостью, причем при фиксированных величинах уязвимости продолжительности их жизней независимы. Условное распределение продолжительностей жизни равно

$$S\{T_1 > t_1, T_2 > t_2 \mid z_1, z_2\} = S\{T_1 > t_1 \mid z_1\} S\{T_2 > t_2 \mid z_2\} \qquad (3.1)$$
$$= \exp\left(-z_1 H_{01}(t_1) - z_2 H_{02}(t_2)\right),$$

где T_1, T_2 - случайные величины, соответствующие длительности жизни первого и второго человека соответственно, $H_{01}(t)$ и $H_{02}(t)$ обозначают соответствующие базовые кумулятивные риски, нормированные так, чтобы среднее значение уязвимости каждого человека было равно единице. Относительно индивидуальных уязвимостей Z_1 и Z_2 сделаем предположение, что они представимы в виде $Z_i = \sum_{j=1}^{n}\left(Y_{0j} + Y_{ij}\right)$, $i = 1,2$, где величины Y_{0j}, Y_{1j} и Y_{2j}, $j = 1,...,n$ являются независимыми гамма распределёнными случайными величинами с параметрами k_{oj}, k_{1j}, k_{2j} и единым параметром l. Таким образом, случайные величины Z_1 и Z_2 имеют гамма распределение с тем же параметром l и параметром $k = l$ поскольку среднее значение уязвимости равно единице. Удобно предположить, что и $k_{1j} = k_{2j}$. Коэффициент корреляции между уязвимостями Z_1 и Z_2 вычисляется по формуле

$$r_z = \frac{1}{s_z^2}\mathrm{E}\left(\left(Z_1 - \mathrm{E}\,Z_1\right)\left(Z_2 - \mathrm{E}\,Z_2\right)\right)$$

$$= \frac{1}{s_z^2}\mathrm{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\left(Z_{oi} - \mathrm{E}\,Z_{oi} + Z_{1i} - \mathrm{E}\,Z_{1i}\right)\left(Z_{oj} - \mathrm{E}\,Z_{oj} + Z_{2j} - \mathrm{E}\,Z_{2j}\right)\right)$$

$$= \frac{1}{s_z^2}\sum_{j=1}^{n} s_{0j}^2$$

где s_z^2 - дисперсия уязвимости, s_{0j}^2 - дисперсия j-того члена в

представлении уязвимости. Формулу для коэффициента корреляции можно переписать в виде

$$r_z = \sum_{j=1}^{n} r_j g_j^2 ,$$

где $g_j^2 = \left(s_{0j}^2 + s_{1j}^2\right) / s_z^2$ есть отношение дисперсии компоненты Y_{0j} к дисперсии уязвимости, $r_j = s_{0j}^2 / \left(s_{0j}^2 + s_{1j}^2\right)$ является коэффициентом корреляции *j*-тых членов в представлении уязвимостей Z_1 и Z_2. Параметры гамма распределений связаны с соответствующими дисперсиями и коэффициентами корреляции соотношениями: $l = s_z^{-2}$, $k_{0j} = r_j \left(g_j / s_z\right)^2$, $k_{1j} = \left(1 - r_j\right)\left(g_j / s_x\right)^2$.

Усредняя двумерную функцию дожития (3.1) по гамма распределённым коррелированным уязвимостям получим двумерное распределение длительности жизни пары из гетерогенной популяции, обладающей различными кумулятивными рисками

$$S(t_1, t_2) = \left(1 + s_z^2 \left(H_{01}(t_1) + H_{02}(t_2)\right)\right)^{-s_z^{-2} \sum_{j=1}^{n} r_j g_j^2}$$
$$\times \left(1 + s_z^2 H_{01}(t_1)\right)^{-s_z^{-2} \sum_{j=1}^{n} \left(1 - r_j\right) g_j^2}$$
$$\times \left(1 + s_z^2 H_{02}(t_2)\right)^{-s_z^{-2} \sum_{j=1}^{n} \left(1 - r_j\right) g_j^2} .$$

