

Глава 1 Основные задачи многомерного статистического анализа

1

совек. объектов - почти неизм. св-ва
даны в матрицу (даны объект и неизм.
его св-ва, характеристики):

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$$

кажд-во
признаков

Обычно таких измерений
много $\Rightarrow \bar{X} = (X_{jk})$, j - номер
измер.
 k - размерность
вектора
измерений

$$j = 1, N$$

$$k = 1, d$$

какими св-вами обладает изучаемая совокупность

почти равно, как признаки введут себя
групп. отнес. группы

1) Изучение зависимости признаков
(есть негов. признаки, по введем измеримые задачи)
 $\Rightarrow$ корреляции

2) если связей нет, то какой признак
изм. по отдельности

есть какие-то признаки связ. с
группами, по
Поиск связей признаков

⇒ Регрессионный анализ

6) реальный зазор d достаточно велик (позволяет как-то уменьшить d)

3) Считаем размерности

⇒ факторный анализ (многомерное инкапсулирование, когда сводят к двумерной картине где наглядности)

4) компактное представление данных

2) Глава 2 Гильбертово пространство сущ. величин

1) Описание сущ. пр-во

2) Задача компактной оценки (все можно учесть и забыть)

§ 1. Гильбертово пр-во сущ. величин

вер. пр-во - основная модель теории вероятностей

Пусть $(\Omega, \mathcal{A}, P)$ - вер. пр-во

Опр. Тр-во L_2 сущ. ин-во сущ. велич. ξ , заданных на $(\Omega, \mathcal{A}, P)$: $E(|\xi|^2) < \infty$
ин-отн.

Задача L_2 - ин. пр-во.

вер-во Хильбертово пространство, когда $E_1, E_2 \in \mathcal{P}$ и $E_1 \perp E_2$

Задача дуги отсюда с. в. ξ_1 и ξ_2

из L_2 , если $P(\xi_1 \neq \xi_2) = 0$

в L_2 дуги класс сущ. велич. которые сводят на ин-во вер. пр-во (или сущ. велич. не дуги)

Определим скалярный продукт:

$$\forall \xi_1, \xi_2 \in L_2 \quad (\xi_1, \xi_2) = E(\xi_1 \cdot \xi_2)$$

скал. произв. по определению

Задача 1) $\forall \xi \quad (\xi, \xi) \geq 0$ и $(\xi, \xi) = 0 \Leftrightarrow$

$\xi = 0$ п.н. (почти наверное, т.е. с вероятностью 1)

2) $\forall \xi_1, \xi_2 \in L_2 \quad (\xi_1, \xi_2) = (\xi_2, \xi_1)$

3) $\forall \xi_1, \xi_2, \xi_3 \in L_2 \quad \forall c_1, c_2 \in \mathbb{R}^1$

$$(c_1 \xi_1 + c_2 \xi_2, \xi_3) = c_1 (\xi_1, \xi_3) + c_2 (\xi_2, \xi_3)$$

⇒ (ξ_1, ξ_2) - скал. произв. в ин. пр-ве L_2 (из ин. аксиом)

по скал. произв. можно построить норму

Определим $\forall z \in L_2 \quad \|z\|^2 := (z, z) = E(|z|^2)$
 $\|z\|$ — норма элемента $z \in L_2$

Тогда можно определить понятие сходимости

Впр. последовательность $\{z_n\} \subset L_2$ сходится к элементу $z \in L_2$ в среднем квадратическом, если $\|z_n - z\|^2 = E(|z_n - z|^2) \rightarrow 0, n \rightarrow \infty$

Обозначение: $z_n \xrightarrow{L_2} z$ или $z_n \xrightarrow{с.к.} z$ — среднее квадратическое

Задача Пр-во L_2 можно считать сход. в среднем квадратическом.

если есть последовательность (выбранных пр-во) то она сходится имеет предел к элементу пр-во (выбранных нет)

L_2 с одной стороны выбранных пр-во

L_2 с другой стороны выбранных пр-во, пр-во нормы

$\Rightarrow L_2$ — матричное пр-во

! нормы элементов сходится нормы элементов сходится

Опр. Два элемента $z_1, z_2 \in L_2$ ортогональны, если $(z_1, z_2) = E(z_1 \bar{z}_2) = 0$

или $(z_1, z_2) = E(z_1 \bar{z}_2) = 0$

Замечание: Если $E(z_1) = E(z_2) = 0$, то (центрированы векторы, т.е. константа из них убрана)

$$(z_1, z_2) = 0 \Leftrightarrow \text{cov}(z_1, z_2) = 0$$

т.е. z_1 и z_2 — некоррелированы (отсутствие вариации независимости)

Определим косинус угла между элементами $z_1, z_2 \in L_2$ как число $\frac{(z_1, z_2)}{\|z_1\| \cdot \|z_2\|} =$

$$= \frac{E(z_1 \bar{z}_2)}{\sqrt{E(|z_1|^2) \cdot E(|z_2|^2)}}$$

— это коэффициент корреляции z_1 и z_2 , если $E(z_1) = E(z_2) = 0$

§2 Задача о каноническом базисе

Пусть L — замкнутое линейное подпространство L_2 .

$\eta \in L_2$.

нормальный элемент

Опр. Канонический линейный элемент $\eta \in L_2$ в пр-ве L норм. элемент η :

$$1) \hat{\eta} \in L$$

$$2) \forall \xi \in L$$

$$\| \eta - \hat{\eta} \|^2 = E(|\eta - \hat{\eta}|^2) \leq E(|\eta - \xi|^2) = \|\eta - \xi\|^2$$

$\hat{\eta}$ - наилучшее линейное по критерию

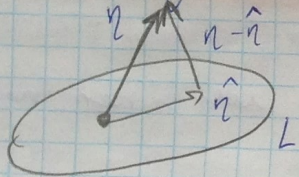
это некоторое экстремальное значение (задача минимизации), можно ли иметь критерий

Лемма о перпендикулярности

$\hat{\eta}$ есть наилучшее линейное значение для $\eta \in L_2$ в пр-ве $L \Leftrightarrow$

$$1) \hat{\eta} \in L$$

$$2) (\eta - \hat{\eta}, \xi) = 0 \quad \forall \xi \in L$$



лучший вектор (точка) - это перпендикуляр на пр-ве

не всегда приближение строится из минимального пр-ва (конкретно) когда выполняется т.е. инвариантность

Глава 3 Многомерное нормальное распределение

Пусть $X = (X_1, \dots, X_n)^T$ - случайный вектор

Случайный вектор X имеет многомерное нормальное (Гауссовское) распределение, если он имеет плотность распределения следующего вида:

$$p(x) = p(x_1, \dots, x_n) = (2\pi)^{-\frac{n}{2}} [\det(A)]^{\frac{1}{2}} \exp \left[-\frac{1}{2} (A(x-m), (x-m)) \right] = (2\pi)^{-\frac{n}{2}} \det(A) \cdot \exp \left[-\frac{1}{2} \sum_{i,j=1}^n a_{ij} (x_i - m_i) (x_j - m_j) \right] \quad (1)$$

где $m = (m_1, \dots, m_n)^T \in \mathbb{R}^n$, $A = (a_{ij})$ - неоп. эрм. матрица (симметричная, положит. определ.)
удов., когда есть х.ф. или плотность (удов. неинтересно, доказывать)

Если $\varphi_X(t) = E[e^{i(X,t)}]$ есть характеристическая ф-ция с.в. вектора X , то $\varphi_X(t) = \exp[i(m,t) - \frac{1}{2}(\Sigma t, t)] =$

$$\exp \left[i \sum_{j=1}^n m_j t_j - \frac{1}{2} \sum_{p,q=1}^n b_{pq} t_p t_q \right]$$

$$t = (t_1, \dots, t_n)^T \in \mathbb{R}^n, \quad \Sigma = A^{-1} \quad (2)$$

известно, что по х.ф. можно считать с помощью производных

Пусть задан х.ф. получаем:

$$m_j = E(X_j) \quad (3)$$

$$b_{pq} = \text{Cov}(X_p, X_q) \quad (4)$$

6 (2) группировки, когда Σ вырожденное
пример

X - имеет много корр. распре с пар. (m, Σ)
 $\Rightarrow$ адекват
 можно показать, что есть не корр. распре-
 в этой паре более низкого порядка

Теорема 1 Если $X \sim N(m, \Sigma)$ и
 $C: R^n \rightarrow R^k$ - линейное
 отображение
 $X \sim N(m, \Sigma)$ то $Y = C \cdot X \sim N(C \cdot m, C \cdot \Sigma \cdot C^T)$

пример:
 X - имеет корр. распре, если $m = (0, \dots, 0)$
 $= 0, \Sigma = I_n$
единичная
матр.

Следствие 1 Если $X \sim N(0, I_n)$ и C есть
 ортогональная проекция X из R^n в R^k , то
 $Y = C \cdot X \sim N(0, I_k)$

кан. метр. coord. извл., если есть орт.
 проекц. (в орт. канонич. базисе) то
 корр. распре. меньшей размерности

Разобьем вектор X на два подвектора
 $X = (X_1, \dots, X_{n_1}, X_{n_1+1}, \dots, X_{n_1+n_2})^T = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}$
 $\Rightarrow m = \begin{pmatrix} m^{(1)} \\ m^{(2)} \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$

Теорема 2 (следствие из предыдущей теоремы)

Если $X \sim N(m, \Sigma) \in R^n$ и $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, то
вектор средних

1) $X^{(1)} \sim N(m^{(1)}, \Sigma_{11})$

2) Во многих задачах интересует, когда эти
 два подвектора независимы

$X^{(1)}$ и $X^{(2)}$ - независимы $\Leftrightarrow \Sigma_{12} = \Sigma_{21}^T = 0$
матр. из нулей
 (некоррелируемость)

при анализе-когда условия не-опт. все-таки
 на группой вектор не зависят

Теорема 3 Если $X \sim N(m, \Sigma)$
вектор
средних
незав. $X^{(1)}$ - не корр.

