Вопросы к экзамену

по курсу «Прикладной многомерный статистический анализ»

(Билеты)

1. Основные задачи многомерного статистического анализа.
2. Гильбертово пространство случайных величин. Задача о наилучшей линейной оценке.
3. Корреляционный и регрессионный анализ.
4. Коэффициенты корреляции.
5. Простая линейная регрессия. Метод наименьших квадратов. Свойства оценок.
6. Множественная линейная регрессия. МНК. Свойства оценок.
7. Т-критерий значимости влияния фактора.
8. Проверка линейных гипотез. F-критерий.
9. Проверка адекватности модели. Коэффициент детерминации.
10. Равенство уравнений регрессии. Тест Чоу.
11. Фиктивные переменные.
12. Модель линейной регрессии с гетероскедастичностью.
13. Модель линейной регрессии с автокорреляцией в ошибках. Критерий Дарбина-Уотсона.
14. Бинарная модель дискретного выбора.
15. Многомерная модель дискретного выбора в случае упорядоченных альтернатив.
16. Многомерная модель дискретного выбора в случае упорядоченных альтернатив.
17. Однофакторный дисперсионный анализ.
18. Двухфакторный дисперсионный анализ.
19. Дискриминантный анализ: постановка задачи и ее решение в случае известных параметров.
20. Решение задачи дискриминантного анализа в случае неизвестных параметров.

21. Кластерный анализ: постановка задачи и основные понятия.
22. Кластерный анализ: схема последовательного построения факторов.

Дополнительные вопросы к экзамену

по курсу «Прикладной многомерный статистический анализ»

(Знать наизусть!)

1. Основные задачи многомерного статистического анализа:
   - корреляционный анализ
   - регрессионный анализ
   - снижение размерности
   - дисперсионный анализ
   - дискриминантный анализ
   - кластерный анализ
2. Гильбертово пространство случайных величин.
3. Что такое наилучшая линейная оценка.
4. Лемма о перпендикуляре.
5. Простой коэффициент корреляции и что он измеряет.
6. Множественный коэффициент корреляции и что он измеряет.
7. Частный коэффициент корреляции и что он измеряет.
8. Множественная линейная регрессия: модель и основные ограничения.
9. Описание МНК для оценки параметров.
10. Явный вид оценок параметров по МНК.
11. Общая схема проверки гипотезы о параметре.
12. Для чего используется Т-критерий.
13. Основное различие Т-критерия и F-критерия в задаче проверки значимости влияния фактора.
14. Адекватность модели. Постановка задачи.
15. Коэффициент детерминации и что он измеряет.
16. Задача о равенстве двух регрессий.
17. Что такое модель с гетероскедастичностью в ошибках?
18. Что проверяет тест Дарбина-Уотсона?
19. Описание модели бинарного выбора?
20. Что такое Пробит-модель?
21. Что такое Логит-модель?
22. В чем различие моделей упорядоченного и неупорядоченного выбора?
23. Основная задача в однофакторном дисперсионном анализе.
24. Основная задача в двухфакторном дисперсионном анализе.

25. Основная задача дискриминантного анализа.
26. Кластерный анализ: постановка задачи.
27. Кластерный анализ: последовательное построение факторов.

Прикладной многомерный статистический анализ (ПМСА)

04.09

Хохлов Юрий Степанович yskhokhlov@yandex.ru

Лекции: "Многомерный статистич. анализ"

Экз. теоретический (без задач)

Введение.

$\vec{X} = (X_1, \dots, X_p)^T$ – случайный вектор с некоторым распределением в p различных характеристиках.

Выборка: $\vec{X}_1, \dots, \vec{X}_N$.

$\vec{X}_j = (X_{1j}, \dots, X_{pj})$

$X = (X_{kj})$ – матрица наблюдений.

Матрица централизованных наблюдений:

$$\overline{X}_k = \frac{1}{N} \sum_{j=1}^{N} X_{kj}$$

$$\vec{x}_j = (x_{1j}, \dots, x_{pj})^T = \left(X_{1j} - \overline{X}_1, \dots, X_{pj} - \overline{X}_p\right)$$

Глава 1. Основные задачи многомерного стат. анализа

В курсе предполагается, что величина – многомерная нормальная.

1) $\exists$ ли и насколько сильна связь между различными характеристиками $\Rightarrow$ корреляционный анализ
2) Найти конкретную функциональную связь одних характеристик от других.
   $\Rightarrow$ функции регрессии,
   регрессионный анализ
3) Задачи снижения размерности
   $\Rightarrow$ анализ главных компонент
   $\Rightarrow$ факторный анализ
4) Есть ли влияние в среднем на нек. группах заданного набора характеристик?
   $\Rightarrow$ дисперсионный анализ
5) Поиск правила для отнесения наблюдений к одному из заданного набора классов [классификация, кластеризация]
   $\Rightarrow$ дискриминантный анализ
6) Если неизв-но кол-во классов $\Rightarrow$ кластерный анализ

Глава 2. Гильбертово пространство случайных величин.

Нужно решить задачу о наилучшем линейном приближении.

§1. Гильбертово пр-во случайных величин

$\exists (\Omega, \mathcal{A}, \mathbb{P})$ – фиксированное вероятностное пр-во.

Обозначим через L_2 пр-во с.в. ξ: $\mathbb{E}(|\xi|^2) < \infty$ (1)

Проверим, что L_2 – линейное пр-во. (конечно)

Для $\forall \xi_1$ и $\xi_2 \in L_2$ определим число $(\xi_1, \xi_2) := \mathbb{E}(\xi_1 \cdot \xi_2)$ (2)

Оно $\exists$ и конечно. Теперь можно проверить, что выполнены следующие св-ва:

1) $(\xi, \xi) \geq 0$ и $(\xi, \xi) = 0 \Leftrightarrow \xi = 0$ п.н.

2) $(\xi_1, \xi_2) \in L_2$: $(\xi_1, \xi_2) = (\xi_2, \xi_1)$

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

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

$\Rightarrow (\xi_1, \xi_2)$ – скалярное произведение в L_2.

Можно определить норму $\forall \xi \in L_2$:

$$\|\xi\| = \sqrt{(\xi, \xi)} = \left(\mathbb{E}(|\xi|^2)\right)^{1/2} \quad (3)$$

Норма – длина элемента в линейном пр-ве

Норма → длина → сх-ть

Опр. Последовательность с.в. $\{\xi_n\}$ **сходится в среднеквадр.** к с.в. $\xi_0 \Leftrightarrow \|\xi_n - \xi_0\|^2 = \mathbb{E}(|\xi_n - \xi_0|^2) \to 0$ при $n \to \infty$ (4)

L_2 – полно относительно сходимости в среднеквадр.

$\Rightarrow L_2$ – гильбертово пр-во.

$\forall \xi_1$ и $\xi_2 \in L_2$ определим $\cos\alpha$ – угла между ними

сх. проуф.

$$\cos(\alpha) = \frac{(\xi_1, \xi_2)}{\|\xi_1\| \cdot \|\xi_2\|} \quad (5)$$

сх. проуф.

Опр. ξ_1 и $\xi_2 \in L_2$ **ортогональны** $\Leftrightarrow \xi_1$ и ξ_2: $(\xi_1, \xi_2) = \mathbb{E}(\xi_1, \xi_2) = 0$.

Замечание! $\exists \xi_1$ и $\xi_2 \in L_2$. $\exists \mathbb{E}(\xi_1) = \mathbb{E}(\xi_2) = 0$.

1) $(\xi_1, \xi_2) = \mathbb{E}(\xi_1 \cdot \xi_2) = \mathrm{cov}(\xi_1, \xi_2) = 0 \Leftrightarrow$ когда ξ_1 и ξ_2 не коррелированы

с.в. ξ_1 и ξ_2 ортогональны – т.с. они некоррелированы

2) $\cos\alpha = \rho(\xi_1, \xi_2)$ – коэфф. корреляции

§2. Задача наилучшей линейной оценки для с.в.

$\exists\, \mathcal{L} \subset L_2$ – замкнутое линейное подпространство

$\eta \in L_2$ (не обязат. $\eta \in \mathcal{L}$)

Опр. с.в. $\hat{\eta}$ наз-ся *наилучшим линейным приближением* для η в пр-ве $\mathcal{L}$, если:

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

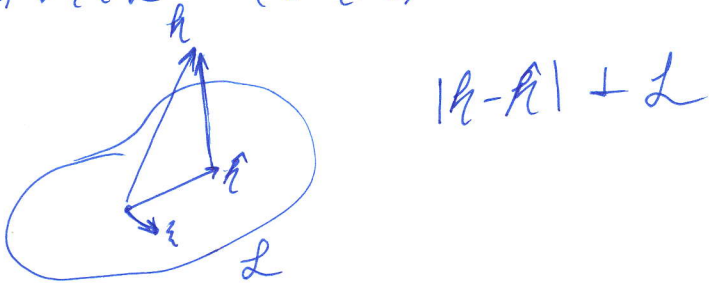
2) $\forall\, \xi \in \mathcal{L} \quad \|\eta - \hat{\eta}\| \le \|\eta - \xi\|$ (7)

Лемма о перпендикуляре

$\hat{\eta}$ есть наилучшее линейное приближение η в $\mathcal{L}$ $\Leftrightarrow$

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

2) $\forall\, \xi \in \mathcal{L} \quad (\eta - \hat{\eta}, \xi) = 0$ (8)



$|\eta - \hat{\eta}| \perp \mathcal{L}$

$\exists\, \mathcal{L} = \xi_0 + \mathcal{L}_0$, где $\mathcal{L}_0$ – лин. подпр-во.

$\mathcal{L}$ – гиперплоскость

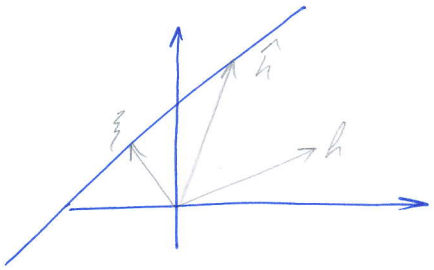
Определение оптимального эл-та дословно то же.

Лемма о перпендикуляре

$\hat{\eta}$ есть наилучшее лин. приближение $\hat{\eta}$ в $\mathcal{L}$ (гиперплоскости) $\Leftrightarrow$

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

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



11.09.

] Заданы 2 с.в. ξ и h. Необходимо найти наилучшее приближение h с помощью линейной ф-ции от ξ.

В качестве лин. пр-ва $\mathcal{L}$ рассмотрим с.в. следующего вида:

$$a+b\cdot\xi = a\cdot 1 + b\cdot\xi = a\cdot\xi_1 + b\cdot\xi_2, \quad \text{где} \quad \xi_1 = 1, \quad \xi_2 = \xi.$$

$\hat{h}$ - наилучшее лин. приближение h в $\mathcal{L}$, т.е. $\hat{h} = \hat{a} + \hat{b}\cdot\xi$

Применим лемму о перпендикуляре:

$$(h-\hat{h}, \xi_1) = \mathbb{E}[(h-\hat{h})\cdot\xi_1] = \mathbb{E}[h-\hat{a}-\hat{b}\cdot\xi] = \mathbb{E}(h) - \hat{a} - \hat{b}\cdot\mathbb{E}(\xi) = 0 \qquad (9)$$

$$(h-\hat{h}, \xi_2) = \mathbb{E}[(h-\hat{h})\cdot\xi_2] = \mathbb{E}[(h-\hat{a}-\hat{b}\cdot\xi)\cdot\xi] = \mathbb{E}[h\cdot\xi] - \hat{a}\cdot\mathbb{E}(\xi) - \hat{b}\,\mathbb{E}(\xi) = 0 \qquad (10)$$

$$\Rightarrow \hat{b} = \frac{\operatorname{cov}(\xi, h)}{\mathbb{D}(\xi)} = \rho(\xi, h)\cdot\frac{\sigma_h}{\sigma_\xi}$$

$$\hat{a} = \mathbb{E}(h) - \rho(\xi, h)\frac{\sigma_h}{\sigma_\xi}\mathbb{E}(\xi)$$

$$\Rightarrow \hat{h} - \mathbb{E}\xi = \rho(\xi, h)\cdot\frac{\sigma_h}{\sigma_\xi}[\xi - \mathbb{E}\xi].$$

Глава 3. Многомерное нормальное распределение.

Случайный вектор $\xi = (\xi_1, \dots, \xi_n)^T$ имеет многомерное нормальное распределение в $\mathbb{R}^n$, если он имеет плотность следующего вида:

$$p_\xi(x_1, \dots, x_n) = p_\xi(x) = (2\pi)^{-\frac{n}{2}} \det(A) \cdot \exp\left\{-\frac{1}{2}(A(x-m), (x-m)\right\} =$$

$$= (2\pi)^{-\frac{n}{2}} \det(A) \exp\left\{-\frac{1}{2} \sum_{i,j=1}^{n} a_{ij}(x_i - m_i)(x_j - m_j)\right\}$$

$m = (m_1, \dots, m_n)^T \in \mathbb{R}^n$

$A = (a_{ij})$ – симметрическая положительно определённая матрица

$\Sigma = A^{-1} = (\sigma_{pq})$

Обозначение: m и Σ – параметры многомерного норм. распред.

Если $t = (t_1, \dots, t_n) \in \mathbb{R}^n$, то

$$\varphi_\xi(t) := \mathbb{E}\left[e^{i(t,\xi)}\right] = \mathbb{E}\left[e^{i(t_1 \cdot \xi_1 - \dots + t_n \xi_n)}\right] \quad (1)$$

$$(1) \Rightarrow \varphi_\xi = \exp\left[i(t, m) - \frac{1}{2}(\Sigma \cdot t, t)\right] \quad (2)$$

$\Rightarrow \mathbb{E}(\xi_k) = m_k$

$\operatorname{cov}(\xi_p, \xi_q) = \sigma_{pq}$.

Если Σ – не вырождена, то многомерное норм. распред. также называют невырожденным.

Если распред. вырожд., то $\exists$ подпространство меньшей размерности, в которой М.Н.Р. не вырождено.

Если $\Sigma = I$ (единичная матрица), а $m = \vec{0}$, то распределение наз-ся стандартным.