Принимая во внимание выражение для вероятности безотказной работы элемента в популяции с гамма распределённой гетерогенностью, полученное в предыдущей главе, запишем соотношение

$H_{0i}(t) = \left(S_i(t)^{s_z^2} - 1\right) / s_z^2$, i=1,2 из которого следует представление двумерной функции дожития через одномерные функции дожития

$$S(t_1, t_2) = \frac{S_1(t_1)^{1-r_z} S_2(t_2)^{1-r_z}}{\left(S_1(t_1)^{-s_z^2} + S_2(t_2)^{-s_z^2} - 1\right)^{r_z / s_z^2}}.$$

Коэффициент корреляции r_z представляет собой комбинацию коэффициентов корреляции отдельных компонент уязвимости $r_z = \sum_{j=1}^{n} r_j g_j^2$. В практических приложениях эти коэффициенты корреляции могут быть известны экспериментально или теоретически. Например, при анализе дожития близнецов возможно рассмотреть компоненты уязвимости, связанные с аддитивным генетическим эффектом, доминантным генетическим эффектом, эпистатическим генетическим эффектом, эффектами общего и индивидуального окружения [25]. Из постулатов количественной генетики следуют следующие значения коэффициентов корреляции для монозиготных и дизиготных близнецов [27]

эффект	аддитивный	доминантный	эпистатический	общеее окружение	индивидуальное окружение
монозиготные	1.0	1.0	1.0	1.0	0.0
дизиготные	0.5	0.25	x	1.0	0.0

Используя эту информацию остальные неизвестные параметры модели определяются из экспериментальных данных.

3.2 LCM - модель латентных классов

Время безотказной работы или жизни является примером непрерывной характеристики, изучаемой в неоднородной популяции. Часто интерес представляют категориальные переменные, выражающие признаки наличия заболевания, состояния здоровья, выраженные в некоторой условной шкале и так далее. При этом удобно представлять результаты единичных наблюдений в виде М – мерных дискретных величин. Каждая координата соответствует градациям своей номинальной шкалы, совокупность всех градаций образует М – мерную таблицу сопряжённости. При анализе подобных данных основной интерес представляет оценка и содержательная интерпретация вероятностей попадания наблюдений в отдельные ячейки таблицы сопряженности. В этой главе опишем модель латентных классов (LCM модель), применяемую для анализа дискретной ненаблюдаемой неоднородности [18].

Сначала допустим, что наблюдаются две дискретные переменные с I и J градациями. Результаты наблюдений обычно представляются двумерной таблицей сопряженности, состоящей из IxJ ячеек. Обозначим через p_{ij} – вероятность попадания наблюдения в соответствующую ячейку. Предположим, что для некоторого K существуют векторы $\Theta_a > 0$, $p_i^{(a)} > 0$, $q_j^{(a)} > 0$ (α=1,…,K; i=1,..,I; j=1,…,J) такие, что $\sum_{a=1}^{K} \Theta_a = 1$, $\sum_{i=1}^{I} p_i^{(a)} = 1$, $\sum_{j=1}^{J} q_j^{(a)} = 1$ и справедливо представление

$$p_{ij} = \sum_{a=1}^{K} \Theta_a \, p_i^{(a)} q_j^{(a)} \, .$$

Такая запись означает, что существуют K ненаблюдаемых классов, в каждом из которых строки и столбцы таблицы сопряженности независимы (условная независимость). Каждый эксперимент, результат которого наблюдается, соответствует одному классу, вектор Θ_α задаёт априорные вероятности принадлежности эксперимента к различным классам, произведения $p_i^{(a)} q_j^{(a)}$ задают условные вероятности исхода, соответствующего ячейке ij в различных классах.

Если таблица сопряженности соответствует не двум, а M переменным, то вероятность попадания наблюдения в конкретную ячейку в рамках LCM модели может быть записана в виде

$$p_{i_1 \cdots i_M} = \sum_{a=1}^{K} \Theta_a \prod_{j=1}^{M} p_{i_j}^{(a)}$$

где $p_{i_j}^{(a)}$ - вероятность попадания наблюдения по переменной j в ячейку i_j при условии, что наблюдение принадлежит классу α. Оценок параметров в модели латентных классов рассмотрены в главе 4.