$X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}$, то условное распределение

$X^{(2)}$ при условии, что $X^{(1)} = x^{(1)}$ описывается

есть многомерное нормальное распределение
 со средним $m^{(2)} + \Sigma_{21} \cdot \Sigma_{11}^{-1} (x^{(1)} - m^{(1)})$

и матрицей ковариации $\Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}$
абсолютно

кто имеет X^2 распределение (то есть
 $g - P_0$, опираясь на предыдущее утверждение)

Теор 4 $X \sim N(m, \Sigma)$, $\det(\Sigma) \neq 0$.

Тогда сугр. величина $(X-m)^T \cdot \Sigma^{-1} (X-m)$ имеет χ^2 -распределение с n степенями свободы
каждое по своему квадратичной форме
 а. в. можно пометить. Кратчай. Максимальные
 линейн. статистики (и в Андерсоне)

(3)

Глава 4. Показатели о корреляционной и регрессионной аналитике.

есть ли зависимость, и если есть, то как ее описать

§ 1. Основные задачи корреляционного регрессионного анализа.

Пусть мы имеем две подбора измерений
 $X = (x_1, \dots, x_n)^T$, $Y = (y_1, \dots, y_n)^T$

Есть две основные задачи:

I. Есть ли связь между X и Y и насколько она сильна?

II. Если связь есть, то найти форму связи

Первая задача решается с помощью тех или иных коэффициентов корреляции.
 => корреляционный анализ

Вторая задача решается методами регрессионного анализа

§ 2. Коэффициент корреляции.

1° (Простой) коэффициент корреляции

Пусть X и Y — случайные величины.

Опр (Простой) коэффициент корреляции независимых случайных величин X и Y называется

$$\rho(X, Y) = \rho_{XY} := \frac{\text{Cov}(X, Y)}{\sqrt{D(X) \cdot D(Y)}} \quad (1)$$
отношение ковариации

Теорема Коэфф. корр. сугр. величин. обладает следующими свойствами

1) $|\rho(X, Y)| \leq 1$

$|\rho| < 0,5$ слабая
 $0,5 < |\rho| < 0,7$ средняя
 $|\rho| > 0,7$ сильная

2) Если X и Y некорр., то $\rho(X, Y) = 0$

ρ^2 где 0,7
 0,5, поэтому
 средний $\rho = 0,7$

3) Если $z_1 = a_1 \cdot X + b_1$, $z_2 = a_2 \cdot Y + b_2$,

$a_1 \cdot a_2 \neq 0$, то $\rho(z_1, z_2) = \rho(X, Y) \cdot \text{sign}(a_1 \cdot a_2)$

4) Если $\rho(X, Y) = +1$, то $\exists a > 0$ и $b \in \mathbb{R}^1$:

$Y = aX + b$ (с вероятностью единицы, без сносок функциональная и почти инвариант)

Если $\rho(X, Y) = -1$, то $\exists a < 0$ и $b \in \mathbb{R}^1$:

$Y = aX + b$

Если $\rho(X, Y) > 0$, то говорят, что X и Y — положительно коррелированы
 $\rho(X, Y) < 0$ — отрицательно коррелируют.

Рассмотрим задачу о нахождении линейной связи $\hat{Y}$ с наименьшей сум. величиной ошибки $\sigma X + b$ (2) в виде сум. величин

Далее эту задачу (см. задачу о перпен. функциях), получаем следующий результат для оптимальной оценки:

$$\hat{Y} = \hat{a} \cdot X + \hat{b} \quad (3)$$

$$\hat{a} = \rho(X, Y) \sqrt{\frac{D(Y)}{D(X)}} \quad (4)$$

$$\hat{b} = E(Y) + \rho(X, Y) \sqrt{\frac{D(Y)}{D(X)}} (X - E(X)) \quad (5)$$

ошибка $e = Y - \hat{Y}$ — минимизируется

В силу задачи о перпен. функциях $e \perp \hat{Y}$

$$Y = \hat{Y} + e$$

$$\text{3.4.6} \quad D(Y) = D(\hat{Y}) + D(e)$$

$$D(\hat{Y}) = [\rho(X, Y)]^2 D(Y) \quad (7)$$

$$D(e) = (1 - [\rho(X, Y)]^2) \cdot D(Y) \quad (8)$$

мера изменчивости суммарной величины — дисперсия

$\hat{Y}$ — то, что можно или через X (X — сум. велич.) и измерить изменчивость за счет X

ρ^2 показывает какую долю изменчивости Y удалось объяснить за счет величины X (из линейных связей) — коэффициент корр.

коэффициент корр. измеряет только линейные связи X на Y

(8) $\Rightarrow 1 - [\rho(X, Y)]^2$ — показывает, какую долю изменчивости Y не удалось объяснить линейными связями X и нужно привлекать другие факторы (тогда объяснить эту часть)

Рассмотрим задачу о нахождении связи между X и Y . Формально реш. $\rho = \rho(X, Y)$ и проверим гипотезу $H_0: \rho = 0$ (9)
 $H_1: \rho \neq 0$

Пуско (X, Y) имеет бivariate нормальное распределение

наим. квадратич. и по критерию zero-модель

Пуско имеет bivariate квадратич. $(X_1, X_2), \dots, (X_n, X_{n+1})$ рассмотрим случай, когда $n=1$ и $X_1 = X, X_2 = Y$
 Естественно рассмотреть bivariate коэффициент корр.

$$R := \frac{S_{XY}}{S_X S_Y} \quad (10) \quad \text{где}$$

по-прежнему на $N-1$, дисперсия на N

$$S_{XY} = \frac{1}{N} \sum_{j=1}^N (X_j - \bar{X})(Y_j - \bar{Y})$$

$$S_x^2 = \frac{1}{N} \sum_{j=1}^N (x_j - \bar{x})^2$$

$$S_y^2 = \frac{1}{N} \sum_{j=1}^N (y_j - \bar{y})^2$$

Удобнее перейти к величине $t_{N-1} :=$
(так удобнее, потому что распределение Стьюдента)

$$= \frac{R}{\sqrt{1-R^2}} \cdot \sqrt{N-1} \quad (11)$$

Теорема Если верна H_0 , то случайная величина t_{N-1} имеет распределение Стьюдента с $(N-1)$ степенями свободы

или закон распределения (то же самое, что закон Гаусса)

Далее по заданному уровню значимости α (это маленькое число) по таблице находим критическое значение $C(\alpha) > 0$

$$: P(|t_{N-1}| > C(\alpha) | H_0) = \alpha \quad (12)$$

Правило принятия решения

1) Если реально полученное значение t_{N-1} (или t_{N-1}^*) статистики t_{N-1} таково, что $|t_{N-1}| > C(\alpha)$, то отвергаем H_0 , т.е. обнаружена зависимость между x и y

2) В противном случае не обнаружена зависимость x и y

В противном случае не обнаружена зависимость x и y

в реальном мире неверно утверждение H_0

Завершение

проверка утверждения с помощью статистики t_{N-1} (или t_{N-1}^*)

Пусть t_{N-1} есть статистика Стьюдента и мы получили в эксперименте некоторое конкретное значение t_{N-1}^*

то значение, которое равно t_{N-1}^*

$$\text{Далее вычисляем } P(|t_{N-1}| \geq |t_{N-1}^*| | H_0) = p$$

- достигнут уровень значимости α (или p value)

Сравниваем p с α , которое мы выбрали

1) Если $p < \alpha$, то отвергаем H_0

2) Если $p \geq \alpha$, то H_0 не противоречит экспериментальным данным

тогда мы можем с уверенностью $1-p$ утверждать, что H_0 верна

$$\alpha = 0.1, 0.05, 0.01$$

Пусть мы внесли по своим методам X и Y значения (т.е. H_0 не отвергается)

А насколько она сильна?

Это измеряется по величине $\rho = \rho(X, Y)$ но ее реально не знаем (измерим по гипергеометрическим)

П.к. ρ оценивается по гипергеометрическим данным, то строим доверительный интервал (какое значение будет наименьшим)

надо учесть, что находить распределение статистики не только при H_0 , но и при альтернативе — тогда строим доверительный интервал

Для того надо знать разность R при $\rho \neq 0$

Финшер предложил рассмотреть следующую статистику:

$$Z := \frac{1}{2} \ln \frac{1+R}{1-R} \quad \text{— } Z\text{-статистика Финшера}$$

(затем ее разбавят средним при H_0)

Финшер показал, что даже при небольших N эта статистика имеет приблизительно нормальное распределение со средним

$$a = \frac{1}{2} \ln \frac{1+\rho}{1-\rho} + \frac{\rho}{2(N-1)}$$

$$\text{и дисперсией } \sigma^2 = \frac{1}{N-3}$$

Z имеет норм. разпр. с центром a и изв. дисперсией $\Rightarrow$ можем найти две границы

По заданной норм. разпр. для заданного $\delta \sim 1$ находим $c(\delta) > 0$:

$$P(|Z-a|/\sigma \leq c(\delta)) \approx \delta$$

асимптотически верно

$$\Rightarrow Z - c(\delta)\sigma < a < Z + c(\delta)\sigma \quad (\text{хотим две границы интервала})$$

сразу есть масса методов приближенного вычисления

$a = a(\rho)$ — монотонно возрастает по ρ

$\Rightarrow \rho$ и получить доверительный интервал

2° Множественный коррел. корреляции

Имеется переменная Y , поведение которой пытаются объяснить с помощью набора факторов $X_1, \dots, X_m$

В нашем курсе будем ограничиваться только линейными моделями

Пусть $\hat{y}$ есть наилучшее линейное приближение y с помощью $X_1, \dots, X_m$, т.е. $\hat{y} = \alpha + \beta_1 X_1 + \dots + \beta_m X_m$

Опр. Множественный коррел. корр. Y и набора факторов $X_1, \dots, X_m$ коэф. это

$$\rho_{Y, X_1, \dots, X_m} = \rho(Y, \hat{y})$$

Обозначим $e = y - \hat{y}$ - ошибка накл. регр.

$$\hat{y} = \rho^2_{y \cdot x_1 \dots x_m} \cdot y$$

$$\mathcal{D}(e) = (1 - \rho^2_{y \cdot x_1 \dots x_m}) \mathcal{D}(y)$$

$\Rightarrow \rho^2_{y \cdot x_1 \dots x_m}$ показывает, какую долю изменчивости y можно объяснить линейными влияниями $x_1, \dots, x_m$

$1 - \rho^2 = 0$, то не удалось объяснить вариационный фактор y ни одним из факторов

3°. Частный коэффициент корреляции

какого-то фактора объясняет одно из факторов, или устранить влияние всех остальных

Хотелось бы измерить величину влияния на y одного из факторов.