Теорема 3.1.

Если $C: \mathbb{R}^n \to \mathbb{R}^m$ – линейное (на всё пр-во), то, если $\xi \sim N(m, \Sigma)$, то $C \cdot \xi \sim N(Cm, C \cdot \Sigma \cdot C^T)$

Следствие. ξ имеет станд. многомерное норм. распред. в $\mathbb{R}^n$, C – ортогональная проекция ξ на $\mathbb{R}^m$, $m < n$ $\Rightarrow$ вектор $h = C \cdot \xi$ имеет станд. норм. распред. в $\mathbb{R}^n$.

Теорема 3.2

$\xi = (\xi_1, \dots, \xi_n)$ имеет норм. распред. с параметрами m, Σ: $\xi \sim N(m, \Sigma)$. Разобьём ξ на 2 подвектора ξ' и ξ'' размерностями n_1 и n_2, $n_1 + n_2 = n$.

Тогда $m^T = (m', m'')^T$ и $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$

1) $\xi' \sim N(m', \Sigma_{11})$

2) ξ' и ξ'' независимы $\Leftrightarrow$ $\Sigma_{12} = \Sigma_{21}^T$ состоят из нулей.

Теорема 3.3.

Если $\xi \sim N(m, \Sigma)$, то условное распред. ξ'' при условии, что $\xi' = X'$ яв-ся многомерным норм. с вектором средних $m'' + \Sigma_{21} \cdot \Sigma_{11}^{-1}(X' - m')$ и матрицей ковариации

$$\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21} \cdot \Sigma_{11}^{-1} \cdot \Sigma_{12}$$

Теорема 3.4.

$\xi \sim N(m, \Sigma)$ – невырожд. Тогда с.в. $(\xi - m)^T \cdot \Sigma^{-1} (\xi - m)$ имеет χ^2-распределение с n степенями свободы.

Глава 4. Понятие о корреляционном и регрессионном анализе.

§1. Основные задачи корреляционного и регрессионного анализа.

Задачи 1-го типа: есть ли связь между с.в. и насколько она сильна. Обычно применяют коэфф. корреляции.

⇒ Корреляционный анализ.

Пусть мы знаем, что связь есть и она достаточно сильна.

Из курса ТВ известно, что наилучшее линейное приближение в средней квадратической с.в. Y с помощью с.в. $X_1, \dots, X_n$ задаётся с помощью условного мат. ожидания = ф-ция регрессии.

$$\hat{Y} = g(X_1, \dots, X_n) + \varepsilon$$

$$g(X_1, \dots, X_n) = \mathbb{E}[Y \mid X_1 = x_1, \dots, X_n = x_n]$$

⇒ Регрессионный анализ

1) Выбор класса моделей (будем изучать только линейные)
2) Получение оценок параметров
3) Проверка гипотез о параметрах модели
4) Проверка адекватности моделей
5) Проверка выполнения основных ограничений

§2. Коэффициент корреляции.

1°. Простой коэфф. корреляции

Опр. Простым коэффициентом корреляции простых невырожденных ($\mathbb{D} \neq 0$) с.в. ξ_1 и ξ_2 называют ρ – коэфф. корр. от ξ_1 и ξ_2:

$$\rho = \rho(\xi_1, \xi_2) = \frac{\operatorname{cov}(\xi_1, \xi_2)}{\sqrt{\mathbb{D}(\xi_1) \cdot \mathbb{D}(\xi_2)}}$$

Теорема 4.1.

1) $|\rho(\xi_1, \xi_2)| \leq 1$

2) $\rho(\xi_1, \xi_2) = 0 \iff \xi_1$ и ξ_2 – некоррелированы, т.е. $\operatorname{cov}(\xi_1, \xi_2) = 0$.

3) Если $h_1 = a_1 \xi_1 + b_1 \quad a_1 \cdot a_2 \neq 0$

$h_2 = a_2 \cdot \xi_2 + b_2$

то $\rho(h_1, h_2) = \rho(\xi_1, \xi_2) \cdot \operatorname{sign}(a_1, a_2)$ – с точностью до знака

4) Если $\rho(\xi_1, \xi_2) = \pm 1$, то $\exists\, a > 0$ и $b \in \mathbb{R}^1$: $\xi_2 = a \cdot \xi_1 + b$

Если $\rho(\xi_1, \xi_2) = -1$, то $\exists\, a < 0$ и $b \in \mathbb{R}^1$: $\xi_2 = a \cdot \xi_1 + b$

тесная функц. связь, линейная

Если $\rho > 0$, то ξ_1 и ξ_2 положит. коррелированы
(если одна растёт, то другая растёт в среднем)
Если $\rho < 0$, то ξ_1 и ξ_2 отрицательно коррелированы

$0 < \rho^2 \leq 0{,}25$ — слабо корр.
$0{,}25 < \rho^2 \leq 0{,}5$ — средне корр.
$\rho^2 > 0{,}5$ — сильно корр.

Пусть нам даны с.в. Y и X. Рассмотрим наилучшее линейное приближение $\hat{Y}$ через X.

$$\mathbb{D}(\hat{Y}) = \rho^2(X,Y) \cdot \mathbb{D}(Y).$$

Если $e = Y - \hat{Y}$ — ошибка приближения, то

$$\mathbb{D}(e) = 1 - \rho^2(X,Y) \cdot \mathbb{D}(Y)$$

$\rho^2(X,Y)$ показывает, какую долю изменчивости с.в. Y можно объяснить линейным влиянием величины X.

Задача. Есть ли влияние X на Y? 18.09.

Формально проверяется гипотеза

$H_0 : \rho = 0$ (3)

$H_1 : \rho \neq 0$

Предположим, что пара (X,Y) имеет двумерное нормальное распределение.
Пусть имеем повторную двумерную выборку $(X_1, Y_1), \dots (X_n, Y_n)$
из двумерного норм. распределения и проверяется гипотеза (3).
Рассмотрим выборочный коэфф. корреляции R:

$$R = \frac{S_{XY}}{S_X \cdot S_Y} \quad (4)$$

$$S_{XY} = \frac{1}{N} \sum_j (X_j - \overline{X})(Y_j - \overline{Y})$$

$$S_X = \frac{1}{N} \sum_j (X_j - X)^2, \quad \overline{X} = \frac{1}{N} \Sigma X_j$$

$$S_Y = \frac{1}{N} \sum_j (Y_j - \overline{Y})^2, \quad \overline{Y} = \frac{1}{N} \Sigma Y_j$$

Рассмотрим с.в. $T_{N-2} = \dfrac{R}{\sqrt{1-R^2}} \sqrt{N-2}$ (5)

Теорема

Если верна H_0, то с.в. T_{N-2} имеет распределение
Стьюдента $T_{N-2} \sim S_t$ с $(N-2)$ степенями свободы.

Для заданного уровня значимости α: $0<\alpha<1$ по таблицам распределения Стьюдента находим критическую константу $t_{N-2}(\alpha)>0$: $\mathbb{P}(|T_{N-2}|>t_{N-2}(\alpha)\,|\,H_0)=\alpha$ (6)

1) Если реально полученное значение T^*_{N-2} статистики T_{N-2}: $|T^*_{N-2}|>t_{N-2}(\alpha)$ => отвергаем H_0 => H_1 верна. (Установлено значимое влияние X на Y)

Если $|T^*_{N-2}|\le t_{N-2}(\alpha)$ => не можем отвергнуть H_0. (значимого влияния не обнаружено).

В пакетах прикладных программ по статистике используют другой подход (на самом деле, эквивалентный рассмотренному).

Пусть T^*_{N-2} – реально полученное значение статистики T_{N-2}. Вычислим вер-ть того, что наша с.в. больше T^*_{N-2}:

$\mathbb{P}(|T_{N-2}|>|T^*_{N-2}|\,|\,H_0)$ (7)

Достигнутый уровень значимости: p-value.
Чем менше p-value, тем более это свидетельство в пользу H_1.

2°. Множественный коэффициент корреляции.

Изучается влияние с.в. Y в зависимости от нескольких факторов $X_1,\dots,X_p$, где $p\ge 2$. Рассмотрим наилучшее линейное приближение

$\hat{Y}=\alpha+\beta_1X_1+\dots+\beta_pX_p$ (8)

через $X_1,\dots,X_p$.

Опр. Множественный коэфф-нт корреляции наз-ся число

$R_{Y\cdot X_1,\dots X_n}:=\rho(Y,\hat{Y})$ (9)

$e=Y-\hat{Y}$ – ошибка приближения.
Ошибка ортогональна ко всем X => e и $\hat{Y}$ – некорр.

$Y=\hat{Y}+e$ => $\mathbb{D}(Y)=\mathbb{D}(\hat{Y})+\mathbb{D}(e)$

=> $\mathbb{D}(\hat{Y})=R^2_{Y\cdot X_1,\dots,X_p}\cdot\mathbb{D}(Y)$ (10)

$\mathbb{D}(e)=(1-R^2_{Y\cdot X_1,\dots,X_p})\,\mathbb{D}(Y)$ (11)

Квадрат множественного коэфф-нта корреляции показывает, какую долю изменчивости Y ($\mathbb{D}(Y)$)

удалось объяснить совместным линейным влиянием выбранных факторов.

Теорема.

1) $|R_{Y \cdot X_1, \dots, X_p}| \le 1$

2) $|R_{Y \cdot X_1, \dots, X_p}| = \max$ по модулю простому коэфф. корр. между Y и $C_0 + C_1 X_1 + \dots + C_p X_p$

3) Если $R_{Y \cdot X_1 \dots, X_p} = 0$, тогда наилучшее линейное приближение $\hat{Y} = \mathbb{E}(Y)$.

Если $R_{Y \cdot X_1, \dots, X_p} = 1$ (или -1), то Y и $\hat{Y}$ совпадают почти наверное.

3°. Частный коэффициент корреляции.

Изучается влияние на с.в. Y набора с.в. $X_1, \dots, X_p$, $p \ge 2$.

Выберем некоторое k: $1 \le k \le p$.

Задача: каково "чистое" влияние X_k на Y?

Обозначим через C набор всех остальных факторов (остальных с.в.).

Рассмотрим наилучшее линейное приближение $\hat{Y}_C$ с.в. Y через C. Аналогично рассмотрим X_k^C – наилучшее линейное приближение X_k через C.

Пусть $Z_Y = Y - \hat{Y}_C$

$Z_{X_k} = X_k - X_k^C$

Опр. Частным коэфф-том корреляции с.в. Y и фактора X_k наз-ся число

$$R_{Y X_k \cdot C} := \rho(Z_Y, Z_{X_k}) \qquad (12)$$

Теорема

1) $|R_{Y X_k \cdot C}| \le 1$

2) $1 - R^2_{Y \cdot X_1, \dots, X_k} = (1 - R^2_{Y \cdot X_1 \dots \cdot X_{k-1}})(1 - R^2_{Y X_k \cdot X_1 \dots X_{k-1}})$

3) Если X_c и X_k некорр. со всеми остальными факторами из C, то $R_{kec} = \rho_{ke}$

Квадрат частного коэфф. корреляции показывает, какую долю необъяснённой ранее изменчивости (дисперсии) удалось объяснить добавлением ещё одного фактора.

$1 - \rho^2_{Y,X}$ – доля необъяснённой изменчивости

Замечание. т.к. не определения используют моменты не выше 2-го порядка и мн. совместности с.в., то и множественный, и частный коэфф-ты корр. можно вычислить, зная простые коэфф-ты корр.

§3. Пример

По итогам работы 37-ми однородных предприятий лёгкой промышленности регистрировались следующие показатели их работы:

$Y = X_0$ – среднемесячная характеристика качества ткани

Факторы:

X_1 – среднемесячное число профилактических проверок

X_2 – среднемесячное число обрывов нити

Вычислим парные (простые) коэфф-ты корр.:

$R_{01} = 0,105$

$R_{02} = -0,024$

$R_{12} = -0,996$

Полученные данные явно противоречат опыту работы в данной области.

Вычислим частные коэфф-ты корреляции:

$R_{01 \cdot 2} = 0,907$

$R_{02 \cdot 1} = -0,906$

Этот результат связан с тем, что X_1 и X_2 сильно связаны между собой и отрицательно корр., поэтому их совместное влияние маскирует влияние каждого фактора в отдельности.

Глава 5. Множественная линейная модель регрессии

§1. Классическая линейная модель.

Имеем одну объясняемую переменную Y и несколько объясняющих факторов или предикторов $X_1, \dots, X_m$.

Для этого произвели N независимых и одновременных измерений этих величин и предполагаем, что они связаны следующим образом: $\forall Y_j$ это какая-то функция от:

$$Y_j = g(X_{j1}, \dots, X_{jm}) + \varepsilon_j, \qquad j = \overline{1, N} \tag{1}$$

Основные ограничения:

1) Модель линейна по параметрам, т.е.

$$Y_j = \alpha + \beta_1 X_{j1} + \dots + \beta_m X_{jm} + \varepsilon_j \tag{2}$$

2) Факторы измерены без ошибок, т.е. X_{jk} – неслучайные величины

3) $\mathbb{E}(\varepsilon_j) = 0 \quad \forall j$, т.е. нет систематических ошибок

4) $\mathbb{D}(\varepsilon_j) = \sigma^2 > 0 \quad \forall j$ – условие гомоскедастичности (одинаковая дисперсия)

5) ε_j и ε_k – некоррелированы (независимы), $j \neq k$

6) ε_j имеет нормальное распределение

Далее введём следующее обозначение:

$$Y = \begin{bmatrix} Y_1 \\ \dots \\ Y_N \end{bmatrix}, \quad \vec{X}_k = \begin{bmatrix} X_{1k} \\ \dots \\ X_{Nk} \end{bmatrix}, \quad \vec{X}_0 = \begin{bmatrix} 1 \\ \dots \\ 1 \end{bmatrix}, \quad \Theta = \begin{bmatrix} \alpha \\ \beta_1 \\ \dots \\ \beta_m \end{bmatrix} = \begin{bmatrix} \Theta_0 \\ \Theta_1 \\ \dots \\ \Theta_m \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_N \end{bmatrix},$$

$$X = (X_{jk}) = \begin{bmatrix} X_{00} & X_{11} & \dots & X_{1m} \\ \dots & \dots & \dots & \dots \\ X_{N0} & X_{N1} & \dots & X_{Nm} \end{bmatrix} \Rightarrow$$

$$\boxed{Y = X \cdot \Theta + \varepsilon} = \Theta_0 \vec{X}_0 + \Theta_1 \vec{X}_1 + \dots + \Theta_m \vec{X}_m + \varepsilon \tag{3}$$

При выполн. основных ограничений:

$$\mathbb{E}Y = X \cdot \Theta$$

$$\Sigma_Y = \Sigma_\varepsilon = \sigma^2 \cdot E \tag{4}$$

(E – единичная матрица)

Основные задачи:

1) Оценка параметров модели, т.е. $\theta_0, \dots \theta_m$ и σ^2.
2) Проверка гипотез о параметрах
3) Отбор значимых факторов
4) Проверка адекватности модели
5) Проверка основных ограничений (как правило, это не делается в силу большой сложности, но это не есть хорошо.)