3.3 GOM - модель степени принадлежности

Обобщением модели латентных классов служит модель, использующая идеи размытых множеств. При конструировании такой модели предполагается, что условия проведения каждого эксперимента соответствуют не одному, а одновременно нескольким клас-

сам. Степень принадлежности эксперимента конкретному классу определяется функцией принадлежности, которая предполагается уникальной для каждого эксперимента. При этом, удаётся избежать противоречий между реальностью и гипотезой о принадлежности эксперимента гипотетическому “чистому классу”, который может быть совершенно условным и не существовать в природе.

При использовании статистических моделей, основанных на размытых множествах, совокупность наблюдений интерпретируется как наблюдения в гетерогенной статистической совокупности, а функция принадлежности конкретного наблюдения различным классам описывает неоднородность соответствующего эксперимента или объекта. Совокупность функций принадлежности, соответствующих всем изучаемым объектам, характеризует неоднородность всей совокупности объектов или популяции. Функции принадлежности отдельных объектов характеризуют дополнительно индивидуальную гетерогенность, присущую конкретному эксперименту или индивидууму, в отличие от популяционной гетерогенности, присущей совокупности объектов – популяции.

Независимые наблюдения

Модели, основанные на размытых множествах, были введены в 70х годах для анализа категориальных данных и получили название GOM-модели (от grades of mambership) [19, 20]. Следуя принятым обозначениям представим каждое наблюдение в виде целочисленного вектор с J координатами, каждая из которых принимает L_j зна-

чений. Это могут быть дискретизованные значения непрерывной величины, либо кодировки номинальных величин как при обработке анкетной информации. В последнем случае каждую координату можно понимать как “вопрос”, а значение координаты как конкретный “ответ” на это вопрос. Пусть имеется K базовых классов. Функцию принадлежности эксперимента i базовому классу k обозначим через g_{ik}. Вероятность того, что в эксперименте, принадлежащем базовому классу k, переменная j примет значение l, обозначим через l_{kjl}. За этими вероятностями закрепилось название *профильные вероятности*. На функции g_{ik} и l_{kjl} накладываются естественные ограничения

$$g_{ik} \geq 0,$$

$$\sum_{k=1}^{K} g_{ik} = 1,\ i = 1,\ldots,I,$$

$$l_{kjl} \geq 0,$$

$$\sum_{l=1}^{L_j} l_{kjl} = 1, k = 1,\ldots K,\ j = 1,\ldots J\ .$$

Применение размытых множеств в статистике основано на двух допущениях первое из которых утверждает, что вероятность исхода конкретного эксперимента является линейной комбинацией профильных вероятностей, то есть справедливо соотношение

$$P\left\{Y_{ijl} = 1\right\} = \sum_{k=1}^{K} g_{ik}\, l_{kjl}\ ,$$

где Y_{ijl} – случайная величина, принимающую значение 1 если в эксперименте i переменная j примет значение l, и значение 0 в противном случае. Второе условие имеет смысл условной независимости и означает, что для фиксированного i вероятность одновременного наблюдения значений различных переменных равна произведению вероятностей наблюдать каждое значение в отдельности

$$P\left\{Y_{ij_1l_1}=1,\ldots,Y_{ij_Jl_L}=1\right\}=\left(\sum_{k=1}^{K}g_{ik}\,l_{kj_1l_1}\right)\mathbf{L}\left(\sum_{k=1}^{K}g_{ik}\,l_{ij_Jl_{L_1}}\right).$$

С учётом этих условий и обозначив реализацию случайной величины через y_{ijl} запишем вероятность наблюдения конкретной реализации значений всех переменных при проведении I независимых экспериментов – функцию правдоподобия

$$L=\prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{l=1}^{L_j}\left(\sum_{k=1}^{K}g_{ik}\,l_{kjl}\right)^{y_{ijl}}.$$