$x_1, \dots, x_m$ - объясняющие факторы

Взберем один из факторов x_k .

Обозначим C - набор всех остальных факторов.

Сначала найдем влияние всех остальных факторов, когда не учтем x_k .

Пусть y^c - велич. накл. регр. y с помощью величин из C .

x_k^c - накл. регр. x_k с помощью величин из C .

Обозначим $z_y = y - y^c$

$$z_{x_k} = x_k - x_k^c$$

Опр. Частный коэффициент корреляции $\rho_{y \cdot x_k \cdot C}$ между y и x_k при условии, что y и x_k объяснены факторами из C :

Можно показать, что для $\rho_{y \cdot x_k \cdot C}$ справедливо:

$$1 - \rho^2_{y \cdot x_1 \dots x_k} = (1 - \rho^2_{y \cdot x_1 \dots x_{k-1}}) (1 - \rho^2_{y \cdot x_k \cdot x_1 \dots x_{k-1}})$$

Это соотношение показывает, что если $\rho_{y \cdot x_1 \dots x_{k-1}} = 0$, то $\rho_{y \cdot x_k \cdot x_1 \dots x_{k-1}} = 0$, т.е. влияние x_k на y не объясняется факторами $x_1, \dots, x_{k-1}$.

Простой пример, когда объясняет то такое прост, имеет и другие факторы, которые не объясняют.

По итогам года регистрировались следующие предприятия легкой промышленности:

$y = X_0$ - количество тканей (мил. в каких-то тканях)

X_1 - число производственных предприятий

X_2 - число работников

X_3 - как и X_1 и X_2 производственные предприятия

X_4 - количество работников (число работников, т.е. работников)

Вспомогательные коэффициенты корреляции:

$$R_{01} = 0.105, R_{02} = -0.024, R_{12} = -0.996$$

X_1 и X_2 по 1% не влияют на X_0 - ошибку

Введем более удобное обозначение для $\beta_{0,2} = 0,907$, $\beta_{0,1} = -0,906$

β_0 — число корр. и корр. корр. введем
группа группа

5) Глава 5 Элементы регрессионного анализа

§1 Постановка задачи

Пусть мы имеем одну зависимость переменной y , которую мы хотим объяснить с помощью нескольких факторов $x_1, \dots, x_m$

Предположим, что мы имеем следующее соотношение: $y = f(x_1, \dots, x_m) + \varepsilon$ (1)

где ε — влияние неизвестных факторов

введем, что y — значение переменной y и x — значения факторов $x_1, \dots, x_m$

Основная задача: найти такую функцию $f(x_1, \dots, x_m)$, которая обеспечивает в некотором смысле наилучшее приближение y .

Если наилучшее приближение понимать в среднеквадратическом смысле, то мы имеем тер. вер. известн., то наилучшее приближение y с помощью факторов $x_1, \dots, x_m$ имеет вид: $g(x_1, \dots, x_m) := E(y | x_1 = x_1, \dots, x_m = x_m)$ (2)

— g — регрессия

в реальной задаче задается числом значений y и x .

далее мы будем рассуждать так: пусть $g(x_1, \dots, x_m) = \alpha + \beta_1 x_1 + \dots + \beta_m x_m$ (3)

где $\alpha, \beta_1, \dots, \beta_m$ — неизвестные параметры. Теперь задача в том, чтобы оценить параметры.

Базисная регрессия (линейная) — базисная регрессия по порядку 1-й степени, полиномиальная — 2-й степени — так далее и т.д.

§2 Классическая линейная модель регрессии

Пусть мы хотим объяснить поведение y с помощью факторов $x_1, \dots, x_m$, т.е. $y = g(x_1, \dots, x_m) + \varepsilon$ (1)

Пусть мы имеем N одновременных измерений $(y_i, x_{i1}, \dots, x_{im})$:

$$y_i = g(x_{i1}, \dots, x_{im}) + \varepsilon_i \quad (2)$$

где ε_i — влияние неизвестных факторов — случайная величина.

Основные ограничения

1) Модель линейна по параметрам, т.е.

$$y_i = \alpha + \beta_1 x_{i1} + \dots + \beta_m x_{im} + \varepsilon_i \quad (3)$$

не важно, что X входит линейно, а важно, что параметр β входит линейно

2) Фактор измерен без ошибок

3) $E(\varepsilon_j) = 0 \forall j$ - нет систематического смещения

4) $D(\varepsilon_j) = \sigma^2 \forall j$ - условие гомосkedастности
это они одинаковы для $\forall j$

5) $Cov(\varepsilon_j, \varepsilon_k) = 0, j \neq k$

6) ε_j - имеет нормальное распределение

введем векторные обозначения

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{pmatrix}$$

$$\vec{X}_k = \begin{pmatrix} x_{1k} \\ \vdots \\ x_{Nk} \end{pmatrix}, \quad k = \overline{1, m}$$

$$\vec{X}_0 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \theta = \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_m \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_m \end{pmatrix}$$

$$X = (X_{jk}), \quad j = \overline{1, N}, \quad k = \overline{0, m}$$

(3) $\Rightarrow Y = X \cdot \theta + \varepsilon$ (4)

или

$$y = \theta_0 \cdot \vec{X}_0 + \dots + \theta_m \cdot \vec{X}_m + \varepsilon \quad (4')$$

ε - имеет многомерное нормальное распределение со средним 0 и мат. ковар. $\Sigma_\varepsilon = \sigma^2 I_N$

$\Rightarrow y$ имеет многомерное нормальное распр.

$$E(y) = X \cdot \theta, \quad \Sigma_y = \Sigma_\varepsilon = \sigma^2 \cdot I_N \quad (5)$$

Основные задачи

1. Оценка параметров
2. Проверка гипотез о параметрах (а вообще при некоем X параметр θ очень мал, или вообще почти одинаков, тогда вместо X_1, X_2 будет $(X, \pm X_2)$)
3. Выбор значимых факторов
4. Проверка адекватности модели
5. Проверка выполнения основных ограничений (важные и самая сложная задача)

§ 3 Оценка параметров. Метод наименьших квадратов.

Имеем $y_j, x_{j0}, \dots, x_{jm}$. Нужно оценить параметр θ .

Будем использовать метод наименьших квадратов (мнк)

$$Q(\theta) := \sum_{j=1}^N [y_j - (x_{j0}\theta_0 + \dots + x_{jm}\theta_m)]^2 = (y - X \cdot \theta)^T \cdot (y - X \cdot \theta) = \|y - X \cdot \theta\|_2^2 \rightarrow \min_{\theta} \quad (1)$$

Экстремальная задача

Запишем необход. условие экстремума:

$$\frac{\partial Q}{\partial \theta_k} = 0, \quad k = \overline{0, m}$$

если предприм. кв. форму, то получится линейная система

После несложных преобразований получаем систему уравнений

$$(X^T X) \cdot \theta = X^T y \quad (2)$$

кв. матрица $(n+1) \times (n+1)$

Если $X^T X$ - невыр. матрица, то получаем

$$\hat{\theta} = (X^T X)^{-1} X^T y \quad (3)$$

$\hat{\theta}$ - оценки по МНК

Замечание

матрица $X^T X$ невырождена $\Leftrightarrow X_0, \dots, X_n$ линейно независ.

Обозначим $\hat{y} = X \cdot \hat{\theta}$ (4) - вектор предсказанных значений

$e = y - \hat{y}$ (5) - вектор остатков

Замечание

Фактически мы решили задачу о нахождении линейной оценки вектора y в пр-ве, порожденном векторами $X_0, \dots, X_m$

Задача решается однозначно, когда matr. невыр. (иначе можно по-разному спроецировать на эти векторы)
 e_j - базисные векторы

В качестве оценок b^2 берем

$$S^2 = \frac{1}{N - (m+1)} \sum_{j=1}^N e_j^2 \quad (6)$$

§ 4 Свойства оценок параметров

Опр. Оценка $\hat{\theta}$ параметра θ по наблюдениям y называется линейной, если $\hat{\theta} = A \cdot y$, где A - некоторая невырожденная матрица.

Можно показать, что оценки по МНК обладают следующими свойствами. Предполагается, что выполняются условия 1-5.