§2. Оценка параметров

Для оценки параметров Θ применяется метод наим. квадратов. Рассмотрим величину

$$Q(\Theta) = \|Y - X\cdot\Theta\|^2 = \sum_{j=1}^{N} [x_j - (x_{j0}\theta_0 + \dots + x_{jm}\cdot\theta_m)]^2 \to \min_{\Theta} \quad (5)$$

Необходимое условие экстремума: $\exists$ т. $Q'(\Theta) = 0$
$\Rightarrow m+1$ уравнение.

После некоторых преобразований получаем:

$$\boxed{X^T\cdot X\cdot\Theta = X^T\cdot Y} \quad \text{– система нормальных уравнений} \quad (6)$$

Пусть матрица $X^T\cdot X$ – невырожд. $\Rightarrow X$ имеет ранг $m+1 \Rightarrow$
$\Rightarrow \vec{X_0}, \vec{X_1}, \dots, \vec{X_m}$ – лин. нез. (базис)

$\Rightarrow$ решение системы

$$\hat{\Theta} = (X^TX)^{-1}X^TY = \Theta + (X^TX)^{-1}X^T\cdot\varepsilon \quad (7)$$

$\hat{\theta}$ – оценки по МНК.

$\hat{y} = X\cdot\hat{\Theta}$ – вектор предсказанных значений

$e = Y - \hat{y}$ – вектор остатков

$Y = \hat{Y} + e$

$\sigma^2 = \mathbb{D}(\varepsilon_j) = \mathbb{E}(\varepsilon_j^2)$

Рассмотрим: $\frac{1}{N}\sum_{j=1}\varepsilon_j^2 \to \sigma^2$, $N\to\infty$

В качестве оценки σ^2 предлагается взять:

$$\hat{\sigma}^2 = S^2 = \frac{1}{N-(m+1)}\sum_{j=1}^{N} e_j^2 \quad (8)$$

Чем хорош МНК? Фактически мы решаем задачу о наилучшей линейной оценке вектора Y в линейном пр-ве L, которое порождено векторами $\vec{X_0}, \dots, \vec{X_m}$

§3. Свойства оценок параметров

Лемма 1.

Пусть Y – случ. вектор с конечными моментами 2-го порядка, A – неслучайная матрица. Образуем новый вектор $Z = A \cdot Y$

Тогда $\mathbb{E} Z = A \cdot \mathbb{E} Y$

$$\Sigma_Z = A \cdot \Sigma_Y \cdot A^T$$

ковариац. матрица

Пусть Y – вектор измерений, γ – набор параметров.
$\hat{X}$ – оценка γ с помощью Y.

$$Y = \theta_0 \vec{X_0} + \ldots + \theta_m \cdot \vec{X_m} + \varepsilon$$

Оценка $\hat{\gamma}$ параметров γ наз-ся линейной (по Y), если она имеет вид $\hat{\gamma} = A \cdot Y$

Лемма 2. Пусть выполнены условия 1)–5) основных ограничений. Линейная оценка $\tilde{\theta} = AY$ параметров θ будет несмещ. $\Leftrightarrow AX = E$ (9)

Док-во: $\mathbb{E}(\tilde{\theta}) = \mathbb{E}(AY) = \mathbb{E}(A(X\theta + \varepsilon)) = AX\theta = \theta \quad \forall \theta \Leftrightarrow AX = E.$ ▨

Предположение 1. Пусть выполнены условия 1)–5) основных ограничений. Тогда $\hat{\theta}$ – оценка по МНК – яв-ся линейной и несмещённой.

Д-во:

$\hat{\theta} = (X^T X)^{-1} \cdot X^T \cdot Y = A \cdot Y$ – линейная

Далее: $A \cdot X = (X^T X)^{-1} \cdot X^T \cdot X = E$ – несмещённая. ▨

Предположение 2. Пусть выполнены 1)–5). Тогда матрица ковариации оценки $\hat{\theta}$ имеет вид:

$$\Sigma_{\hat{\theta}} = \sigma^2 \cdot (X^T \cdot X)^{-1} \quad (10)$$

Д-во!

$\hat{\theta} = (X^T \cdot X)^{-1} \cdot X^T \cdot Y = AY$

По лемме 1: $\Sigma_{\hat{\theta}} = A \cdot \Sigma_Y A^T = (X^T X)^{-1} \cdot X^T \cdot \sigma^2 \cdot E \cdot X (X^T \cdot X)^{-1} = \sigma^2 (X^T \cdot X)^{-1}$ ▨

Предположение 3. Пусть выполнены 1)–5) и диагональные элементы стремятся к 0 при $n \to \infty$. Тогда оценки $\hat{\theta}$ параметров θ по МНК яв-ся состоятельными.

(это следует из условия несмещённости и нер-ва Чебышёва)

02.10.

Теорема (Гаусса-Маркова)

Пусть выполнены условия 1)-5) основных ограничений, тогда оценка $\hat{\theta}$ по МНК пар-ра θ яв-ся оптимальной в средн. квадр. в классе всех линейных и несмещённых оценок.

Д-во:

$\hat{\theta} = (X^T X)^{-1} \cdot X^T \cdot Y = A \cdot Y$

$\tilde{\theta} = B \cdot Y$ – другая лин. несмещ. оценка

$B = A + C$

В силу несмещ. $CX = 0$

$\Sigma_{\tilde{\theta}} = B \cdot \Sigma_y \cdot B^T = (A+C)\sigma^2 \cdot E \cdot (A+C)^T = (A \cdot A^T + A \cdot C^T + C \cdot A^T + C \cdot C^T) \cdot \sigma^2$

$A \cdot C^T = (C \cdot A^T)^T = (X \cdot X^T)^{-1} \cdot X^T \cdot C^T = (X \cdot X^T)^{-1} \underbrace{(C \cdot X)^T}_{=0} = 0$

$\Rightarrow \Sigma_{\tilde{\theta}} = (A \cdot A^T + C \cdot C^T)\sigma^2 = \Sigma_{\hat{\theta}} + C \cdot C^T \cdot \sigma^2$

$C \cdot C^T$ – неотриц. опред. матрица

$\Rightarrow \Sigma_{\tilde{\theta}} - \Sigma_{\hat{\theta}} = (C \cdot C^T)\sigma^2$ – неотр. опред.

$\Rightarrow$ на диагонали стоят неотр. числа

$\Rightarrow \mathbb{D}(\tilde{\theta}_k) \geq \mathbb{D}(\hat{\theta}_k) \quad \forall k$ ▨

Рассмотрим оценку для $\hat{\sigma}^2 = S^2 = \frac{1}{N-(m+1)} \sum_{j=1}^{N} e_j^2$

Предложение 4. Пусть выполнены условия 1)–6). Тогда оценка S^2 по МНК для σ^2 яв-ся несмещённой и состоятельной

§4. Построение доверительных интервалов

теорема о нормальной корреляции ?

Всюду далее считаем, что выполнены св-ва 1)-6).

Теорема.

Пусть $\hat{\theta}$ – оценка по МНК пар-ра θ, тогда справедливы след. св-ва:

1) $\hat{\theta}$ имеет многомерное нормальное распред. с вектором средних: $X \cdot \theta$ и матрицей ковариации: $\Sigma_{\hat{\theta}} = (X^T \cdot X)^{-1} \cdot \sigma^2$

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

3) с.в. $\sum_j (y_j - \hat{y}_j)^2 / \sigma^2 = \sum_j e_j^2 / \sigma^2$ имеет χ^2-распределение с $(N-(m+1))$ степенями свободы

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

Фиксируем некот. k: $0 \leq k \leq m$. Обозначим $B = (X^T X)^{-1} = (b_{jk})$.

В силу теорем с.в. $\hat{\theta}_k$ имеет одномерное норм. распред. со средним θ_k и дисперсией $b_{kk} \cdot \sigma^2$.

$\Rightarrow \frac{\hat{\theta}_k - \theta_k}{\sqrt{b_{kk}\cdot\sigma^2}} : \frac{\sigma}{S} = \frac{\hat{\theta}_k - \theta_k}{S\cdot\sqrt{b_{kk}}} = \frac{\hat{\theta}_k - \theta_k}{S_k}$ – имеет распр. Стьюдента с $(N-(m+1))$ степенями свободы

$S_k = S\cdot\sqrt{b_{kk}}$ – стандартная ошибка оценки пар-ра θ_k.

По заданному доверительному уровню $\gamma \sim 1$ находим по таблицам критическую константу $t_{N-(m+1)}(\gamma) > 0$:

$$P\left(|T_k| < t_{N-(m+1)}(\gamma)\right) = \gamma$$

$$|T_k| = \left|\frac{\hat{\theta}_k - \theta_k}{S_k}\right| < t_{N-(m+1)}(\gamma)$$

$$\Rightarrow \quad \hat{\theta}_k - t_{N-(m+1)}(\gamma)\cdot S_k < \theta_k < \hat{\theta}_k + t_{N-(m+1)}(\gamma)\cdot S_k$$ (с вер-ю γ)

§5. Проверка значимости одного фактора.

Рассмотрим задачу: фиксируем k (номер фактора): $0 \le k \le m$.

Проверяется гипотеза $H_0: \theta_k = \theta_{k0}$

$H_1: \theta_k \neq \theta_{k0}$

В силу теоремы из §4 утверждается, что с.в. $T_k = \frac{\hat{\theta}_k - \theta_k}{S_k}$ имеет при верной H_0 распределение Стьюдента с $(N-(m+1))$ степенями свободы.

Для заданного уровня значимости $\alpha \sim 0$ находим критическую const $t^*_{N-(m+1)}(\alpha)$ $\left(= t_{N-(m+1)}(\gamma = 1-\alpha)\right)$

$$P\left(|T_k| > t^*_{N-(m+1)}(\alpha) \,\middle|\, H_0\right) = \alpha$$

Далее:

1) Если реально полученное значение T_k^* статистики T_k: $|T_k^*| > t^*_{N-(m+1)}(\alpha)$, то отвергаем H_0

2) В противоположной ситуации H_0 не отвергается.

] $\theta_{k0} = 0$, т.е. фактор не оказывает влияния на Y.

$\Rightarrow$ Гипотеза о значимости влияния фактора X_k. (Один фактор!)

$\Rightarrow$ T-критерий Стьюдента. (см. выше)

Если фактор признан незначимым, то его удаляют из модели и всё оценивают заново.

§6. Проверка линейных гипотез

1°. Общая линейная гипотеза

Имеем модель измерений:

$$Y = X \cdot \theta + \varepsilon \qquad (1)$$

Всюду далее выполнены условия 1)–6).

Проверяем гипотезу:

$$H_0: A \cdot \theta = a$$
$$H_1: A \cdot \theta \neq a \qquad (2)$$

A – заданная неслучайная матрица размера $p \times (m+1)$.

a – заданный неслучайный вектор размера p.

Предполагаем, что $p \leq m+1$.

$\operatorname{rank}(A) = p$ (p лин/нез ограничений)

Описание алгоритма проверки общей линейной гипотезы:

1) Оцениваем модель (1) без учёта ограничений (2) и находим сумму квадратов остатков ESS_{UR}

2) Оцениваем модель (1) с учётом ограничений (2) и находим сумму квадратов остатков ESS_R

3) При верной H_0 с.в. ESS_{UR}/σ^2 и $(ESS_R - ESS_{UR})/\sigma^2$ независимы и имеют χ^2-распределение с $(N-(m+1))$ и p степенями свободы

4) При верной гипотезе H_0 с.в. $F = \dfrac{(ESS_R - ESS_{UR})/p}{ESS_{UR}/(N-(m+1))}$ — имеет распределение Снедекора–Фишера с $(p, N-(m+1))$ степенями свободы.

5) Для заданного $\alpha > 0$ находим по таблицам критическую const $F(\alpha) > 0$: $P(F > F(\alpha) \mid H_0) = \alpha$

6) Если реально полученное значение F^* статистики F таково, что $F^* > F(\alpha)$, то отвергаем H_0. В противном случае H_0 не противоречит экспериментальным данным.

$\Rightarrow$ F-критерий. (см. выше) (тесное родство этого факта)

F-критерий

2°. Проверка значимости влияния одного фактора

Проверяем

$H_0: \theta_k = 0$

против

$H_1: \theta_k \neq 0$

В случае простой линейной регрессии F- и T-критерии дают один и тот же ответ.

3°. Проверка значимости влияния группы факторов.

Выберем $1 \le i_1 < i_2 < \dots < i_p \le m$

Проверяемая гипотеза

$H_0: \theta_{i_1} = \theta_{i_2} = \dots = \theta_{i_p} = 0$

против $\exists k$:

$H_1: \theta_{i_k} \neq 0.$

Если H_0 не отвергается, то все факторы θ_{i_j} одновременно исключаются (их значимого влияния не обнаружено).

4°. Проверка адекватности модели.

В нашей задаче под адекватностью будем понимать, что выбранный набор факторов охватывает значимое влияние на исследуемую переменную.