Строго говоря, эта функция является функцией правдоподобия только если индивидуальные функции принадлежности g_{ik} заданы априори. По этой причине её называют условной функцией правдоподобия в отличие от безусловной функции правдоподобия

$$L_u=\int\prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{l=1}^{L_j}\left(\sum_{k=1}^{K}l_{ijlk}x_k\right)^{y_{ijl}}dP(x),$$

которая получается путём усреднения условной функции правдоподобия по случайной k-мерной вектор- функции принадлежности x, имеющей распределение, удовлетворяющее ограничению

$x_k \geq 0,\ \sum_{k=1}^{K} x_k = 1$. Индивидуальные функции принадлежности g_{ik} рассматриваются при этом как независимые реализации случайного вектора *x*.

Зависимые наблюдения

При конструировании условной функции правдоподобия предполагалось, что результаты исходов различных экспериментов независимы. При анализе данных о состоянии индивидуумов, связанных родственно и генетически, наблюдения оказываются зависимыми и большой интерес представляет изучение степени этой связи в зависимости от генетических и родственных факторов. Один из возможных подходов к применению моделей, основанных на размытых множествах, для анализа данных о родственниках использован в [27] для поиска генов, связанных с наследственным диабетом. При этом в качестве исхода “эксперимента” принималась информация о двух родственниках, сами же пары считались независимы друг от друга. Информация о неполных парах не использовалась, что является существенным ограничением применения этого подхода.

Другая возможность заключается в непосредственном учёте структуры зависимости между наблюдаемыми переменными. Пусть x^1 и x^2 обозначают функции принадлежности для двух статистически зависимых индивидуумов. Безусловная функция правдоподобия для такой пары имеет вид

$$L_u = \int \prod_{j=1}^{J} \prod_{l=1}^{L_j} \left(\sum_{k=1}^{K} x_k^1 l_{kjl} \right)^{y_{1jl}} \left(\sum_{k=1}^{K} x_k^2 l_{kjl} \right)^{y_{2jl}} dP(x^1, x^2) .$$

Допустим, что существуют две такие случайные ненаблюдаемые величины T_1 и T_2, называемые liability, такие, что совместное распределение вектор- функций принадлежности x^1 и x^2 можно представить в виде

$$P(x^1, x^2) = \int F\left(x^1 | t_1\right) F\left(x^2 | t_2\right) h(t_1, t_2) dt_1 dt_2 ,$$

где $F\left(x|t\right)$ - условное многомерное распределение функции принадлежности при фиксированной величине ненаблюдаемой liability. Безусловная функция правдоподобия примет вид

$$L_u = \int \int \prod_{j=1}^{J} \prod_{l=1}^{L_j} \left(\sum_{k=1}^{K} x_k^1 l_{kjl} \right)^{y_{1jl}} \left(\sum_{k=1}^{K} x_k^2 l_{kjl} \right)^{y_{2jl}} f\left(x^1 | t_1\right) f\left(x^2 | t_2\right) h(t_1, t_2) dt_1 dt_2 dx^1 dx^2 .$$

Перегруппировав, запишем

$$L_u = \int \left(\int \prod_{j=1}^{J} \prod_{l=1}^{L_j} \left(\sum_{k=1}^{K} x_k^1 l_{kjl} \right)^{y_{1jl}} \left(\sum_{k=1}^{K} x_k^2 l_{kjl} \right)^{y_{2jl}} f\left(x^1 | t_1\right) f\left(x^2 | t_2\right) dx^1 dx^2 \right) h(t_1, t_2) dt_1 dt_2$$

$$= \int \left(\int \prod_{j=1}^{J} \prod_{l=1}^{L_j} \left(\sum_{k=1}^{K} x_k^1 l_{kjl} \right)^{y_{1jl}} f\left(x^1 | t_1\right) dx^1 \right) \left(\int \prod_{j=1}^{J} \prod_{l=1}^{L_j} \left(\sum_{k=1}^{K} x_k^2 l_{kjl} \right)^{y_{2jl}} f\left(x^2 | t_2\right) dx^2 \right) h(t_1, t_2) dt_1 dt_2 .$$

Функция распределения величин liability $h\left(t_1, t_2\right)$ отражает зависимость между близнецами.