- 1) $\hat{\theta}$ - линейная оценка (см. 3)
- 2) $\hat{\theta}$ - несмещенная оценка, т.е. $E_{\theta}(\hat{\theta}) = \theta$ $\forall \theta$ где θ - истинный параметр
- 3) $\hat{\theta}$ - имеет matr. ковариации $\Sigma_{\hat{\theta}} = \sigma^2 (X^T X)^{-1}$
- 4) Если диагональные элементы матрицы $(X^T X)^{-1}$ стремятся к нулю при $n \rightarrow \infty$, то $\hat{\theta} = \hat{\theta}_n$ состоятельная оценка (зависит от числа измерений) (первое условие бн)
- 5) Теорема (Гаусса - Маркова)
Оценки $\hat{\theta}$ для θ по МНК являются оптимальными в смысле квадратичности в классе всех линейных и несмещенных оценок
- 6) S^2 - несмещенная и состоятельная оценка для σ^2

§5 Построение доверительных интервалов
 поди узнать насколько точно была оценена истинная
 параметр ^{ошибка может с б.м.с. оценок} ^{ошибка может с б.м.с. оценок} ^{ошибка может с б.м.с. оценок}

Все дальнейшее рассуждение опирается на следующий результат

Пусть имеем классическую линейную модель и выполним основные ограничения 1) - 6)!

Теорема Если $\hat{\theta}$ есть оценка θ по мнк, то

1) $\hat{\theta}$ имеет многомерное нормальное распределение:
 $E(\hat{\theta}) = \theta$, $\Sigma_{\hat{\theta}} = \sigma^2 (X^T X)^{-1}$

2) сур. величина $(\hat{\theta} - \theta)^T \cdot (X^T X) (\hat{\theta} - \theta) / \sigma^2$ имеет χ^2 распределение с $(m+1)$ степенями свободы

3) с.в. $\sum_{j=1}^N (y_j - \hat{y}_j)^2 / \sigma^2 = \sum_{j=1}^N e_j^2 / \sigma^2$ - имеет χ^2 -распр с $N - (m+1)$ степеней своб.

^{размерность N}
^{спроектированы (m+1)}
^{остатки N - (m+1)}

4) с.в. S^2 и $\hat{\theta}$ (или $\hat{y}$) независимы

Фиксируем некоторое $k: 0 \leq k \leq m$

В силу теоремы θ_k имеет одномерное нормальное распределение со средним θ_k и дисперсией $\sigma^2 v_{kk}$, где v_{kk} есть диагональ. элемент матрицы $(X^T X)^{-1}$. Тогда с.в.

$$T_k := \frac{\hat{\theta}_k - \theta_k}{\sqrt{v_{kk}} \cdot S} = \frac{\hat{\theta}_k - \theta_k}{S_k} \quad (1) \text{ - имеет распр.}$$

Стюдента с $(N - (m+1))$ степенями свободы

станд. норм. распр. и интервал в две стороны ± 2

S_k - стандартная ошибка оценки $\hat{\theta}_k$ параметра θ_k

Для заданного $\delta (\approx 1)$ по таблице распр.

Стюдента находим $c(\delta) > 0: P(|T_k| \leq c(\delta)) = \delta \quad (2)$

$$\Rightarrow \hat{\theta}_k - c(\delta) \cdot S_k < \theta_k < \hat{\theta}_k + c(\delta) \cdot S_k \quad (3)$$

ошибка точности оценки

§6 Проверка значимости влияния отдельного фактора

Фиксируем некоторое $k: 0 \leq k \leq m$

Вспомогательная задача о том, оказывает ли фактор X_k значимое влияние на переменную Y .

Формально проверим

$H_0: \theta_k = 0$ не оказывает значимого влияния против (1)

$H_1: \theta_k \neq 0$ оказывает

$$\text{Вновь рассмотрим статистику } T_k := \frac{\hat{\theta}_k}{S_k} \quad (2)$$

которая при верной H_0 имеет распределение Стюдента с $(N - (m+1))$ ст. св.

проверка значимости (под влиянием фактора k отклика)

Для заданного $\alpha (\approx 0)$ по табл. распр. Стюдента находим $c(\alpha) > 0: P(|T_k| > c(\alpha) | H_0) = \alpha \quad (3)$

принимать H_0

$\Rightarrow$ Пусть реально получим некоторое T_k^* (не сур. вел., а то, что мы реально измерили)

1) Если $|T_k^*| > c(\alpha)$, то отвергаем H_0 , т.е. обнаружено значимое влияние фактора

- 2) Если $|T_k^*| \leq c(k)$, то говорят, что не обнаружено значимого влияния X_k
 $\Rightarrow$ в том случае X_k удален из уравнения и оцениваем модель заново !!!
 $\Rightarrow$ Т-критерий Стюарта удалит только по одной переменной

8) § 7 Проверка линейной гипотезы.

1° Общие линейная гипотеза.

Пусть мы имеем классическую линейную модель

$$y = X \cdot \theta + \varepsilon \quad (1)$$

и выполняем основные ограничения 1) - 6).

Рассматривается задача проверки гипотезы

$$H_0: R \cdot \theta = \gamma \quad (2)$$

против

$$H_1: R \cdot \theta \neq \gamma$$

где R - есть матрица $p \times (m+1)$ - матрица
 γ - соответствующий p -вектор

(2) - p линейных соотношений для параметров θ .

Алгоритм проверки общей линейной гипотезы:

- 1) Оцениваем модель (1) без учета ограничений (2) и находим сумму квадратов остатков: ESS_{un} (свободный регресс) error sum of squares unrestricted (unstr)
- 2) Оцениваем модель (1) по мнк с учетом ограничений (2) и находим сумму квадратов остатков ESS_R restricted

ошибка меньше, точность выше $\Rightarrow ESS_R > ESS_{un}$

- 3) с.в. ESS_{un} / σ^2 и $(ESS_R - ESS_{un}) / \sigma^2$ (здесь на σ^2 , чтобы было стандартизовано)

- независимы и имеют χ^2 -распр. с $(N - (m+1))$ и p степенями свободы.
разница чисел на две единицы, тогда по берну

- 4) с.в.

$$F = \frac{(ESS_R - ESS_{un}) / p}{ESS_{un} / (N - (m+1))} \quad (3)$$

при верной H_0 (!) имеет распределение Снедекера - Фишера с $(p, N - (m+1))$ степеней своб.
т.е. мы уже можем проверить гипотезу H_0

- 5) Для заданного уровня значимости α по таблице находим критическую константу $F(\alpha) > 0: P(F > F(\alpha) | H_0) = \alpha \quad (4)$
критическое значение для верной гипотезы H_0 ?

- 6) Если реально наблюдаемое значение F_N статистики F таково, что $F_N > F(\alpha)$, то обнаружено значимое отличие от гипотезы H_0 (т.е. её отвергают)

Если $F^H \leq F(1)$, то утверд, что не
одновременно значимого отличия от H_0 .

F-критерий Фишера

рассмотрим несколько частных случаев:

частный случай 1 2° Проверка значимости внешней
отдельного фактора (интересен ли H_0 данному фактору)

Фиксируем k -ый. $0 \leq k \leq m$
(выделим какой-то фактор)

Рассмотрим заданную значимость внешней
фактора X_k на Y

Формально проверим:

$H_0: \theta_k = 0$ (по спец. случаю (2))

против

$H_1: \theta_k \neq 0$

$p=1$. Имеем частный случай (2).

Применяем описанную выше процедуру,
 $\Rightarrow$ F-критерий

H_0 и H_1 как есть. T-критерий для
той же задачи.

T-критерий проверяет внешние X_k
в присутствии остальных факторов,

F-критерий оценивает много
внешних X_k (т.е. все остальные факторы)

хотелось бы узнать, а можно ли сразу
выкинуть неск. факторов, как проверяли,
то они совместно не вых. внешне
(вспомни пример)

3° Проверка значимости внешних групп
факторов.

Фиксируем $0 \leq i_1 < i_2 < \dots < i_r \leq m$

Рассматриваем заданную значимость совместного
многого внешних на Y групп факторов

$X_{i_1}, \dots, X_{i_r}$

Формально проверим $H_0: \theta_{i_1} = \theta_{i_2} = \dots = \theta_{i_r} = 0$
против

$H_1: \exists k: \theta_{i_k} \neq 0$

Если выяснится, что эта группа
факторов не оказывает значимого внешнего
на Y , то их все вместе можно удалить
из уравнения

4° Проверка адекватности модели. (9)

много ли каким есть модели, то принимая всего
какими заданы каждый фактор, которые
ли взгляды, значимое содержание переходим Y

Будем понимать под адекватностью модели
то, что выбранный каждый фактор
оказывает значимое внешнее на Y

Формально проверим

$H_0: \theta_1 = \dots = \theta_m = 0$
против

$H_1: \exists k \theta_k \neq 0$

Пусть y_j - реально наблюдаемое значение

$\hat{y}_j$ - предсказанные значения

Можно показать, что верно следующее тождество

$$\sum_j (y_j - \bar{y})^2 = \sum_j (\hat{y}_j - \bar{y})^2 + \sum_j (y_j - \hat{y}_j)^2$$

$$TSS = RSS + ESS \quad (5)$$

total sum of squares
полная сумма кв.

regression sum of squares
объясненная сумма кв.

error ...
остаточная сумма кв.