$H_0: \theta_1 = \theta_2 = \dots = \theta_m = 0$

против $\exists k$: $1 \le k \le m$

$H_1: \theta_k \neq 0.$

Можно показать, что справедливо следующее тождество:

$$\sum_j (Y_j - \bar{Y})^2 = \sum_j (Y_j - \hat{Y}_j)^2 + \sum_j (\hat{Y}_j - \bar{Y})^2$$

$$TSS = ESS + RSS$$

TSS – полная сумма квадратов – полная изменчивость Y под влиянием как учтённых, так и неучтённых факторов. (total sum of squares)

ESS – остаточная сумма квадратов – та часть изменчивости Y, которая определяется неучтёнными факторами (error)

RSS – объяснённая сумма квадратов – та изменчивость Y, которая объясняется влиянием учтённых факторов (regression)

Бывает $\frac{ESS}{explained} \leftrightarrow \frac{RSS}{residual}$ (в некоторых книгах)

Справедлива следующая

Теорема

Если верна H_0, то

1) ESS и RSS – независимы

2) ESS/σ^2 и RSS/σ^2 имеют χ^2-распред. с $(N-(m+1))$ и $(m+1)$ степенями своб.

3) с.в. $F = \frac{RSS/(m+1)}{ESS/(N-(m+1))}$ имеет распр. Снед-Фишера с $((m+1), (N-(m+1)))$ степ. своб.

Дальше проверка Но проводится по стандартной схеме.

Опр. Величина

$$R^2 := \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}$$ назыв. коэффициентом детерминации.

Он хар-зует качество подгонки модели.

Для хороших моделей близок к 1.

$$\Rightarrow F = \frac{N-m-1}{m+1} \cdot \frac{R^2}{1-R^2}$$

5° Пример.

Изучается индекс импорта товаров и услуг (Y) в зав-сти от индекса ВВП (X_1) и отношения индексов цен на импортные товары и ВВП (X_2).

Данные за 9 лет:

Y	X_1	X_2	
100	100	100	← базовый уровень
106	104	99	← % к базовому периоду
107	106	110	
120	111	126	
110	111	113	
116	115	103	
123	120	102	
133	124	103	
137	126	98	

Рассматриваем модель измерений: (лин. рег. модель)

$$Y_j = \alpha_1 + \beta_1 X_{j1} + \beta_2 X_{j2} + \xi_j$$

Применяем к имеющимся данным и нашей модели МНК:
(Например, пакет EViews)

Получим таблицу:

Var	Coef	SE	t-stat	p-value
X_0	-49.34	24.06	-2.0506	0.0862
X_1	1.364	0.143	9.5299	0.001
X_2	0.114	0.143	0.7943	0.4573

$S^2 = 12.92$

$F^* = 45.78$

$\Rightarrow$ p-value $= 0.0002$

$R^2 = 0.9385$

Выводы:

1) Выявлено значимое совместное влияние факторов.
2) Вторичные факторы объясняют поведение Y на 94%.
3) Фактор X_1 оказывает значимое влияние. X_2 нет
4) X_2 является незначимым фактором. $\Rightarrow$

=> Удалим X_2 из уравнения и оценим заново.

Var	coef	SE	t-stat	p-value
X_0	-35.08	24.06	-2.25	0.059
X_1	1.345	0.137	9.8	0.000...

$R^2 = 0.92$ $F^* = 96$, p-value = 0.000...

6°. Выбор модели

В нашей методике мы подбирали по величине R^2, но только по ней судить нельзя (если 5 лишних факторов дают лишь 1% точности, то они не нужны), а ещё она слегка завышенная. => Введём несмещенную

$R^2_{adj} = 1 - (1-R^2)\frac{N-1}{N-m-1}$ – исправленный коэфф. детерминации.

adjusted

7°. Сравнение двух регрессий

(Когда собираются данные из разных совокупностей, то вообще говоря, неочевидно, что они из одной регрессии (подходят модели).)

Собрано два набора наблюдений.

В первом имеем след. модель изначально.

$$Y_j = \beta_1' \cdot X_{j1} + \beta_2' X_{j2} + \dots + \beta_m' X_{jm} + \varepsilon_j', \quad j = \overline{1, N_1} \qquad (1)$$

Во втором

$$Y_j = \beta_1'' \cdot X_{j1} + \beta_2'' \cdot X_{j2} + \dots + \beta_m'' X_{jm} + \varepsilon_j'', \quad j = \overline{N_1+1, N_1+N_2} \qquad (2)$$

$\sigma'^2 = \sigma''^2$

Проверяется гипотеза

$H_0: \beta_k' = \beta_k'' \quad \forall k = \overline{1, m}$

против (3)

$H_1: \exists k: \beta_k' \neq \beta_k''$.

Рассмотрим вспомогательную задачу:

$$Y_j = \beta_1' \cdot \tilde{X}_{j1} + \dots + \beta_m' \cdot \tilde{X}_{jm} + \beta_1'' \cdot \tilde{X}_{j,m+1} + \dots + \beta_m'' \cdot \tilde{X}_{j,2m} + \varepsilon_j, \quad j = \overline{1, N_1+N_2} \qquad (4)$$

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

$$\tilde{X}_{jk} = \begin{cases} 0, & 1 \le j \le N_1 \\ X_{j,k-m}, & N_1+1 \le j \le N_1+N_2 \end{cases} \text{ для } m+1 \le k \le 2m.$$

В рамках модели (4) проверяем гипотезу (3) с помощью F-критерия.

=> Критерий Чоу (Chow). Для него нужно знать момент N_1 – момент потенциального изменения коэф-та наблюдений

Пример. Изучается разница в уровне зарплат мужчин и женщин. Δ 2 фактора.

$X_1 = age$ - возраст

$X_2 = edu$ - уровень образования

N_1 мужчин, N_2 женщин. $N_1 = N_2 = 75$.

$W = wage$ - уровень зарплаты. (рублей/час)

Рассматриваем ур-ния регрессии след. вида: 16.10

$$w_j = \alpha + \beta_1 \cdot age_j + \beta_2 \cdot edu_j + \varepsilon_j$$

Оценим по МНК это уравнение для данных по мужчинам.

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

$$ESS_1 = 5672.328$$

Для данных по женщинам получаем:

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

$$ESS_2 = 1788.344$$

$N_1 = N_2 = 72$, $m = 3$ 3 фактора

2 разных ур-я это как одно, но без учёта ограничений. Общая ESS:

$$ESS_{UR} = ESS_1 + ESS_2 = 7460.672$$

Оценим без учёта ограничений, как будто это 2 разных ур-я.

Теперь оценим ур-ние по всем данным.

$$w = -3.06 + 4.78 \cdot age + 3.254 \cdot edu$$

$$ESS_R = 8080.443$$

Найдём наблюдаемое значение статистики F:

$$F^* = \frac{(ESS_R - ESS_{UR})/m}{ESS_{UR}/(N-2m)} = 3.342$$

150-6

По таблице распр-я Снедекора-Фишера с (3, 144) степ. свободы получаем, что это отклонение значимо на уровне $\alpha = 0.05$.

⇒ Действительно, в ф-ровании з/п мужчин и женщин обнаружено значимое различие.

§10. Ортогональная матрица плана.

Имеем мн. модель

$$Y = X \cdot \Theta + \varepsilon \qquad (1)$$

X – матрица плана эксперимента.

Если оценить Θ по МНК, то

$$\hat{\Theta} = (X^T \cdot X)^{-1} \cdot X^T \cdot Y$$

$$X^T X \cdot \Theta = X^T \cdot Y \qquad (2)$$

Опр. Матрица плана X называется ортогональной, если ортогональны столбцы этой матрицы.

$\Rightarrow X^T \cdot X$ – диаг. матрица.

В этом случае СУ (2) разбивается на $n+1$ независимых ур-ий. Тогда при удалении какого-то фактора ур-я для других параметров останутся теми же самыми $\Rightarrow$ не нужно будет пересчитывать при удалении/добавлении факторов. Если $X^T X$ близка к вырожденной, то решение очень неустойчиво. Это условие мультиколлинеарности: факторы между собой сильно связаны. $\Rightarrow$ надо убирать какие-то факторы, чтобы её не было.

§11. Фиктивные переменные.

Когда есть категор. признаки, вводят фиктивные (dummy) пер-ки.

Пример 1.

Изучается уровень з/п Y в зав-ти от некоторых колич. факторов $X_1, \dots X_p$ а ещё от наличия/отсутствия высшего образования.

$$Y_j = \beta_1 \cdot X_{j1} + \dots + \beta_m \cdot X_{jm} + \varepsilon_j$$

$$d_j = \begin{cases} 1, \text{ если есть высш. об.} \\ 0, \text{ в прот. случае} \end{cases}$$

$\Rightarrow$ рассматриваем след. модель

$$Y_j = \beta_1 \cdot X_{j1} + \dots + \beta_m \cdot X_{jm} + \gamma \cdot d_j + \varepsilon_j$$

Но не всегда интересующий фактор принимает только 2 значения...

Пример 2.

Изучается спрос на мороженое в разные сезоны.

$$Y_j = \beta_1 X_{j1} + \dots + \beta_m \cdot X_{jp} + \varepsilon_j$$

$\Rightarrow$ Введём несколько фик. перемен., а не одну со значениями 0,1,2,3, т.е. коэфф. перед ней один и тот же, а смысл перехода между сезонами будет неправильное (осень-зима > весна-лето).

Нужно ли нам 4 пер-ки? Нет, они мн/зав, достаточно 3-х.

Введём пуп фикт. перем:

$d_j^{(1)} = \begin{cases} 1, \text{ если жена} \\ 0, \text{ в пр. случае} \end{cases}$

$d_j^{(2)} = \begin{cases} 1, \text{ если вена} \\ 0, \text{ в пр случае} \end{cases}$

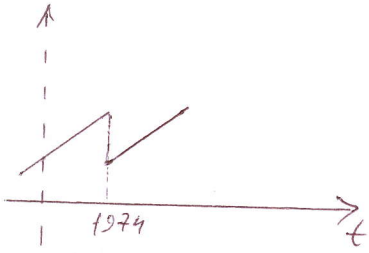
$d_j^{(3)} = \begin{cases} 1, \text{ если лето} \\ 0, \text{ в пр. случае} \end{cases}$

$$Y_j = \beta_1 \cdot x_{j1} + \dots + \beta_p \cdot x_{jp} + \gamma_1 \cdot d_j^{(1)} + \gamma_2 \cdot d_j^{(2)} + \gamma_3 \cdot d_j^{(3)} + \varepsilon_j$$

Пример 3

1973 - арабо-израильская война

$\Rightarrow$ 1974 г.:



$\Rightarrow$ уже нелин. модель $\Rightarrow$ введём фикт. перем.

$d_j = \begin{cases} 1, & t < 1974 \\ 0, & \text{иначе} \end{cases}$

Такой сдвиг тоже можно учесть: $\backslash\!/$

$\Rightarrow$ учёт структурных изменений.

Пример 4

К задаче о формир. з/п в Болонии для м и ж. (см. § 9).

Изучается уровень з/п W в зависимости от стажа (возраста) и пола работника.

age - стаж

S - пол работника

$S_j = \begin{cases} 1, \text{ если женщина} \\ 0, \text{ в прот. случае} \end{cases}$

Рассматриваем след. модель:

$W_j = \alpha + \beta \cdot age_j + \gamma \cdot S_j + \varepsilon_j$

Применим МНК (ко всем данным)

coef	Est	SE	t	p-value
α	0.11	2.39	2.56	0.011
β	0.53	0.06	8.58	0.000...
γ	-3.73	1.35	-2.76	0.0065

$\Rightarrow W = -0.11 + 0.53 \cdot age - 3.73 \cdot S$

$R^2 = 0.39$

$F^* = 0.000...$

$\boxed{\hat{\gamma} = -3.73}$

$\Rightarrow$ з/п для м и ж отличается при прочих равных условиях на 3.73 ушедена/час

Глава 6. Обобщения классической линейной модели.

§1. Обзор результатов для классической модели.

Имеем одну зависимую переменную и m объясняющих факторов. Провели N измерений, которые связаны след. образом:

$$Y_j = \alpha + \beta_1 X_{j1} + \dots + \beta_m \cdot X_{jm} + \varepsilon_j \quad (1)$$

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

Основные ограничения.

1) модель лин. по параметрам
2) X – неслучайная матрица
3) $\{\varepsilon_j\}$ – некорр. (незав.)
4) $\mathbb{E}(\varepsilon_j) = 0 \ \forall j$ – нет систематических ошибок
5) $\mathbb{D}(\varepsilon_j) = \sigma^2 \ \forall j$ – условие гомоскедастичности = точность измерений одна и та же
6) $\{\varepsilon_j\}$ – имеют норм. распред.

Если матрица плана X невыр., то

$$\hat{\theta} = (X^T \cdot X)^{-1} \cdot X^T \cdot Y = \theta + (X^T X)^{-1} \cdot X^T \cdot \varepsilon$$

Оценка для σ^2:

$$S^2 = \frac{1}{N-(m+1)} \cdot \sum_{j=1}^{N} e_j^2, \text{ где } \hat{Y} = X \cdot \hat{\theta}, \ e_j = Y_j - \hat{Y}_j.$$

Свойства оценок.

1) $\hat{\theta}$ – лин. по Y.
2) $\hat{\theta}$ – не смещ.
3) $\hat{\theta}$ – оптим. в кв. квадр. в классе лин. и несмещ. оценок.
4) Если $\frac{1}{N} \cdot X^T \cdot X \to \Sigma_X$, $N \to \infty$, и Σ_X – невыр. матр. размера $(m+1)$, то $\hat{\theta}_N$ – сост.
5) S^2 – несмещ. и сост. оценка для σ^2.

§2. Модель со стохаст. факторами

1° Описание модели.

Провели N измерений.

1) модель лин. по параметрам, т.е.

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

2) X – случ. матрица, но $X_{j0} = 1 \ \forall j$.
и $\mathbb{E} X_{jk} = 0 \ \forall k = \overline{1,m}$

3) Потребуем: X и ε – независимы.

4) $\mathbb{E}(\varepsilon_j) = \mathbb{E}(\varepsilon_j | X) = 0 \quad \forall j$

5) $\Sigma_\varepsilon = \mathbb{E}(\varepsilon \cdot \varepsilon^T) = \mathbb{E}(\varepsilon \cdot \varepsilon^T | X) = \sigma^2 \cdot E.$

6) сл. вектор $(Y_j, X_{j1}, \dots X_{jm})$ имеет одно и то же многомерное норм. невырожд. распред.