Следуя моделям количественной генетики [27] представим случайные величины T_1 и T_2 в виде разложения на слагаемые, ответственные за аддитивную генетическую компоненту – A, общее окружение – C, и индивидуальное окружение – E, то есть

$T_1=A_1+C_1+E_1$, $T_2=A_2+C_2+E_2$. Возможны и более сложные представления, дополнительно учитывающие, например, доминантный и эпистатический генетические эффекты.

Компоненты A, B и C являются результатом взаимодействия большого числа генов и микро событий, в результате чего каждая из компонент может рассматриваться как нормально распределённая случайная величина. Считая компоненты C, E и A независимыми, а компоненты A_1 и A_2 коррелированными между собой с коэффициентом корреляции r, получим, что величины T_1 и T_2 распределена по нормальному закону с корреляционной матрицей

$$\operatorname{cov}(T_1, T_2) = \begin{pmatrix} s_a^2 + s_c^2 + s_e^2 & r * s_a^2 + s_c^2 \\ r * s_a^2 + s_c^2 & s_a^2 + s_c^2 + s_e^2 \end{pmatrix}$$

где s_a^2, s_c^2 и s_e^2 обозначают соответственно дисперсии аддитивной генетической компоненты, общего и индивидуального окружения. Эти дисперсии считаются одинаковыми для обоих близнецов. Поскольку монозиготные близнецы имеют полностью общий хромосомный набор, а у дизиготных близнецов лишь половина хромосом общая, то коэффициент корреляции r равен единице для монозиготных близнецов и 0.5 для дизиготных [27].

Для описания условного распределения функции принадлежности можно использовать представление $x_i = \Phi(t, a_i)$ (i=1,…,K-1), где $\Phi(t,a)$ - функция распределения нормального закона со средним значением a и единичной дисперсией. При ограничениях

$x_j \geq 0;\ \sum_{j=1}^{K} x_j = 1$ справедливо выражение

$$F(x|t) = P\{x_1 < x_1,...,x_K < x_K\} = \Phi\left(\min_{j=1,...,K} \Phi^{-1}(x_j, a_j)\right).$$

Величины a_j являются параметрами модели, подлежащими определению по данным, причем важны не их абсолютные значения, а расстояние до среднего значения величины liability, которое, следовательно, можно принять равным нулю.

Функция правдоподобия принимает вид

$$L = L_S * L_{MZ} * L_{DZ},$$

где L_S обозначает функцию условного правдоподобия для близнецов, представленных без брата или сестры

$$L_S = \prod_{i \in I(S)} \prod_{j=1}^{J} \prod_{l=1}^{L_j} \left(\sum_{k=1}^{K} g_{ik}\, l_{ijl}\right)^{y_{ijl}}.$$

Безусловные функции правдоподобия L_{MZ} и L_{DZ} для монозиготных и дизиготных пар близнецов определяются выражением

$$L_{zyg} = \prod_{i \in I(zyg)} \int \left(\int \prod_{j=1}^{J} \prod_{l=1}^{L_j} \left(\sum_{k=1}^{K} x_k\, l_{kjl}\right)^{y_{i_1 jl}} f(x|t_1) dx \right)$$
$$\times \left(\int \prod_{j=1}^{J} \prod_{l=1}^{L_j} \left(\sum_{k=1}^{K} y_k\, l_{kjl}\right)^{y_{i_2 jl}} f(y|t_2) dy \right) N(t_1, t_2, \mathrm{cov}_{zyg}) dt_1 dt_2$$

где $N(t_1, t_2, \mathrm{cov})$ - двумерная плотность вероятности нормального распределения с нулевым средним и ковариационной матрицей cov.