(в некоторых книжках исп. explanation и residual, т.е. RSS и ESS меняют места)

Опр. Величина $R^2 := \frac{RSS}{TSS} \quad (6)$

называется **коэффициент детерминации**

R^2 показывает, насколько точно модель объясняет совместный линейный эффект всех факторов

Замечание

1) Модель хороша (адекватна), если $R^2 \sim 1$
(теоретически, но на практике считается 40% хорошо)

2) Если имеем две разные модели, то модель по которой R^2 больше

проверим, что
критерий значимости θ
никогда не бывает нулевым

3) чем больше факторов (заметно больше), тем R^2 увеличивается, это плохо

Используется так называемый **исправленный (adjusted) коэффициент детерминации**
(в некоторых источниках в некотором смысле интерпретируется за увеличение измерений)
но это тоже не выход

§ 8 Сравнение двух регрессий (10)

очень важно сравнивать данные
данные одной совокупности и групп (компр. у
мужского и женского пола)

сравнивать данные в разные дни

Пусть имеем две модели измерений, которые удовлетворяют следующим уравнениям:

$$y_j = \theta_1' \cdot X_{j,1} + \dots + \theta_m' \cdot X_{j,m} + \varepsilon_j, \quad j = \overline{1, N_1} \quad (1)$$

$$y_j = \theta_1'' \cdot X_{j,1} + \dots + \theta_m'' \cdot X_{j,m} + \varepsilon_j, \quad j = \overline{N_1+1, N_1+N_2} \quad (2)$$

ε_j - н.о.р. с.в. и $D(\varepsilon_j) = \sigma^2$

Проверяется гипотеза:

$$H_0: \theta_k' = \theta_k'' \quad \forall k$$

против

$$H_1: \exists k \theta_k' \neq \theta_k''$$

(3)

Рассмотрим более общее уравнение

$$y_j = \theta_1' \tilde{X}_{j,1} + \dots + \theta_m' \tilde{X}_{j,m} + \theta_{m+1}'' \tilde{X}_{j,m+1} + \dots + \theta_{m+N_2}'' \tilde{X}_{j,m+N_2} + \varepsilon_j \quad (4)$$

$$j = \overline{1, N_1 + N_2}$$

$$\text{где } \tilde{X}_{jk} = \begin{cases} X_{jk}, & 1 \leq j \leq N_1, 1 \leq k \leq m \\ 0, & \text{в противном случае} \end{cases}$$

$$\tilde{X}_{jk} = \begin{cases} 0, & \text{в противном случае} \\ X_{j, k-m}, & N_1 + 1 \leq j \leq N_1 + N_2, m+1 \leq k \leq 2m \end{cases}$$

в рамках модели (4) гипотеза H_0 об отсутствии линейной регрессии

Если оценить (4) без учета (3), то это эквивалентно независимой оценке (1) и (2)

$$\Rightarrow ESS_{ur} = ESS_1 + ESS_2$$

Если в (4) оценить с учетом (3), то функционально оценивается (4) по $N_1 + N_2$ наблюдений $\Rightarrow ESS_R$

$$F = \frac{(ESS_R - ESS_{ur}) / m}{ESS_{ur} / (N - (m+1))}$$

- имеет при верной H_0 распределение Стьюдента - Фишера

$\Rightarrow$ Тест Чоу (Chow)

Пример

Изучается разница в оплате труда мужчин и женщин в Голландии, что влияет и образование, стаж.

Опросили 75 мужчин и 75 женщин, которые работают на полную ставку, получены информации о следующих величинах:

w - зарплата

age - возраст (дней считая, что безработ и стаж примерно один и тот же)

edu - уровень образования

Оценим уравнение регрессии w на c , age , edu только для мужчин

$$\Rightarrow w^{(1)} = -3.37 + 0.479 \cdot age + 3.943 \cdot edu$$

коэффициенты $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ в м.у. $\hat{\beta}_0$ - константа, $\hat{\beta}_1$ - эффект от возраста, $\hat{\beta}_2$ - эффект от образования.

$$ESS_1 = 5672.328$$

аналогично только для женщин:

$$w^{(2)} = -0.20 + 0.414 \cdot age + 2.305 \cdot edu$$

$$ESS_2 = 1788.344$$

Применим тест Чоу

Оценим уравн. по всем наблюдениям.

$$w = -3.06 + 0.478 \cdot age + 3.254 \cdot edu$$

$$ESS_R = 8080.113$$

Вспомогат. значение статистики $F \Rightarrow$

$$F^* = 3.342$$

$$\Rightarrow \text{так } \alpha = 0.05$$

т.е. значение статист. уровн.

сейчас изур. подход, кет. заверш. не только о существующих различиях, но и о том "какие резим"

(11)

§ 9 Фиктивные переменные

Ранее явно предполагалось, что факторы - это количественные характеристики

Есть ситуации, когда факторы имеют качественный характер

Примеры

1. Изменить наличие или отсутствие влияния образования на величину зарплат в некоторой отрасли
2. Изучается сезонное влияние на продажу товара.
3. Структурные перестройки в модели.

Вводит некоторые вспомогательные (и фик- тивные "дипту") переменные

Пусть некое кач. фактор имеет два

уровня. Обозначим один из них - 1
второй - 0

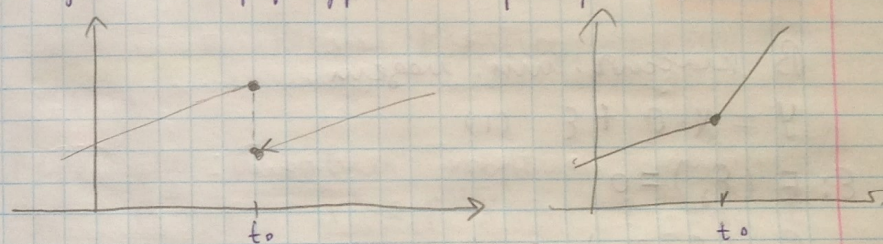
Пусть фактор = сезон

Рассмотрим какую переменную со значениями 0, 1, 2, 3

При таком подходе все переходы между соседними уровнями считаются равноправными.

В том случае более разумно ввести несколько бинарных переменных - каждая для своего перехода.

Случай структурных перестроек:



используя график переи. линии в рамках линейной регрессии такие резиме перестройки учесть

Пример Зарплата мужчин и женщин в Голландии

В дополнение к известным факторам добавим график переи.:

$$d = \begin{cases} 1, & \text{если мужчина} \\ 0, & \text{если женщина} \end{cases}$$

$$w = 6.11 + 0.53 \text{ age} - 3.73 d$$

Фактор d - значимый $\Rightarrow$ обнаружено
значимое различие в зарплате

- 3.73 - коэффициент оценки этой переменной

"-" - значит, что она меньше

симметричная корр. ρ классиф. линейной
модели $\Rightarrow$ всегда один

6

Глава 6 Обобщенный МНК

В классиф. линейной модели

$$y = x \cdot \theta + \varepsilon \quad (1)$$

$$\varepsilon: E(\varepsilon_j) = 0$$

$$\text{Cov}(\varepsilon_j, \varepsilon_k) = 0 \quad j \neq k$$

$$D(\varepsilon_j) = \sigma^2$$

$$\Rightarrow \Sigma_{\varepsilon} = \sigma^2 \cdot I_n \quad (2)$$

$$\text{Пусть } \Sigma_{\varepsilon} = \sigma^2 \cdot \Omega \quad (3)$$

где Ω - известная симметричная положительно
определенная матрица

Примеры к (1) обобщенный МНК

1) $\hat{\theta}$ - неслучайные элементы θ

2) $\hat{\theta}$ - (при некотором θ) - случайные

3) $\hat{\theta}$ - неслучайные.

4) Проблема преобразования возникает при
преобразовании модели

Из линейной алгебры известен след. результат:
симм. матрица P - невырождена:

$$\Omega = P \cdot P^T \quad (4)$$

Замечание: Матрица P опреz. не единична
обращена

Перейдем к новым переменным:

$$y^* = P^{-1} y \quad x^* = P^{-1} x \quad \varepsilon^* = P^{-1} \varepsilon$$

$$\text{Тогда вместо (1) получим } y^* = x^* \cdot \theta + \varepsilon^* \quad (5)$$

Здесь θ - такое же как в (1)!
с тем же значением параметров, то есть значения их здесь

Приведу в пример можно показать,
что (5) - классическая линейная модель

$$\text{т.е. } \Sigma_{\varepsilon^*} = \hat{\sigma}^2 I \quad (2')$$

Применим к (5) обобщенный МНК

В старых переменных это имеет вид:

$$(y^* - x^* \theta)^T \cdot (y^* - x^* \theta) = (y - x \cdot \theta) \Omega^{-1} \cdot (y - x \cdot \theta) \rightarrow \min_{\theta} \quad (6)$$

(6) - некоторая квадратичная форма по θ .
 $\Rightarrow$ Обобщенный МНК

$\Rightarrow \hat{\theta}_{GLS}$ - оценки по обобщ. МНК
 Generalized least squares

Преимущества (по сравн. с МНК)

- 1) Эти оценки оптимальны
- 2) Применяют обобщенные методы проверки гипотез (в новых переменных!)

Важный частный случай:

Пусть Ω - невырожденная диагональная матрица, у которой на диагонали элементы $\omega_1, \omega_2, \dots, \omega_n > 0$

В этом случае ω^{-1} есть диагональная матрица с элементами $\frac{1}{\omega_1}, \dots, \frac{1}{\omega_n}$

А тогда (6) можно переписать в виде

$$\sum_{j=1}^n \left[y_j - (\theta_0 + \theta_1 \cdot x_{j1} + \dots + \theta_m x_{jm}) \right]^2 \frac{1}{\omega_j}$$

$\rightarrow \min_{\theta} \quad (6')$

- обобщенный МНК

$\hat{\theta}_{GLS}$ - оценки Гаусса
 все взаимосвяз. уравнен. когда Ω известна
 В реальных задачах Ω неизвестна
 $\Rightarrow \frac{N(N+1)}{2}$ дополнительных параметров

А измерений только N !

нужно делать разные предпос. о матрице Ω

В общем случае надо учесть два обстоятельства:

- 1) Дисперсии ошибок могут быть разными
- 2) Ошибки могут быть коррелированными

Глава 7. Гетероскедастичность и автокорреляция ошибок.

(12)

В классической линейной модели

$$y = x \cdot \theta + \varepsilon \quad (1)$$

$$\Sigma \varepsilon = \sigma^2 \cdot I_N \quad (2) \quad (\text{мы предполагаем})$$

$$\Sigma \varepsilon = \sigma^2 \cdot \Omega \quad (3) \quad \text{где } \Omega - \text{известно}$$

(связи или преобр. к I_N , т.е. обобщ. метод наим. кв.)