Если выполнен перечисленные выше ограничения, то в силу ЗБЧ имеем: 23.10.

$$\frac{1}{N} \varepsilon^T \cdot \varepsilon = \frac{1}{N} \sum_{j=1}^{N} \varepsilon_j^2 \to \sigma^2 \qquad (2)$$

$\leftarrow cov(X,X)$

$$\frac{1}{N} X^T \cdot X \to \Sigma_{XX} \qquad (3)$$

← аналог закона больших чисел для моментов и оценок

$$\frac{1}{N} X^T \cdot \varepsilon \to 0 \qquad (4)$$

при $N \to \infty$.

2°. Оценки параметров и их свойства.

Применим к (1) МНК. $\Rightarrow$

$$\hat{\theta} = (X^T \cdot X)^{-1} \cdot X^T \cdot Y = \theta + (X^T \cdot X)^{-1} \cdot X^T \cdot \varepsilon \qquad (5)$$

Обозначим $\hat{Y} = X \cdot \hat{\theta}$, $e = Y - \hat{Y} \Rightarrow$

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

Пусть вып-ны основные огр. 1)–6).

Свойства оценок.

1) $\hat{\theta}$ – линейна по Y

2) $\hat{\theta}$ – несмещ. оценка θ. (доказывается через функцию $\theta + \delta(\omega x)$)

3) $\hat{\theta} = \hat{\theta}_N$ – состоятельная.

XX – невырожд. $\Rightarrow (XX)^{-1} \exists$

Д-во:

$$\hat{\theta} = \theta + (X^T \cdot X)^{-1} \cdot X^T \cdot \varepsilon = \theta + \left(\frac{1}{N} X^T \cdot X\right)^{-1} \cdot \left(\frac{1}{N} X^T \cdot \varepsilon\right) \Rightarrow \theta + \Sigma_{XX}^{-1} \cdot 0 = \theta$$

Важным понятием дек-ва яв-ся то, что факторы не коррелируют с ошибками. Если это не так, то св-во состоятельности нарушается. ∎

4) $\hat{\theta}$ – оптимальная в среднеквадр. в классе всех лин. и несмещ. оценок.

Д-во: напишем нер-во для дисперсий для какого X, а потом усредним по X.

5) S^2 – есть несмещ. и сост. оценка для σ^2.

3°. Метод инструментальных переменных.

Если некот. факторы корр. с ошибками, то их надо заменить на другие факторы, которые не будут корр-ны.
Пусть имеем новую матрицу Z размера $N \times (m+1)$, $Z_{j0} \equiv 1$, причём

$$\frac{1}{N} Z^T \cdot \varepsilon \to 0 \qquad (7)$$

и

$$\frac{1}{N} Z^T \cdot X \xrightarrow{P} \Sigma_{ZX}, \qquad (8)$$

где Σ_{ZX} – невыр. квадр. матрица.

Запишем новую оценку

$$\tilde{\theta} = (Z^T \cdot X)^{-1} \cdot Z^T \cdot Y = \theta + (Z^T \cdot X)^{-1} \cdot Z^T \cdot \varepsilon = \theta + \left(\frac{1}{N} Z^T \cdot X\right)^{-1} \cdot \left(\frac{1}{N} Z^T \cdot \varepsilon\right) \to$$

$$\to \theta + \Sigma_{ZX}^{-1} \cdot 0 = \theta.$$

Новые переменные в Z – инструментальные переменные.
Экономисты в кач-ве таких берут экзогенные переменные.

§3. Обобщённый МНК.

В класс. мн. модели:

$$Y = X \cdot \theta + \varepsilon \qquad (1)$$

причём

$$\Sigma_{YY} = \Sigma_{\varepsilon\varepsilon} = \sigma^2 \cdot E \qquad (2)$$

Пусть мы имеем

$$\Sigma_{\varepsilon\varepsilon} = \sigma^2 \cdot \Omega, \qquad (3)$$

где Ω – известная симмет. положит. опред. матрица.
Применим обычный МНК. $\Rightarrow$

$$\hat{\theta} = (X^T \cdot X)^{-1} \cdot Y = \theta + (X^T \cdot X)^{-1} \cdot X^T \cdot \varepsilon \quad \Rightarrow$$

1) $\hat{\theta}$ – мн. по Y.
2) $\hat{\theta}$ – несмещ.
3) при тех же условиях $\hat{\theta}$ – состоят.
4) св-во оптимальности не выполняется.

Из линейной алгебры известно, что $\exists$ невыр. матрица P:

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

Перейдём к новым переменным: (всё умножим на P^{-1})

$$Y_* = P^{-1} \cdot Y, \quad X_* = P^{-1} \cdot X, \quad \varepsilon_* = P^{-1} \cdot \varepsilon \quad \Rightarrow$$

$$Y_* = X_* \cdot \theta + \varepsilon_* \qquad (5)$$

Покажем, что (5) – это класс. мн. модель.

$$\Sigma_{\varepsilon^*\varepsilon^*} = \mathbb{E}\left(\varepsilon_* \varepsilon_*^T\right) = \mathbb{E}\left(P^{-1} \cdot \varepsilon \cdot \varepsilon^T \cdot (P^{-1})^T\right) = P^{-1} \cdot \mathbb{E}(\varepsilon \cdot \varepsilon^T) \cdot (P^{-1})^T =$$

$$= \sigma^2 \cdot P^{-1} \cdot \Omega \cdot (P^{-1})^T \overset{(4)}{=} \sigma^2 \cdot P^{-1} \cdot P \cdot P^T (P^T)^{-1} = \sigma^2 \cdot E$$

Применим к (5) обычный МНК. Полученные оценки будут оптимальными.

$$(Y_* - X_* \cdot \theta)^T \cdot (Y_* - X_* \cdot \theta) = (Y - X \cdot \theta)^T \cdot \Omega^{-1} (Y - X \cdot \theta) \to \min_{\theta} \quad (6)$$

– обобщенный МНК.

$\hat{\theta}_* = (X_*^T \cdot X_*)^{-1} \cdot X_*^T \cdot Y_* = (X^T \Omega^{-1} \cdot X)^{-1} \cdot X^T \cdot \Omega^{-1} \cdot Y$ – оценка Эйткена

Важный частный случай:

Пусть Ω – диаг. матрица с элементами $w_1, w_2, \ldots, w_N$.

Т.е. ошибки некорр., но, возможно, имеют разные дисперсии и мы знаем соотношения между их дисперсиями.

Ω^{-1} – диаг. матр. с эл-ами $\frac{1}{w_1}, \ldots, \frac{1}{w_N}$.

$$(6) \Rightarrow \sum_{j=1}^{N} [Y_j - (X\theta)_j]^2 / w_j \to \min_{\theta}. \quad \text{– взвешенный МНК.} \quad (7)$$

Глава 7. Гетероскедастичность и автокорреляция ошибок.

$$Y = X\cdot\theta + \varepsilon \qquad (1)$$

В классич. лин. модели $\Sigma_{\varepsilon\varepsilon} = \sigma^2\cdot E$ (2)

Мы рассм-м случай, когда $\Sigma_{\varepsilon\varepsilon} = \sigma^2\cdot\Omega$, (3)

где Ω известна.

Если Ω неизвестна, то имеем $\frac{N(N+1)}{2}$ дополнительных параметров, кроме θ и σ^2.

А измерений всего N. (!)

2 трудности в $\Sigma_{\varepsilon\varepsilon}$: когда дисперсии разные (на диаг не равные числа) и корр-ты: вне диаг не нули => ∆ 2 случая по очереди.

§1. Гетероскедастичность. (когда ошибки некорр но с разн. дисп.)

1. Описание модели

Задан. модель

$$Y = X\cdot\theta + \varepsilon \qquad (1)$$

и

$$\Sigma_{\varepsilon\varepsilon} = \sigma^2\cdot\Omega, \qquad (2)$$

где Ω – диаг. матрица.

Пусть Ω – неизвестна!

Применим обычный МНК. Получим оценки

1) лин.
2) несмещ.
3) состоят.
4) не оптимальны!

Вновь имеем N доп. парам., кроме θ и σ^2 => Опять проблема так не можем такое решить и будем ∆ спец. случаи

Модель описанного выше типа наз-ся моделью лин. регрессии с гетероскедастичностью в модели измерений.

2°. Модели гетероскедастичности

Рассмотрим модель, в которой ошибки пропорциональны одному из факторов, т.е.

$$\sigma_j^2 = E(\varepsilon_j^2) = \sigma^2\cdot|x_{jl}|^2 \qquad (3)$$

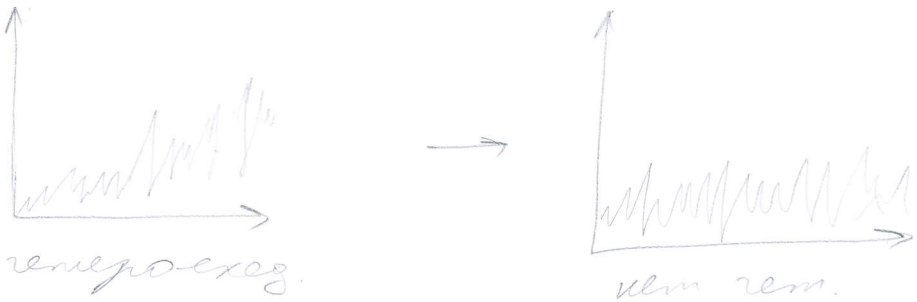
Фактически, это ситуация, где применим взвешенный МНК.

Реально оценку производят следующим образом: переходим к новым переменным:

$Y_j^* = Y_j/|x_{jl}|$, $x_{jk}^* = x_{jk}/|x_{jl}|$, а далее применяем обычный МНК.

Существует простой визуальный метод проверки того, что модель имеет форму (3). Не аккуратный, просто чтобы определить вид, потом исп-ся более аккур. методы

1) Применяем обычный МНК к набору данных, где не упорядочено по величине $|X_{je}|$.
2) Находим остатки e_j и строим график $|e_j|$.
3) Если на графике видна тенденция роста величин $|e_j|$, это в пользу (3).
4) Переходим к новым переменным, дел на $|X_{je}|$. Снова оцениваем (1) по МНК, находим e_j и строим график $|e_j|$.
5) Если на графике $|e_j|$ видим хаотическое колебание (т.е. нет явной тенденции), то это опять в пользу модели (3).



Другая модель.

$$\sigma_j^2 = \gamma_0 + \gamma_1 z_{j1} + \ldots + \gamma_p \cdot z_{jp} \qquad (4)$$

Предлагается след. процедура для оценки модели:

1) Оцениваем (1) по обычному МНК и находим остатки e_j.
Остатки должны себя вести похоже на ошибки
2) Используя остатки e_j, оцениваем вспомогательную регрессию следующего вида:

$$e_j^2 = \gamma_0 + \gamma_1 z_{j1} + \ldots + \gamma_p \cdot z_{jp} + u_j \qquad (5)$$

Далее берём

$$\hat{\sigma}_j^2 = \hat{\gamma}_e + \hat{\gamma}_1 \cdot z_{j1} + \ldots + \hat{\gamma}_p \cdot z_{jp} \qquad (6)$$

3) Применяем к (1) взвешенный МНК с весами $\hat{\sigma}_j^2$.
4) По новым остаткам оцениваем дисперсии оценок $\hat{\theta}$.

– Двухшаговый МНК.

Тесты на гетероскедастичность.

Нулевая гипотеза имеет вид:

$H_0: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_N^2 = \sigma^2$

H_1: это не так

Тест Гольдфельда–Куандта

$H_1: \sigma_j^2 = \sigma^2 \cdot X_{je}^2.$

Хотим выстроить дисперсии по возрастанию. Для этого выстр. X_{je} по возраст.

Описание процедуры:

1) Все переменные упорядочиваются по возрастанию переменной $|X_e|$.
2) Из всех данных удаляем среднее d измерений. (обычно $d \sim N/2$ – порядка четверти).
3) Строим 2 независимые регрессии по первым $(N-d)/2$ измерениям и последним $(N-d)/2$ измерениям и находим и там, и там вектора остатков e_1 и e_2.
4) По остаткам e_1 и e_2 вычисляем суммы квадратов остатков ESS_1 и ESS_2 и статистику $F = \frac{ESS_2}{ESS_1}$.
5) Если верна H_0, то статистика F имеет распределение Снедекора–Фишера с $\left(\frac{N-d}{2} - p, \frac{N-d}{2} - p\right)$ степ. свободы. p – число параметров в уравн. регрессии
6) Далее процедура проверки строится по станд. схеме: строится H_0, вычисляется p-value, ...

Если обнаружена гетероскедастичность, то применяется взвешенный МНК с весами $\sigma^2 X_{je}^2$ (из гипотезы H_1).

§2. Модели с автокорреляцией в ошибках.

1°. Описание модели.

Рассматриваем линейную модель измерений:

$$Y = X \cdot \Theta + \varepsilon \qquad (1)$$

Теперь предположим, что ошибки измерений удовлетворяют след. соотношению:

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

– модель авторегрессии (след. наблюдение выраж-ся через себя же на пред. шаге)

1) $|\rho| < 1$

2) $\delta_j \sim N(0, \sigma^2)$

(2) – модель авторегрессии 1-го порядка

Пусть ε_0 независима от набора $\{\delta_j\}$ и имеет норм. распред., также, что $\mathbb{E}(\varepsilon_0)=0$, $\mathbb{D}(\varepsilon_0)=\frac{\sigma^2}{1-\rho^2}=\sigma_\varepsilon^2$.

Тогда $\mathbb{E}(\varepsilon_j)=0$, $\mathbb{D}(\varepsilon_j)=\sigma_\varepsilon^2 \quad \forall j$

$\operatorname{cov}(\varepsilon_j, \varepsilon_{j+m}) = \sigma_\varepsilon^2 \cdot \rho^m$ (3)

$\Rightarrow \{\varepsilon_j\}$ – стационарная в широком смысле случ. послед-ть.

Структура матрицы ковариации:

$$\Sigma_\varepsilon = \sigma_\varepsilon^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{N-1} \\ \rho & 1 & \rho & \dots & \rho^{N-2} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \end{pmatrix}$$

Если применить обычный МНК, то оценки для θ будут несмещёнными, но не будут оптимальными.

Пусть ρ известно. Рассмотрим новую модель:

$$Y_j^* = Y_j - \rho Y_{j-1} = \theta_0(1-\rho) + \theta_1(X_{j1} - X_{j-1,1}\rho) + \dots + \theta_m(X_{jm} - X_{j-1,m}\rho) + \delta_j \quad (4)$$

$j \geq 2$. или теряем 1-е наблюдение

(4) – авторегрессионная модель.

2° Оценивание модели с авторегрессией.

Процедура Кохрейна – Оркатта:

1) Применим обычный МНК к модели (1) и найдём остатки e.
2) Рассмотрим модель регрессии вида:
$$e_j = \rho e_{j-1} + \eta_j, \quad j \geq 2$$
$\Rightarrow$ находим оценку $\hat{\rho}$ для ρ по МНК
3) Подставляем $\hat{\rho}$ вместо ρ в (4). Оцениваем модель (4) по обычному МНК.
4) Находим новый вектор остатков $\tilde{e} = y - X\hat{\theta}$.
5) Повторяем 2)–4), пока не будет достигнута нужная точность оценки ρ.

Процедура Хилдрета – Лу

1) Разбиваем $[-1; 1]$ на некую сетку
2) Для каждого значения ρ проводим авторегрес. преобразование.
3) Оцениваем по МНК преобр. модель
4) Выбираем то значение ρ, где была наименьшая сумма квадратов остатков.
5) В окрестности найденного ρ строим более мелкую сетку и повторяем процедуру, пока не достигнем необх. точности.

Процедура для проверки, есть ли автокорреляция:

3°. Критерий Дарбина – Уотсона.

"Доктор Ватсон"

Рассмотрим линейную модель:

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

Ошибки модели:

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

H_0: $\rho = 0$ – нет автокорр.

H_1: $\rho > 0$.

Оценим (1) по обычному МНК.

$\Rightarrow \hat{\theta}$,

$\hat{Y} = X \cdot \hat{\theta}$,

$e = Y - \hat{Y}$.

Рассмотрим статистику DW:

$$DW = \frac{\sum_{j=2}^{N} (e_j - e_{j-1})^2}{\sum_{j=1}^{N} e_j^2} \approx 2 \cdot (1 - r),$$ где r – выборочный коэфф. корреляции

При H_0: $DW \sim 2$

H_1: $DW < 2$.

Численными методами удалось доказать, что распред. этой статистики зависит от факторов, входящих в модель.

Сложность в том, что таблица заранее неизвестна $\Rightarrow$ сделать ничего нельзя!

Для заданных α, N, m существуют такие константы:

(α – уров. значимости; N – кол-во измерений; m – кол-во факторов)

$d_e < d_v$, тогда $d_e \le d(\alpha) \le d_v$,

где $d(\alpha)$ – <u>истинная</u> критическая константа.

d_e и d_v – константы, не зависящие от факторов, так что мы можем оценить $d(\alpha)$.

Далее:

1) Если $DW < d_e$, то принимаем H_1

2) Если $DW > d_v$, то H_0 не противоречит экспериментальным данным (автокорреляция не обнаружена)

3) $d_e < d(\alpha) < d_v \Rightarrow$ решения нет

Пример

$N = 100$

$m = 1$

$\alpha = 0,05$

$d_e = 1,65$

$d_{кр} = 1,69.$

Глава 8. Модели дискретного выбора.

06.11

§1. Бинарные модели выбора

1°. Описание модели.

Делаем выбор – 0 или 1.

Y – принятое решение.

$X = (X_1, \dots, X_p)$ – факторы.

Основная задача: найти $\mathbb{P}(Y=1|X)$. $\mathbb{P}(Y=0|X) = 1 - \mathbb{P}(Y=1|X)$

Предположим, что

$$\mathbb{P}(Y=1|X) = F(X\cdot\theta) = F(\theta_1 X_1 + \dots + \theta_p X_p), \quad (1)$$

где $F(z)$ – неубывающая функция со значениями из $[0,1]$.

Далее предполагается, что $F(z)$ – функция распределения.

"А с чего так? А пусть. Модель такая. Далее я способа оправдания, почему это адекватно."

Предположим, что $\exists$ некоторая <u>ненаблюдаемая</u> (латентная) переменная Y^*, которая подчиняется следующей модели:

$$Y^* = X\cdot\theta + \varepsilon, \quad (2)$$

ε – с.в. с ф.р. $F(z)$, которая симметрична отн-но нуля.

Пусть $Y = \begin{cases} 1, & \text{если } Y^* \geq 0, \\ 0, & \text{если } Y^* < 0. \end{cases}$ Можно и $c_0 a$ const, но её всегда можно сузнать в (2)

$$\Rightarrow \mathbb{P}(Y=1|X) = \mathbb{P}(Y^* = X\cdot\theta + \varepsilon \geq 0 \mid X) = \mathbb{P}(\varepsilon \geq -X\cdot\theta) = 1 - F(-X\cdot\theta) = F(X\cdot\theta)$$

Другой подход к объяснению:

Пусть принятие каждого решения имеет полезность u_0 или u_1, причём

$u_0 = X\cdot\theta^{(0)} + \varepsilon_0$

$u_1 = X\cdot\theta^{(1)} + \varepsilon_1$

ε_0 и ε_1 – влияние неучтённых факторов.

Естественно принять то решение, у которого полезность больше. $\Rightarrow$

$$\mathbb{P}(Y=1|X) = \mathbb{P}(u_1 - u_0 \geq 0 \mid X) = \mathbb{P}(\varepsilon_1 - \varepsilon_0 \geq -X\cdot(\theta^{(1)} - \theta^{(0)}))$$

Обозначим $\theta^{(1)} - \theta^{(0)} = \theta$, $\varepsilon_1 - \varepsilon_0 = \varepsilon$ – имеет ф.р. $F(z)$, которая симм. отн-но нуля. $\Rightarrow$ получаем модель (1).

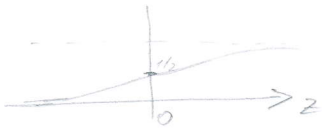
2°. Примеры моделей бинарного выбора

1) Если $F(z) = \Phi(z)$ – ф.р. станд. норм. закона, то имеем Probit-модель.

Не стандартное, но нормальное будится к этому случаю

2) Если $F(z) = \Lambda(z)$ – ф.р станд. логистического распределения, то имеем Logit-модель.

$$\Lambda(z) = \frac{e^z}{1+e^z} = 1 - \frac{1}{1+e^z}$$



3°. Оценивание модели.

МНК здесь неприменимо, 0 и 1 лишь обозначения, не значения.

В этой задаче для оценки параметров используют метод наибольшего правдоподобия:

Пусть имеем N ~~независимых~~ одновр. наблюдений

$y_j, x_{j1}, \dots, x_{jp}$.

Запишем ф-цию правдоподобия:

$$\mathcal{L}(\theta) = \prod_{j=1}^{N} [F(x_j\theta)]^{y_j} [1 - F(x_j\theta)]^{1-y_j} \qquad (3)$$

$$l(\theta) = \log \mathcal{L}(\theta) = \sum_{j=1}^{N} \left(y_j \log F(x_j\theta) + (1-y_j) \log (1 - F(x_j\theta)) \right) \longrightarrow \sup_{\theta} \qquad (4)$$

Теоретический анализ показывает, что в случае Logit-модели ф-ция $l(\theta)$ является глобально вогнутой.

$\Rightarrow$ У этой ф-ции $\exists!$ точка max.

Если ф-ция "хорошая", а эта хорошая $\Rightarrow$ асимптотически норм. оценки.

4°. Выбор модели.

Есть две модели:

- основная с парам. θ_0
- альтернативная, которая имеет некот. число дополнит. параметров (факторов) $\rightarrow \theta_A$.

Надо выбрать ту или иную.

Макфаддин (McFadden) предложил следующую модель детерминации (аналог коэфф-та детерминации):

$$R^2_M := 1 - \frac{l(\hat{\theta}_A)}{l(\hat{\theta}_0)} \qquad (5)$$

Не может быть $=1$, но чем ближе к 1, тем лучше основная. Хотим уметь предсказывать рез-т выбора (в основной модели)! $\Rightarrow$

5°. Задача прогноза.

Все наблюдения разбиваем на 2 части размера N_1 и N_2:

$N_1 + N_2 = N$.

По первым N_1 измер. оцениваем модель и находим оценку $\hat{\theta}$.

Далее полагаем

$$\hat{y}_j = \begin{cases} 1, \text{ если } F(x_j\hat{\theta}) \geq c \\ 0, \text{ в пр. случае} \end{cases}, \quad j = \overline{N_1+1, N_1+N_2}$$

По умолчанию $c = \frac{1}{2}$, если хотим сделать прогноз на какой-то выборке, то смотрим.

Для оценки качества прогноза строим:
Таблица предсказаний и реализаций.

	предсказания $\hat{y}_j = 1$	$\hat{y}_j = 0$	
наблюд. $y_j = 1$	P_{11}	P_{10}	$P_{1\cdot}$
$y_j = 0$	P_{01}	P_{00}	$P_{0\cdot}$
	$P_{\cdot 1}$	$P_{\cdot 0}$	

сумма 1, предсказано 0

сумма по строке

Качество прогноза можно оценить по интенсивности правильных прогнозов = $P_{11} + P_{00}$.

Мера Макфаддена: $F_1 = \dfrac{P_{11} + P_{00} - P_{11}^2 - P_{00}^2}{1 - P_{11}^2 - P_{00}^2}$.

6°. Выбор между двумя марками товара

На рынке есть 2 сорта кетчупа: Heinz и Hunts.
Собраны данные по 2798 покупкам для 300 семей.
Для каждой семьи отмечена последняя покупка ($N_2 = 300$).
На этом полном будем тестировать качество.
Оцениваем модель по 2498 измерениям. В качестве факторов выбираем следующие величины:

1) Лог отношения цен:
$$lp10 = \log\left(\frac{\text{price Heinz}}{\text{price Hunts}}\right)$$

1 – Heinz
0 – Hunts

2) X_{11} – был ли представлен сорт Heinz на витрине.
(фиктивные переменные (0 или 1, dust))
X_{12} – была ли реклама Heinz в печати
X_{13} – были ли оба типа рекламы одновременно.
X_{21} | аналогично для Hunts.
X_{22} |
X_{23} |

Оценим Logit-модель и получим:

Перем	Коэфф	...	p-value
C	1.85		0.0000
X_{11}	0.27		0.0357
X_{12}	0.19		0.2334
X_{13}	0.25		0.3051
X_{21}	−0.37		0.0125
X_{22}	−0.57		0.0036
X_{23}	−1.09		0.0001
$Z=\lg 10$	−3.27		0.0000

X_{12} и X_{13} – незначимы.
X_{22} и X_{23} – значимы.
Почему? Надо ↓ доли сортов на рынке => оказ-ся сорт Heinz в несколько раз > сорта Hunts => более раскрученный товар, про него и так все знают => реклама не принесла новой инф-ции => факторы для Heinz оказались незначимы.

Если выписать таблицу предсказанных / наблюдаемых значений, то получим:
Доля прав. предсказаний = <u>0.77</u>. => хорошие цифры.
У экономистов и объяснение уровня 40% считается уже неплохим.
Логистическая регрессия

§2. Многомерные модели неупорядоченного выбора. 13.11

Как раз модель от Макфаддена.

1°. Описание модели ∃ несколько моделей, ↓ одну.

Пусть необходимо выбрать одно из нескольких решений: $1, 2, \ldots, J$.

Имеется несколько объясняющих факторов: $X_1, X_2, \ldots, X_m$.

Пусть принятие решения j имеет полезность u_{ij} для i-го индивидуума, причём

$$u_{ij} = X_i \cdot \Theta_j + \varepsilon_{ij} = \overline{u_{ij}} + \varepsilon_{ij} \qquad (1)$$

здесь $X_i = (X_{1i}, \ldots, X_{mi})$ и $\Theta_j = (\Theta_{1j}, \ldots, \Theta_{mj})$

ε_{ij} – влияние неучт. факторов, случ. вел.

Принимается то решение, у которого полезность самая большая.

$$\mathbb{P}(Y_i = j \mid X_i) = \mathbb{P}(u_{ij} \geq u_{ik} \ \forall k \mid X_i) = \mathbb{P}(u_{ij} = \max_k u_{ik} \mid X_i) \qquad (2)$$

Надо как-то конкретизировать распределение X_i
В ТВ есть предельные теоремы в схеме максимума (max)

Пусть ε_{ij} – н.о.р.с.в., имеющие распределение экстрем. значений:

$$\mathbb{P}(\varepsilon_{ij} < z) = e^{-e^{-z}}, \quad z \in \mathbb{R}^1.$$

Макфадден показал, что в такой модели

$$\mathbb{P}(Y_i = j \mid X_i) = \frac{e^{\overline{u}_{ij}}}{\sum_k e^{\overline{u}_{ik}}} = \frac{e^{x_i \cdot \theta_j}}{\sum_k e^{x_i \cdot \theta_k}} \quad — \qquad (3)$$

– многомерная Logit-модель

Почему Logit, если тут нет Logit-распределения? Потому что если вер-ти решения 2, то получаются такие же вер-ти, как в Logit-модели.

Замечание.

Рассмотрим 2 решения: j и ℓ и вычтем

$$\frac{\mathbb{P}(Y_i = j \mid X_i)}{\mathbb{P}(Y_i = \ell \mid X_i)} = \frac{e^{x_i \theta_j}}{e^{x_i \cdot \theta_\ell}} \quad — \qquad (4)$$

– независимость от посторонних альтернатив

Есть примеры, которые показывают, что с экономической точки зрения это неправильно (как же так, что на рынке появился 3-й альтернатив, с другими св-вами не влияют на это отношение?) ⇒ модель не очень хорошая это всё из-за предположения н.о.р.с.в.

Хаусман и Со – придумали модель ~3 года назад, где ошибки зависимы и вер-ть тоже есть в явном виде.