Если $\Sigma \varepsilon$ - другой вид и неизвест., то много параметров

§1 Модель с гетероскедастичностью

1°. Описание модели.

Пусть имеем модель (1). где

1) ε_j - некоррелированы

2) $D(\varepsilon_j) = \sigma_j^2$ (разные для разных j)

- модель с гетероскедастичностью

опять параметров $(N + (N+1))$ больше, чем измерений N , значит делаем предположения о σ_j^2 :

Предположим, что выбран один из измерений X_k (она n -ная σ_j^2 в модели, но все значения реально неизвестны (например), и $\sigma_j^2 = \sigma^2 \cdot |X_{jk}|^2$ (4)

Здесь мы находимся в рамках модели (3) с весами $w_j = |X_{jk}|^2$

Если применить взвешенный МНК, то получим оптимальные оценки и разработать все процедуры для проверки гипотез.

В новых переменных $y_j^* = y_j / |X_{jk}|$
 $x_{jp}^* = x_{jp} / |X_{jk}|$

2°. Проверка на гетероскедастичность

Фактически мы хотим проверить гипотезу

$$H_0: \Sigma \varepsilon = 0^2 \cdot I_N$$

против

$$H_1: \Sigma \varepsilon - \text{виза (4)}$$

Имеется Тест Голдфелда-Кванта
(Goldfeld-Quandt)

Описание процедуры

1. Все измерения упорядочиваем по возрастанию $|X_{jk}|$

2. Из всех данных исключаем d измерений (обычно $d \sim \frac{N}{4}$)

3. Сформируем две регрессии по первым $(N-d)/2$ и последним $(N-d)/2$ измерениям и получим остатки $e^{(1)}$ и $e^{(2)}$.

4. По остаткам находим сумму квадратов остатков ESS_1 и ESS_2 и статистику $F = \frac{ESS_2}{ESS_1}$ (5)
(числ. степ. свободы сократились) $\rightarrow$

5. Для заданного α при верной H_0 статистика F имеет распределение Фишера с $((N-d)/2 - (m+1))$, $\text{числ. степ. свободы}$ и $(m+1)$, $\text{знамен. степ. свободы}$

6. Для заданного λ по табл. Найдите $c(\lambda)$
 $P(F > c(\lambda) | H_0) = \lambda$

λ - ошибка 1-го рода (очень маленькое число)

7. Если реально получ. знел. F^N стат. F таково, что $F^N > c(\lambda)$, то отвергаем H_0 , т.е. обнаружены признаки гетер.

В противном случае утверд., что не обнаруж. признаков.

(если гетероген, не типа (4), например синусоид. колеблется, то этот тест не работает)

13 §2 Модель с авторегрессионной ошибкой.

1° Описание модели

$$y = X \cdot \theta + \varepsilon(1)$$

Пусть

$$\varepsilon_j = \rho \cdot \varepsilon_{j-1} + \delta_j \quad (2) \quad j = \overline{1, N}$$

где

1) δ_j - н.о.р. (независимо одинаков. распр.) с нормальным распр.

$$E(\delta_j) = 0, D(\delta_j) = \sigma^2$$

$$\text{Cov}(\delta_j, \delta_k) = 0, k \neq j.$$

$$2) \varepsilon_0 - \text{с.в. с норм. распр.}, E(\varepsilon_0) = 0, D(\varepsilon_0) = \frac{\sigma^2}{1 - \rho^2} = \sigma_\varepsilon^2$$

Предполагаем, что $|\rho| < 1$ (иначе ошибки будут накаплив.)

$$\Rightarrow E(\varepsilon_j) = 0 \quad \forall j \quad (3)$$

$$\text{Cov}(\varepsilon_j, \varepsilon_k) = \sigma_\varepsilon^2 \rho^{|k-j|} \quad (4)$$

$$D(\varepsilon_j) = \sigma_\varepsilon^2 \quad (5) \quad \forall j$$

(ε_j) со св-вами (2)-(5) наз. модель авторегрессии первого порядка (объясняется через саму себя)

Параметры модели:

$$\theta = (\theta_0, \theta_1, \dots, \theta_m), \sigma_\varepsilon^2, \rho$$

2° Оценивание модели.

Пусть ρ - известно.

Перейдем к новым переменным $y_1^* = \sqrt{1 - \rho^2} y_1 \quad (6)$

$$y_j^* = \theta_0 + \theta_1 x_{j,1} + \dots + \theta_m x_{j,m} + \varepsilon_j \quad (1')$$

$$y_j^* = y_j - \rho y_{j-1}^* =$$

$$= \theta_0(1 - \rho) + \theta_1(x_{j,1} - \rho x_{j-1,1}) + \dots + \theta_m(x_{j,m} - \rho x_{j-1,m}) + \delta_j$$

$\Rightarrow$ (6) - класс. линейная модель (т.к. ошибки δ_j - н.о.р.) т.е. достаточно применить мнк

(6) - авторегрессионное преобразование

В реальных задачах ρ - неизвестно.
 см в (6) применяю мнк при неизв. ρ , но
 получив наименьшую
 сумму квадратов $\rightarrow$ как ρ было известно (предположительно)
 отрезок на 10 точек $\rightarrow$ увеличен $\rightarrow$ смотрим еще
 лучше всего
 далее найденный отрезок делим на 10 т. и т.д.

II процедура Кимберга - Лу (K_1 до K_{10} - L_1)

- 1) Отрезок $(-1, 1)$ разбиваем на мелкую
сетку возможных значений
- 2) Для каждого возможного ρ делаем
преобразование (6)
- 3) Проводим систему параметров
предсказания системы (при фиксированном
 ρ)
- 4) Выбираем то ρ , где сумма квадратов
остатков минимальна
- 5) В окрестности найденного ρ строим
более мелкую сетку

Процедура продолжается до тех пор, пока не
будет достигнута нужная точность.

3° Критерий Дарбина - Уотсона (Durbin-Watson)

Задача состоит в том, чтобы проверить
наличие автокорреляции. Формально
проверяем

$$H_0: \rho = 0$$

против

$$H_1: \rho > 0$$

Дарбин и Уотсон в 1951 году предложили след.
процедуру

Оценим (1) по абсолютному мнк и найдем
остатки e_j .

Рассмотрим статистику $DW := \frac{\sum_{j=2}^N (e_j - e_{j-1})^2}{\sum_{j=1}^N e_j^2} \quad (7)$

$$\approx 2(1 - \Gamma)$$

где Γ - оценка для ρ

приведенной выше к виду

H_1 меньше 2

Внимательно, что эта величина зависит от
оценки $\hat{x}$ - предельная (т.к. заранее не
не сечено)

$$DW \approx 2 \text{ для } H_0$$

$$< 2 \text{ для } H_1$$

Для построения критерия необходимо для
заданного L найти $c(L)$:

$$P(DW < c(L) | K_0) = L$$

Эта вероятность зависит от конкретных значений факторов $X_1, \dots, X_m$. (т.е. таблица заранее составлена не была)

Дарбин и Уотсон применили метод наименьших квадратов, где константы d_e и d_u зависят только от m, N, L , такие, что

$$d_e \leq c(L) \leq d_u$$

$\begin{matrix} \text{наиб.} & \text{наиб.} \\ \text{мал.} & \text{мал.} \\ \text{знач.} & \text{знач.} \end{matrix}$

Описание процедуры:

конкретное значение

1) Находим наблюдаемое значение DW^* статистики DW - с.в.

2) Если $DW^* < d_e \leq c(L)$, то принимаем H_1 , т.е. обнаружена значимая положительная автокорреляция в ошибках.

3) Если $DW^* > d_u \geq c(L)$, то гипотеза H_0 по крайней мере не отвергается, т.е. не обнаружена...

4) Если $d_e \leq DW^* \leq d_u$, то не можем принять решение.

где d_e и d_u есть таблица

Пример $N=100; m=1; L=0,05$

$$\Rightarrow d_e = 1,65; d_u = 1,69$$

Глава 7 Модель дискретного выбора.

(14)

Есть зависимая (объясняемая) переменная Y и один или несколько факторов $X_1, \dots, X_m$, с помощью которых мы пытаемся объяснить поведение Y .

Во всех моделях ранее Y принимало непрерывные значения.

напр. марка - 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

Y принимает конечное мн-во значений.

§ 1 Модель бинарного выбора.

Пусть Y принимает только два разных значения: 1 и 0.

Y - случайная величина!

Для описания распределения с.в. Y достаточно знать $P(Y=1 | X)$, где X - набор объясняющих факторов.

$(X_1, \dots, X_m)$

Итак, мы хотим найти функцию, которая принимает значения между 0 и 1, рассмотрим следующую модель

$$P(Y=1 | X) = F(X \cdot \theta)$$

$$\text{где } X \cdot \theta = X_0 \cdot \theta_0 + X_1 \cdot \theta_1 + \dots + X_m \cdot \theta_m$$

где θ - набор параметров

$F(z)$ - некоторая строго возрастающая функция,

таким, что $0 \leq F(z) \leq 1$

Обычно $F(z)$ - некоторая ор. ф. распределения

Обычно мы имеем N независимых наблюдений $y_j, x_j = (x_{j0}, \dots, x_{jm})$

$$P(y_j = 1 | x_j) = F(x_j; \theta) \quad (1)$$

т.е. какому из двух исходов

Природа состоит в том, что наблюдаются не вероятности, а значения величин y_j

Пусть есть некоторая ненаблюдаемая (латентная) переменная y_j^* , на основе которой принимается решение по следующему правилу:

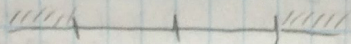
$$y_j = \begin{cases} 1, & y_j^* > 0 \\ 0, & y_j^* \leq 0 \end{cases} \quad (2)$$

$$\text{где } y_j^* = x_j \cdot \theta + \varepsilon_j \quad (3)$$

- максим. лн. модель, где $F(z) = P(\varepsilon_j < z)$ (4)

- строго монотонна и симметрична относительно нуля

$$\begin{aligned} \Rightarrow P(y_j = 1 | x_j) &\stackrel{(2)}{=} P(y_j^* > 0 | x_j) = \\ &\stackrel{(3)}{=} P(x_j \theta + \varepsilon_j > 0) = P(\varepsilon_j > -x_j \theta) = \\ &= 1 - F(-x_j \theta) = F(x_j \theta) \end{aligned} \quad \begin{matrix} \text{а это} \\ \text{именно} \\ \text{модель (1)} \end{matrix}$$



Предмет, что ~~каждое~~ из возможных решений (0 или 1) имеет свою полезность

$$u_{j1} = x_j \theta^{(1)} + \varepsilon_{j1} \quad (4)$$

$$u_{j0} = x_j \theta^{(0)} + \varepsilon_{j2}$$

Модель "полезности" принятия решений u_{j0} .