2°. Оценивание модели.

Пусть имеем N независимых одновременных измерений Y_i, X_i. Применяем для оценки параметров МНП (метод наиб. правдоподобия). Запишем ф-цию правдоподобия:

$$L(\theta) = \prod_{i=1}^{N} \prod_{j=1}^{J} [\mathbb{P}(Y_i = j \mid X_i)]^{\mathbb{1}(Y_i = j)} \qquad (5)$$

Перейдём к логарифмам:

$$\ell(\theta) = \sum_{i=1}^{N} \sum_{j=1}^{J} \mathbb{1}(Y_i = j) \cdot \log \mathbb{P}(Y_i = j \mid X_i) \to \max_{\theta} \qquad (6)$$

Это очень задача на экстремум численными методами, в явном виде тут плохо.

Для рассм. модели $\ell(\theta)$ оказ-ся глобально вогнутой ⇒ ед. точка max.

Ни в одном пакете эта процедура не реализована (ну, в большинстве вариантов...) ⇒ если не хотите программировать сами :)

3°. Выбор одного из четырёх сортов товара – пример

Есть 4 сорта печенек на рынке: 1, 2, 3, 4.

Собрали данные для $N = 3292$ покупок в 136 семьях.

Что известно (факторы):

- цена купленного товара A_1
- цены на другие товары (для 3-х не выбранных) A_2
- наличие рекламы → фиктивные переменные: A_3

  d_1 – презентация товара B_1

  d_2 – реклама в газете B_2

  d_3 – то и другое вместе B_3

Была проведена оценка и оказалось, что просто презентация товара – незнач., остальные перем. → значим.

Нашли оценки пар-ров, по ф-ле (3) прогнозировали выбор того или иного товара.

Составляем табличку предсказанных и реальных значений, кач-во прогноза оцениваем по % правильно предсказанных значений.

В данной задаче доля правильно предсказанных значений $= 58\%$.

§3. Многомерные модели упорядоченного выбора.

1°. Описание модели

Варианты $1 < 2 < 3 < \ldots < J$. (↖ хуже) ~ оценки на экзамене, $(2 < 3 < 4 < 5)$ оценки

Более простая ситуация, чем в §2. Решается с помощью латентной переменной.

Пусть имеем некоторую латентную (ненаблюдаемую) переменную Y_i^*, которая подчиняется модели:

$$Y_i^* = X_i \cdot \Theta + \varepsilon_i \tag{1}$$

$X_i = (X_{i1}, \ldots, X_{im})$ – измер. факторов,

$\Theta = (\Theta_1, \ldots, \Theta_m)$ – набор параметров.

Предполагается, что ε_i – н.о.р.с.в с ф.р. $F(z) = \mathbb{P}(\varepsilon_i < z)$.

Пусть выбраны некоторые пороги:

$$-\infty = a_0 < a_1 < \ldots < a_{k-1} < a_k < \ldots < a_J = +\infty$$

Предполагаем, что

$$Y_i = k \mid X_i \iff a_{k-1} \le Y_i^* < a_k \tag{2}$$

$$\Rightarrow \mathbb{P}(Y_i = k \mid X_i) = F(a_k - X_i \cdot \Theta) - F(a_{k-1} - X_i \cdot \Theta) \tag{3}$$

Для того, чтобы закончить описание модели, надо в явном виде задать F.

Есть несколько вариантов:
1) Logit - модель
2) Probit - модель
3) модель для экстр. значений

Параметры:

a_k, θ, σ

Нетрудно видеть, что модель не меняется при масштабных преобразованиях. (если все пар-ры * const) ⇒ считаем $\sigma = 1$.

Если в пар-ры входит const, то вычитаем её из $\overset{всех}{d_k}$ ⇒ модель не меняется и при сдвигах. ⇒ ставим $a_1 = 0$ или $a_5 = 0$.

Пусть $a_1 = 0$.

Данные для оценки модели применяется МНП. Здесь и в Logit-модели ответ в явном виде получить не удаётся, поэтому решаются числ. методами (это уже в большинстве пакетов есть.)

2°. Пример.

Опрос о роли женщин в формировании рынка труда.

Вопрос: место женщины дома, а не на работе?

Вар. ответов:
1) категорически не согласна
2) не согласна
3) согласна
4) полностью согласна

Обследование проведено среди 3705 белых женщин из США.

Y - ответ (1, 2, 3, 4).

Факторы:
1) возраст
2) фикт. переменная =1, если до 14 лет семья распалась
3) фикт. перем. = 1, если мать работала, когда ребёнку было 14 лет
4) число лет, затраченных матерью на обучение
5) доход семьи
6) число братьев и сестёр
7) pro = 1, если протестантское воспитание
8) cath = 1, если католическое воспитание
9) $\overset{south}{SO}$ = 1, если из южных штатов
10) urb = 1, если живёт в крупном городе

Оценивалось по МНП в той и другой модели.

8–10 факторов оказались неопознанными. Оказалось, что d_2 и d_3 неизвестны (т.е. ответы "больна" и "скорее больна" не столь различаются, чтобы защитить на "не знаю").

Оценим Θ, вернемся решению, сформулированному.

Глава 9. Дисперсионный анализ

Первым стал рассматривать Фишер.

Классическая литература: Шеффе, "Дисперсионный анализ".

§1. Однофакторный дисперсионный анализ

Имеем следующую модель измерений:

$$Y_{kj} = \mu_k + \varepsilon_{kj} \quad (1)$$

(μ_k – среднее для уровня фактора k)

k – уровень фактора, $k = \overline{1, m}$

j – номер измерения, $j = \overline{1, N}$

$n = m \cdot N$

ε_{kj} – с.в., для которых предполагаем:

1) $\varepsilon_{kj} \sim N(0, \sigma^2) \quad \forall k, j$

2) ε_{kj} – независ. для $\forall k, j$.

Выделяют общее среднее:

$$\mu = \frac{1}{m} \sum_{k=1}^{m} \mu_k$$

и вводят новые переменные:

$$\alpha_k = \mu_k - \mu.$$

$$\Rightarrow \quad Y_{kj} = \mu + \alpha_k + \varepsilon_{kj} \quad (2)$$

Основная задача:

Выяснить, есть ли различия в средних для различных генеральных совокупностей.

Формально проверяем гипотезу

$H_0: \alpha_k = 0 \quad \forall k$

против альтернативы

$H_1: \exists k$, для которого $\alpha_k \neq 0$.

Пример. Y – количество вдыхаемого азота для 4 режимов питания.

Применим МНК, т.е. рассмотрим следующую экстремальную задачу:

$$Q = \sum_k \sum_j (Y_{kj} - \mu - \alpha_k)^2 \to \min_{\mu, \alpha_k} \quad (3)$$

$$\Rightarrow \hat{\mu} = \overline{Y}_{..} = \frac{1}{n} \sum_k \sum_j Y_{kj} \quad (4)$$

$$\hat{\alpha}_k = \overline{Y_{k\cdot}} = \frac{1}{N}\sum_j Y_{kj} \qquad (5)$$

Справедливо следующее соотношение

$$SS_T = SS_B + SS_R \qquad (6)$$

Total Between Residual

$SS_T = \sum_k \sum_j (Y_{kj} - \overline{Y_{\cdot\cdot}})^2$ – полная сумма квадратов

$SS_B = \sum_{k=1}^{m} (\overline{Y_{k\cdot}} - \overline{Y_{\cdot\cdot}})^2 \cdot N$ – разброс между группами

$SS_R = \sum_k \sum_j (Y_{kj} - \overline{Y_{k\cdot}})^2$ – разброс внутри группы

При верной Но SS_B должен быть мал на фоне SS_R.

Теорема При верной Но

1) с.в. SS_B и SS_R – независимы и имеют χ^2-распределение с $(m-1)$ и $n-m = (N-1)\cdot m$ степенями свободы

2) с.в. $F = \dfrac{SS_B/(m-1)}{SS_R/(n-m)}$ имеет распределение Снедекора-Фишера с $((m-1),(n-m))$ степенями свободы

Далее проверка гипотезы проходит по стандартной схеме (сравниваем значение стат-ки с крит. const, ...)

Другая постановка задачи, в которой

$\alpha \sim N(0, \sigma_0^2)$

Тогда задача о влиянии фактора сводится к проверке

$H_0: \sigma_0^2 = 0$ против

$H_1: \sigma_0^2 > 0$.

Пример Изучается норма прибыли на обыкновенную акцию для 3-х ~~факторов~~ групп: сталелитейной, судостроительной, энергетич.:

Y_1, Y_2, Y_3

$m = 3$

$N = 9$.

Получен след. рез-ты для средних

$\overline{Y_1} = 9.78$, $\overline{Y_2} = 18.93$, $\overline{Y_3} = 13.1$

$\overline{Y_{\cdot\cdot}} = 13.93$

$SS_B = 128.73$, $SS_R = 496.62$

Стат. F имеет распр. Снедекора – Фишера с (2,24) степ. своб.

$F^* = 3.11$

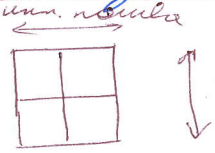
По таблицам достигнутый уровень значимости равен:

$p\text{-value}_F = 0.063$

$\Rightarrow$ Даже на 5% уровне нет <u>значимого</u> различия.

§2. Двухфакторный дисперсионный анализ

Пример: мясо — с травой



(Далее 5 схем одноф., где контроль для каждого уровня 1 измерение.)

Имеем следующую схему измерений:

$$Y_{kj} = \mu_{kj} + \varepsilon_{kj} \quad (7)$$

k - уровень первого фактора

j - уровень второго фактора.

Предположения:

1) ε_{kj} – нез. $\forall k, j$

2) $\varepsilon_{kj} \sim N(0, \sigma^2) \quad \forall k, j$

$k = \overline{1, m}$, $j = \overline{1, N}$, $n = m \cdot N$

Введём следующие величины:

$$\mu_{\cdot\cdot} = \frac{1}{n} \sum_{k,j} \mu_{kj} \quad (8)$$

$$\mu_{\cdot j} = \frac{1}{m} \sum_{k} \mu_{kj} \quad (9)$$

$$\mu_{k\cdot} = \frac{1}{N} \sum_{j} \mu_{kj} \quad (10)$$

$\tau_j = \mu_{\cdot j} - \mu_{\cdot\cdot}$

$\rho_k = \mu_{k\cdot} - \mu_{\cdot\cdot}$

τ_j и ρ_k – <u>эффекты строки и столбца</u> соотв-но.

<u>Предполагается (!)</u>, что

$$\mu_{kj} = \mu_{\cdot\cdot} + \tau_j + \rho_k, \quad (11)$$

т.е. предполагается отсутствие взаимодействия между эффектом строки и эффектом столбца.

Запишем модель (7) с учётом (11).

Для оценки модели применим МНК.

$$\Rightarrow \hat{\mu}_{..} = \overline{Y}_{..} = \frac{1}{n}\sum_{k,j} Y_{kj}$$

с точки зрения теор. вер-ва с.в. разложили в сумму независимых ортогональных (в смысле скал. произв.) величин

$$\hat{\rho}_k = \frac{1}{N}\sum_j (Y_{kj} - \overline{Y}_{.j})$$

$$\hat{\tau}_j = \frac{1}{m}\sum_k (Y_{kj} - \overline{Y}_{k.})$$

$$\Rightarrow Y_{kj} = \overline{Y}_{..} + (\overline{Y}_{k.} - \overline{Y}_{..}) + (\overline{Y}_{.j} - \overline{Y}_{..}) + e_{kj} \qquad (12)$$

Можно показать, что это нез-мые гаусс. величины.

Обозначим

$S_1^2 = \sum_k \sum_j e_{kj}^2$ – сумма квадратов остатков (13)

В двухфакторном анализе есть 2 постановки задачи:

Первая – есть ли разница в средних по строкам?

Формально проверяем

$H_0: \rho_1 = \ldots = \rho_m = 0$ (14)

против альтернативы

$H_1: \exists k: \rho_k \neq 0.$

$\Rightarrow$ частный случай линейной гипотезы, проверяется через F-критерий.

Пример. Изучается %, выплачиваемый по срочным вкладам в 3-х фед. округах в различных по размеру банках.

Факторы: округ, размер банка. $N=3, m=5.$

Получены след. рез-таты

Округ / Размер	Нью-Йорк	Атланта	Сан-Франциско	Среднее по строкам
<10	4.65	4.55	4.44	4.547
10-50	4.76	4.40	4.99	4.717
50-100	4.60	4.92	5.00	4.840
100-500	4.83	4.70	4.97	4.833
>500	4.99	4.46	4.93	4.793
среднее по столб.	4.766	4.606	4.866	
Общее среднее				4.746

Проведем гипотезу

$H_0: \rho_1 = \dots = \rho_N = 0$

против

$H_1: \exists k: \rho_k \neq 0.$

Наблюдаемое значение статистики Фишера $F^* = 1.11$

Число степеней свободы (2, 8).

Достигнутый уровень значимости

p-value = 0.58

Как указать, в ППП двухфакторный или однофакторный анализ?

one-way

two-way

§3. Учёт эффекта взаимодействия.

27.11

$$Y_{kjt} = \mu_{kj} + \varepsilon_{kjt} \quad (1)$$

$k = \overline{1,m}, \quad j = \overline{1,N}, \quad t = \overline{1,T}, \quad n := m \cdot N \cdot T$

$$\mu_{kj} = \mu_{..} + \rho_k + \tau_j + I_{kj} \quad (2)$$

I_{kj} – оценивает эффект взаимодействия факторов

Оценивая модель по МНК. $\Rightarrow$

$$\hat{\mu}_{..} = \frac{1}{n} \sum_{k,j,t} Y_{kjt} = \overline{Y}_{..}$$

$$\hat{\rho}_k = \frac{1}{N \cdot T} \sum_{j,t} Y_{kjt} - \overline{Y}_{..} = \overline{Y}_{k.} - \overline{Y}_{..}$$

$$\hat{\tau}_j = \frac{1}{m \cdot T} \sum_{k,t} Y_{kjt} - \overline{Y}_{..} = \overline{Y}_{.j} - \overline{Y}_{..}$$

$$\hat{I}_{kj} = \frac{1}{T} \sum_{t} Y_{kjt} - \overline{Y}_{k.} - \overline{Y}_{.j} - \overline{Y}_{..} = \overline{Y}_{kj} - \overline{Y}_{k.} - \overline{Y}_{.j} - \overline{Y}_{..}$$

$$\Rightarrow \quad Y_{kjt} = \overline{Y}_{..} + (\overline{Y}_{k.} - \overline{Y}_{..}) + (\overline{Y}_{.j} - \overline{Y}_{..}) + (\overline{Y}_{kj} - \overline{Y}_{k.} - \overline{Y}_{.j} - \overline{Y}_{..}) + e_{kjt} \quad (3)$$

Можно показать, что это разложение на ортог. компоненты.