Естественно использовать следующее правило:

$$y_j = \begin{cases} 1, & u_{j1} \geq u_{j0} \\ 0, & \text{в противном случае} \end{cases}$$

$$\text{Обозначим } \theta = \theta^{(1)} - \theta^{(0)}$$

$\varepsilon_j = \varepsilon_{j1} - \varepsilon_{j0}$, $F(z)$ - ор. ф. с. в. ε_j которая строго монотонна и симметрична относительно нуля

$$\begin{aligned} \text{Легко видеть, что } P(y_j = 1 | x_j) &= P(u_{j1} \geq u_{j0} | x_j) = \\ &= F(x_j \theta) \end{aligned}$$

Пример конкретн. моделей:

$$1) \text{ Если } F(z) = \Phi(z) = \int_{-\infty}^z \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx \quad \begin{matrix} \text{норм. уд. ф.} \\ \text{очень часто} \end{matrix}$$

то получим, что имеем Probit-модель

$$2) \text{ Если } F(z) = \Lambda(z) := \frac{e^z}{1 + e^z} \quad \begin{matrix} \text{стандартное} \\ \text{логистическое} \end{matrix}$$

логистическое распределение)

$\Rightarrow$ Logit - модель

3) Если $F(z) = e^{-e^{-z}}$ (экстремальное распределение)

Основные задачи:

- 1) Оценка параметров модели.
- 2) Интерпретация результатов
как влияют факторы на принятое решение
- 3) Процедура прогноза
- 4) Сравнение реальных и предсказанных значений
предсказывает ли наша модель то, что мы хотим

§ 2 Интерпретация модели.

как факторы, входящие в эту модель, влияют на принятое решение

Основная задача состоит в том, что вероятность принятия решения конкретным образом зависит от факторов.

Для Логит-модели (т.к. она проще)
статистически реализуется, как и у логист

$$P(y_j = 1 | x_j) = \Lambda(x_j; \theta) = \frac{e^{x_j \theta}}{1 + e^{x_j \theta}} \quad (1)$$

Зависимость от факторов преобразовать с помощью т.н. предельных эффектов:
хотим поделить, как это повлияет на интерес, как величина

$$\text{Вот так } \frac{\partial}{\partial x_k} P(y_j = 1 | x_j) \stackrel{(1)}{=} \Lambda(x_j; \theta) \cdot$$

$$\cdot [1 - \Lambda(x_j; \theta)] \cdot \theta_k \quad (2)$$

В этой выработке присутствуют все факторы и все параметры.

Для интерпретации модели используем отношение марков как одно делит на другое

$$\frac{P(y_j = 1 | x_j)}{P(y_j = 0 | x_j)} = \frac{\Lambda(x_j; \theta)}{1 - \Lambda(x_j; \theta)} = e^{x_j \theta} \quad (3)$$

$$\text{Число используем величину } \ln \left[\frac{P(y_j = 1 | x_j)}{P(y_j = 0 | x_j)} \right] = x_j \theta \quad (4)$$

Для сравнения относительного влияния факторов используем следующую величину

$$\frac{\partial P(y_j = 1 | x_j) / \partial x_{jk}}{\partial P(y_j = 1 | x_j) / \partial x_{jl}} \stackrel{(2)}{=} \frac{\theta_k}{\theta_l} \quad (5)$$

§ 3 Оценивание модели.

ранее не только мы, сейчас это не хочется использовать

$y_j, x_j = (x_{j0}, x_{j1}, \dots, x_{jn})$ - набор измерений, которые нужны для разных j

$$P(y_j = 1) = F(x_j; \theta) \quad (1)$$

Необходимо оценить параметр θ .

Мы наблюдаем величину $y_j = 0 \vee 1$.

Для оценки θ применим метод макс. правдоподобия (МП).

Функция правдоподобия в нашей задаче имеет вид $L(\theta) = L(\theta, y, x)$ (2)

$$= \prod_{i=1}^N [F(x_i; \theta)]^{y_i} [1 - F(x_i; \theta)]^{1-y_i}$$

Перейдем к логарифмической форме правдоподобия $\ell(\theta) = \ln L(\theta) = \sum_{i=1}^N [y_i \ln F(x_i; \theta) + (1 - y_i) \ln [1 - F(x_i; \theta)]]$ (3)

$$\ell(\theta) \rightarrow \sup_{\theta} \ell(\theta)$$

$\Rightarrow \hat{\theta}$ - оценка по ММП

Если $F(z) = \Lambda(z)$ - логист. ф.р., то $\ell(\theta)$ модально выпукла

$\Rightarrow$ например, метод Ньютона - Рундса сходится к локальному максимуму при малых начальных условиях

§ 4 Вандер модели.

Пусть есть две модели: некоторая исходная (нулевая) и более сложная (альтернативная) с некоторыми дополнительными параметрами

q -мь деп. парам.

1) Оцениваем обе модели по ММП.

2) Для сравнения моделей используется тест отношения правдоподобия. При верной нулевой гипотезе статистика критерия имеет асимптотически $(N \rightarrow \infty)$ χ^2 -распр.

q степен. своб.

Мак-ражден предложим следующий анализ коэфф. детерминации.

$$R_M^2 = 1 - \frac{\ell(\hat{\theta}_M)}{\ell(\hat{\theta}_0)} \quad (5) \quad R_M^2 \text{ никогда не равна 1}$$

то есть, в ней всегда максимизируется

Замечание В классической линейной модели

$$R_M^2 = R^2$$

т.е. одно и то же для всех структур

Замечание Если модель не является временными, то сравн. моделей производится с помощью т.н. информативных критериев

Крит. Акаике AIC

Крит. Шварца BIC

чем меньше знач. критерия, тем лучше модель.

§ 5 Прогноз.

Всего N измерений. N_1 - из них N_1 для оценки модели. N_2 - для оценки качества прогноза. $N_1 + N_2 = N$.

Пусть мы оценили θ и получили $\hat{\theta}$.

Тогда прогноз для y_i по некоторым x_i данным строим по правилу

$$\hat{y}_i = \begin{cases} 1, & \text{если } F(x_i; \hat{\theta}) > c \text{ (обычно } \frac{1}{2}) \\ 0, & \text{в пр. и.} \end{cases} \quad (6)$$

Матрица Оценки $c = \frac{1}{2}$

$\hat{y}_j$ - прогноз по модели

y_j - факт. зн.

Как оценить прогнозные св-ва модели?

Составим таблицу предсказаний и реализаций

Предск. факт.	$\hat{y}_j = 1$	$\hat{y}_j = 0$	Всего N_j факт.
$y_j = 1$	P_{11}	P_{10}	$P_{1.}$ - прогн. св-ва
$y_j = 0$	P_{01}	P_{00}	$P_{0.}$
	$P_{.1}$	$P_{.0}$	

P_{11} - вер. то, что при модели пред. правильный прогноз, когда $\lambda = 1$

Модель хороша, если величина индекса качества правильных прогнозов $= P_{11} + P_{00}$

Максимум предост. группировки

$$F_1 = \frac{P_{11} + P_{00} - P_{1.}^2 - P_{.0}^2}{1 - P_{1.}^2 - P_{.0}^2} \quad (7)$$

§ 6 Логистическая регрессия в задаче скоринга.

Score

Скоринг - оценка кредитоспособности некоторого физического лица в банке.

Мы имеем некоторую совокупность данных по различным критериям.

$y_j = 1$ - вернул кредит

$y_j = 0$ - не вернул

Для решения задачи используется некоторый набор объясняющих факторов.
 $X = (\dots)$

1) Возраст

2) Пол

3) Семейное положение

4) Число иждивенцев

5) Совокупный располагаемый доход

6) Тип работы

7) Срок проживания в регионе

8) Финансовая стабильность недвижимости в собственности

9) Ежемесячные платежи по кредиту

Входные данные не являются независимыми (неодинаковое распределение)

Таким образом

мы имеем лог. регр. в задаче скоринга

516 стр - 14 табл. данных

нужно их классифицировать (если да, то 12 мес)

потенциально $\rightarrow$ pdf вопросы

предположительно, некоррелируемые и все сгруппировано

данные вопросы по классификации (каждый, ...)

данные вопросы по классификации

данные вопросы по классификации

данные вопросы по классификации

данные вопросы по классификации

данные вопросы по классификации

данные вопросы по классификации

данные вопросы по классификации

17) Тема 8 Дискриминантный анализ

есть некое число различий
предела. что-то считать...

фактически - здесь классификации
современная, они такие
недоприменяются правильно

§1 Задача классификации

Постановка задачи

Пусть имеется некоторый случайный
вектор $X = (X_1, \dots, X_p)$ с некоторым
неизвестным распределением F

Нам представлено несколько классов
 $W_1, W_2, \dots, W_k$ распределений

Основная задача необходимо на основе
некоторого правила отнести к какому
из этих классов принадлежат F

Пример 1. Необходимо отнести абитуриента
к одному из двух классов: W_1 - успешный
зачетник, W_2 - будущий отчисленный

2. Врач на основе некоего показателя
должен определить какой из k
заболеваний есть у пациента.

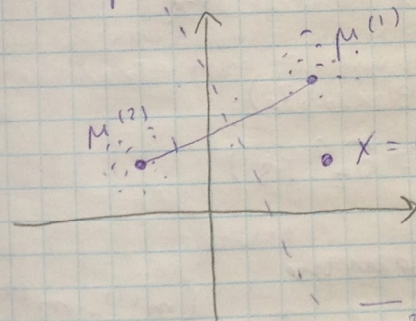
§2 Решение задачи классификации
для нормальных распределений с
известными параметрами.

Всегда далее $k=2$

W_1 - класс p -мерных нормальных распределений
с параметрами $\mu^{(1)} = (\mu_1^{(1)}, \dots, \mu_p^{(1)})^T$ - вектор
средних

$\Sigma = (\sigma_{ij})$ - матрица ковариаций

пусть уже совокупности отличаются только
векторами средних (а не стандартными отклонениями),
тогда проще было бы делать преобразование



линии, которая разделяет точки
перед и после (перед по средним
средних точек)
разрешает отнести точку X ? (или нет...)
как далеко находится среднее?
куда отнести точку X ? (или нет...)

Предлагается принимать решение с
помощью некоторой линейной функции
от наблюдений:

$$z = \alpha_1 \cdot x_1 + \dots + \alpha_p \cdot x_p \quad (1)$$

Принимается решение по следующему правилу:

1) Если $z \geq c$

то $F \in W_1$

2) Если $z < c$

то $F \in W_2$

Если принять решение сгруппировали, теперь
Основная задача:

Найти коэффициенты $\lambda_1, \dots, \lambda_r$, которые приведут к оптимальному решению.

Если $F \in W_1$, то с.в. $Z_1 = \sum_{i=1}^p \lambda_i X_i$

имеет мин. распр.

$$E(Z_1) = \sum_{i=1}^p \lambda_i M_i^{(1)} = m^{(1)} \quad (2)$$

$$D(Z_1) = \sum_{i,j=1}^p \lambda_i \lambda_j \sigma_{ij} = \sigma_z^2 \quad (3)$$

для класса W_2 аналогично

$$\text{Определим величину } \Delta^2 := \frac{(m^{(1)} - m^{(2)})^2}{\sigma_z^2} \quad (4)$$

если ошибки малы, то перенос миним. наим. расстояние
если большие, то и расстояние г.д. большие

Будем искать такие $\lambda_1, \dots, \lambda_r$ для которых получим минимальное Δ^2 .

Δ^2 - расстояние Махалабиса

Вонг Фиттер доказал, что решение этой задачи удовлетворяет линейной системе линейных уравнений

$$\lambda_1 \sigma_{11} + \lambda_2 \sigma_{12} + \dots + \lambda_r \sigma_{1r} = m^{(1)} - m^{(2)}$$

$$\lambda_1 \sigma_{r1} + \lambda_2 \sigma_{r2} + \dots + \lambda_r \sigma_{rr} = m^{(1)} - m^{(2)} \quad (5)$$

z - дискриминантная функция
 $c = ?$

В задаче классификации ошибок I и II -го рода означают и распр. вероятностей допустить эту ошибку.

$P(1|2)$ - принять W_1 , когда верна W_2

$P(2|1)$ - принять W_2 , когда верна W_1

$$\text{Будем искать } c \text{ из условия, что} \\ P(1|2) + P(2|1) \quad (6)$$

т.е. сум. вер-го ошибочной классиф. для W наименьшей

(Но сами же мы не будем восстанавливать задан. напр. 2 раз не восстанавливаем, то в 20% случаев, в 80% случаев) верно.

Предположим, что мы имеем данные
постоянному заданы:

q_1 - априорная вер-ть того, что мы в
реальности имеем W_1 ,

q_2 - априор. вер. того, что мы в реальности
имеем W_2

$\Rightarrow$ По теореме Байеса имеем:

$$P(W_i | X) = \frac{q_i P(X | W_i)}{q_1 P(X | W_1) + q_2 P(X | W_2)} \quad (7)$$

а posteriorная ^{вер-ть} реш-ия принять W_1 или W_2
при получении измерения X

Естественно принимать то реш-ие,
где апостер. вер-ть больше.

$\Rightarrow$ Если $f_1(x)$ - плотность в W_1
 $f_2(x)$ - плотность в W_2

то

$$\frac{q_1 f_1(x)}{q_2 f_2(x)} \geq 1 \Rightarrow W_1 \quad (8)$$

$$< 1 \Rightarrow W_2$$

Можно показать, что это решение
минимизирует ожидаемую вероятность
ошибочной классификации:

$$q_1 P(2|1) + q_2 P(1|2) \quad (9)$$

В терминах дискриминантной функции
процедура (8) эквивалентна

$$P \sum_{i=1}^n d_i x_i \geq \frac{m^{(1)} + m^{(2)}}{2} + \ln \left(\frac{q_1}{q_2} \right) \quad (10)$$

$< \dots$

где d_i решение системы (5)

Если $q_1 = q_2$, то точка раздела - среднее
априор. средних соответствующих
дискримин. функций
(если Б и Р известны)

В реальной задаче Б и Р известны.

Чем отличается от кластерного анализа?
(Здесь ранее сказано, что эти
классы, а там нужно самому выбирать
классы и отнести к ним)

На самом деле по дискр. функциям, как
мы найдем в реальности данные заданы,
то не изв-е. наиболее распространенное

Задана ^{первое} система признаков

(18)

§3 Задача дискриминации для двух совокупностей с неизвестными параметрами

Вектор наблюдений имеет вид
 $X = (X^{(1)}, \dots, X^{(p)})^T$

Имеем две совокупности W_1 и W_2 с нормальными распределениями $N(\mu_1, \Sigma)$ и $N(\mu_2, \Sigma)$

Параметры μ_1, μ_2 и Σ — неизвестны!
 оценивать пар-ры по выборкам, когда я знаю из какой совокупности принадлежит

Пусть имеем две отдельные выборки

$$X_{11}, \dots, X_{1n_1} \sim N(\mu_1, \Sigma)$$

$$X_{21}, \dots, X_{2n_2} \sim N(\mu_2, \Sigma)$$

набор случайных векторов

$$X_{11} = (X_{11}^{(1)}, \dots, X_{11}^{(p)})^T$$

По этим отдельным выборкам оцениваем параметры:

$$\hat{\mu}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} X_{1j} = \bar{X}_1$$

$$\hat{\mu}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} X_{2j} = \bar{X}_2 \quad (1)$$

можно верить по комб. выбор. оценим $\bar{X}, \dots$
 на 1 меньше чем (т.е. $n-1$)

$$S_1 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (X_{1j} - \bar{X}_1) (X_{1j} - \bar{X}_1)^T = \hat{\Sigma}$$

столбец строка

оцениваем одну и ту же Σ , поэтому

$$S_2 = \dots$$

оценим по всем наблюдениям

$$\Rightarrow S = \frac{1}{n_1 + n_2 - 2} [(n_1 - 1) S_1 + (n_2 - 1) S_2] \quad (2)$$

$$S = (S_{ik})$$

используя $\Rightarrow$ формируем процедуру с известными параметрами формально применим

Пусть заданы априорные вероятности q_1 и q_2 для появления W_1 и W_2

Алгоритм решения задачи:

Пусть имеем очередное измерение

$$X = (X^{(1)}, \dots, X^{(p)})^T$$

1) Решаем систему уравнений

$$\begin{aligned} \alpha_1 S_{11} + \alpha_2 S_{12} + \dots + \alpha_p S_{1p} &= \hat{\mu}_{11} - \hat{\mu}_{21} \\ \dots &\dots \\ \alpha_1 S_{p1} + \dots &+ \alpha_p S_{pp} = \hat{\mu}_{1p} - \hat{\mu}_{2p} \end{aligned} \quad (3)$$

$$\Rightarrow \hat{L}_1, \dots, \hat{L}_p$$

2) Строим оценку дискриминантной функции

$$\hat{Z} = \hat{L}_1 X^{(1)} + \dots + \hat{L}_p X^{(p)} \quad (4)$$

3) По одытоким вндеркам построим

$$\hat{z}_1 = \frac{1}{n_1} \sum_{j=1}^n \hat{z}_{1j} \quad (5)$$

$$\hat{z}_2 = \frac{1}{n_2} \sum_{j=1}^n \hat{z}_{2j}$$

где $\hat{z}_{1j}$ одытоким вндеркам по формуле (4)

нормируем $\hat{z}_1, \dots, \hat{z}_2$ по 1-ой и 2-ой одытоким вндеркам

4) Оцениваем S_z^2 по формуле

$$S_z^2 = \sum_{i=1}^p \sum_{k=1}^p \hat{L}_i S_{ik} \hat{L}_k \quad (6)$$

5) Относим $X = (X^{(1)}, \dots, X^{(p)})$ к $\hat{z}_1 + \hat{z}_2$ совокупности W_1 , если $\hat{z} = \sum_{i=1}^p \hat{L}_i X_i \geq \frac{\hat{z}_1 + \hat{z}_2}{2}$

$$+ \ln \frac{q_2}{q_1}$$

и к совокупности W_2 в противном случае (7)

Die ersten Δ^2 unabhängigen Variablen

$$D^2 = \frac{(\hat{z}_1 - \hat{z}_2)^2}{S_z^2} \quad (8)$$