Обозначим

$$S_1^2 = \sum_k \sum_j \sum_t e_{kjt}^2$$ – полная сумма квадратов

S_1^2 после оценки параметров имеет $n - m \cdot N$ степ. свободы.

В такой модели гипотезы об эффектах столбцов и строк проверяются так же, как до этого.

Гипотеза о взаимодействии:

Формально проверяем

$H_0: I_{kj} = 0 \quad \forall (k,j)$

против

$H_1: I_{kj} \neq 0$ для некоторой пары $(k_0 j_0)$

Если учесть огран-ия, задаваемые гипотезой H_0 и применить МНК, то получится, что сумма квадр. остатков

$$S_I^2 = T \cdot \sum_k \sum_j \left(\bar{Y}_{kj} - \bar{Y}_{k\cdot} - \bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot}\right)^2$$ - имеет $(m-1)\cdot(N-1)$ степ. свободы. (4)

Теорема

При верной H_0 статистика

$$F = \frac{S_I^2 / (m-1)(N-1)}{S_1^2 / (n - m \cdot N)} \qquad (5)$$

имеет распределение Снедекора – Фишера с $\left((m-1)(N-1), (n - m \cdot N)\right)$ степ. св.

Далее гипотеза H_0 проверяется по стандартной схеме.

(Реализовано в базовой части ППП)

18.12 – последняя лекция

Глава 10. Дискриминантный анализ (классы заранее знаем)

Это некоторый статистич. метод, который позволяет изучать различия между 2 и более группами (классами) по одновременным измерениям нескольких переменных (характеристик).

Примеры

1) тестирование при приёме на работу
2) анализ переписи населения
3) изучение эффекта при анализе метода лечения
4) разделение студентов по уровню знаний

Возникают 2 задачи:

I. Интерпретация - можно ли по имеющемуся набору хар-к провести разделение на предложенные классы

II. Классификация - надо предъявить нек. кол-во функций от характеристик, по которым по определённому правилу будем относить к тому или иному классу.

Всюду далее предполагаем, что разделить можно, и будем решать задачу II.

§1. Постановка задачи

Пусть каждый объект характеризуется набором хар-к $(X_1, \dots, X_p) = X$. X имеет многомерное нормальное распределение с вектором средних $\mu = (\mu_1, \dots, \mu_p)$ и матрицей ковар. $\Sigma = (\sigma_{ij})$.

Пусть классов 2 (всюду далее). =>

$\mu^{(1)} = (\mu_{11}, \dots, \mu_{1p})$

$\mu^{(2)} = (\mu_{21}, \dots, \mu_{2p})$

и одна и та же матрица ковар.

Рассмотрим линейную функцию

$$Z = \alpha_1 \cdot X_1 + \dots + \alpha_p \cdot X_p \qquad (1)$$

Все решающие правила будут иметь следующий вид:

Правило решения:

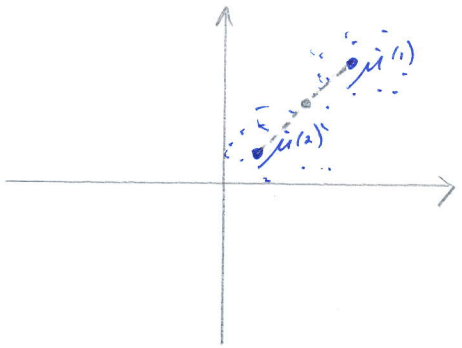
1) Если $Z \geq c$, то относим объект к первой совокупности
2) Если $Z < c$, то ко второй совокупности

$\alpha_1, \dots, \alpha_p, c$ - набор параметров решающего правила.

Хотим выбрать такое правило, которое минимизирует вер-ть ошибочной классификации. ← постановка задачи

§2. Классификация в случае известных параметров.

Нам явно заданы $\mu^{(1)}$, $\mu^{(2)}$ и Σ.



$W^{(1)}$ – первая совокупность

$W^{(2)}$ – вторая

Если X из $W^{(1)}$, то с.в. Z имеет одномерное нормальное распределение со средним

$$m_1 = \alpha_1 \cdot \mu_{11} + \dots + \alpha_p \cdot \mu_{1p} \quad (2)$$

и дисперсией

$$\sigma_z^2 = \sum_{i,j=1}^{p} \alpha_i \cdot \sigma_{ij} \cdot \alpha_j \quad (3)$$

Аналогично, если X из $W^{(2)}$, то с.в. Z имеет одномерное нормальное распределение со средним

$$m_2 = \alpha_1 \cdot \mu_{21} + \dots + \alpha_p \cdot \mu_{2p} \quad (4)$$

и той же дисперсией σ_z^2.

На интуитивном уровне понятно, что для оптимального правила желательно развести $\mu^{(1)}$ и $\mu^{(2)}$ как можно дальше, но с учётом дисперсии σ_z^2.

Этой задачей занимался индиец Махаланобис (Mahalanobis). Он предложил расст. стержневую величину

$$\Delta^2 = \frac{(m_1 - m_2)^2}{\sigma_z^2} \quad (5)$$

– расстояние Махаланобиса

Рональд Фишер (в 1936г) показал, что максимум Δ^2 достигается на наборе $\alpha_1, \dots, \alpha_p$, который есть решение системы

$$\alpha_1 \cdot \sigma_{i1} + \dots + \alpha_p \cdot \sigma_{ip} = \mu_{1i} - \mu_{2i}, \quad \leftarrow \text{СЛАУ} \quad (6)$$

$$i = \overline{1, p}$$

Далее ищем $c = ?$ Обозначим

$\mathbb{P}(1|2)$ – вероятность отнести объект к $W^{(1)}$, когда он из $W^{(2)}$,

$\mathbb{P}(2|1)$ – -//- к $W^{(2)}$, когда он из $W^{(1)}$.

Из МС знаем, что минимизировать одновременно обе вер-ти не удаётся.

Зафиксируем $\mathbb{P}(1|2) + \mathbb{P}(2|1) \to \min\limits_{c}$. Можно показать, что min достигается при

$$c = \frac{m_1 + m_2}{2}$$

Это эвристическое решение.

Строгое решение этой задачи в байесовской постановке ранней Вальд. (Когда заранее знаем частоты)

Пусть q_1 – априорная вер-ть события $X \sim W^{(1)}$.

q_2 – ... $X \sim W^{(2)}$.

Пусть X имеет плотности распределения $f_1(x)$ и $f_2(x)$ для $W^{(1)}$ и $W^{(2)}$.

Вычислим по формуле Байеса апостериорные вероятности:

$$\mathbb{P}(W^{(i)} \mid x) = \frac{q_i \cdot f_i(x)}{q_1 \cdot f_1(x) + q_2 \cdot f_2(x)} \tag{7}$$

Будем выбирать ту совокупность, для которой апостериорная вер-ть больше. Получаем следующий результат:

Правило решения:

1) Принимаем $W^{(1)}$, если $\frac{q_1 f_1(x)}{q_2 f_2(x)} \geq 1$ (8)

2) Принимаем $W^{(2)}$, если $\frac{q_1 f_1(x)}{q_2 f_2(x)} < 1$ (9)

Из общей теории байесовских правил следует, что такое правило минимизирует ожидаемую вер-ть ошибочной классификации.

$$q_1 \cdot \mathbb{P}(2|1) + q_2 \cdot \mathbb{P}(1|2) \tag{10}$$

Для нормальных распред. из (8) и (9) $\Rightarrow$

1) Принимаем $W^{(1)}$, если $\sum_i a_i x_i \geq \frac{m_1 + m_2}{2} + \ln\left(\frac{q_2}{q_1}\right)$ (11)

2) Принимаем $W^{(2)}$, если $\sum_i a_i x_i < \frac{m_1 + m_2}{2} + \ln\left(\frac{q_2}{q_1}\right)$ (12)

Если $q_1 = q_2 \Rightarrow$ получаем эмпирическое правило.

Введём две положит. величины: 04.12.

$c(2|1)$ – сложность потерь, если принято $W^{(2)}$, а на самом деле имеем $W^{(1)}$

$c(1|2)$ – ... аналогично

Ищем процедуру, которая минимизирует ожидаемую стоимость потерь, т.е.

$$c(2|1) \cdot q_1 \cdot \mathbb{P}(2|1) + c(1|2) \cdot q_2 \cdot \mathbb{P}(1|2)$$

Приходим к след. процедуре:

1) Принимаем $W^{(1)}$, если

$$\sum_{i=1}^{p} a_i \cdot x_i \geq \frac{m_1 + m_2}{2} + \ln\left(\frac{q_2 \cdot c(1|2)}{q_1 \cdot c(2|1)}\right)$$

2) Принимаем $W^{(2)}$, если

$$\sum_i \lambda_i \cdot x_i < \frac{m_1+m_2}{2} + \ln\left(\frac{q_2 \cdot C(1|2)}{q_1 \cdot C(2|1)}\right)$$

В случае нормального распределения вероятности ошибочной классификации имеют вид:

$$P(2|1) = \Phi\left(\frac{K - \frac{1}{2}\Delta^2}{\Delta}\right)$$

и

$$P(1|2) = \Phi\left(\frac{-K - \frac{1}{2}\Delta^2}{\Delta}\right),$$

$$K = \ln \frac{q_2 \cdot C(1|2)}{q_1 \cdot C(2|1)},$$

Δ^2 - расстояние Махаланобиса

При $c(1|2) = C(2|1)$ возвращаемся к байесовской постановке задачи.

Если (при этом) $q_1 = q_2$, то получаем эвристическую процедуру $\Rightarrow$

$$P(2|1) = P(1|2) = \Phi\left(-\frac{\Delta}{2}\right)$$

Условный пример

↑ "потому что ни к чему отношения не имеет"

$X = (X_1, X_2)$ - двумерн. норм. распред. с вектором средних $\mu_1 = (1,1)$, $\mu_2 = (3,3)$, $\Sigma = \begin{pmatrix} 1 & 1 \\ 1 & 1{,}25 \end{pmatrix}$

В случае эвристической процедуры:

$$\begin{cases} \lambda_1 + \lambda_2 = -2 \\ \lambda_1 + 1.25 \cdot \lambda_2 = -2 \end{cases} \Rightarrow \begin{cases} \lambda_2 = 0 \\ \lambda_1 = -2 \end{cases} \Rightarrow Z = -2 \cdot X_1$$ - дискриминантная функция.

$\Rightarrow m_1 = -2$, $m_2 = -6$, $\sigma_z^2 = 4$

$$c = \frac{m_1+m_2}{2} = -4.$$

Правило классификации имеет вид:

1) принимаем $W^{(1)}$, если

$-2 x_1 x_1 \geq -4 \iff x_1 \leq 2$

2) принимаем $W^{(2)}$, если

$x_1 > 2$.

Очередной объект имеет набор характеристик $x = (2.5, 1)$.

$\Rightarrow 2.5 > 2 \Rightarrow$ наблюдение из $W^{(2)}$.

← относим к

§3. Классификация в случае неизвестных параметров

Пусть μ_1, μ_2, Σ – неизвестные!

Заданы $q_1, q_2, c(2|1), c(1|2)$.

Пусть заданы две обучающих выборки

$X_{11}, \dots, X_{1n_1}$ и $X_{21}, \dots, X_{2n_2}$,

где $X_{ij} = (X_{ij}^{(1)}, \dots, X_{ij}^{(p)})^T$ – независ. и имеют многомер. норм. распред. с параметрами μ_i, Σ

Находим оценки:

$$\hat{\mu}_1 = \overline{X_1} = \frac{1}{n_1} \sum_j X_{1j}$$

$$\hat{\mu}_2 = \overline{X_2} = \frac{1}{n_2} \sum_j X_{2j}$$

$$\hat{\Sigma} = S = \frac{1}{n_1 + n_2 - 2} [(n_1 - 1) \cdot S_1 + (n_2 - 1) S_2]$$

← для несмещённости

где $S_i = \frac{1}{n_i - 1} \sum_j (X_{ij} - \overline{X_i})(X_{ij} - \overline{X_i})^T$, $\quad i = 1, 2$

(В случае многом. норм. распред. это несмещ. и состоят. оценки.)

Далее применяем обобщённую байесовскую процедуру, где настоящие параметры заменим на их оценки.

Алгоритм решения задачи классификации.

1) Решаем систему уравнений

$$\alpha_1 \cdot S_{i1} + \dots + \alpha_p \cdot S_{ip} = \hat{\mu}_{1i} - \hat{\mu}_{2i}, \quad i = \overline{1, p}$$

$\Rightarrow$ находим $\hat{\alpha}_1, \dots, \hat{\alpha}_p$.

2) Оцениваем дискриминантную функцию:

$$\hat{Z} = \hat{\alpha}_1 \cdot X_1 + \dots + \hat{\alpha}_p \cdot X_p$$

3) по обучающим выборкам вычисляем оценки для дискриминантных функций

$\hat{Z}_{1j}, \hat{Z}_{2j}$.

4) Заменяем m_1 и m_2 на их оценки

$$\hat{m}_1 = \frac{1}{n_1} \cdot \sum_{j=1}^{n_1} \hat{Z}_{1j},$$

$$\hat{m}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} \hat{Z}_{2j}$$

5) Оцениваем σ_z^2:

$$\hat{\sigma}_z^2 = S_z^2 = \sum_{i,k=1}^{p} \hat{\alpha}_i \cdot S_{ik} \cdot \hat{\alpha}_k$$

6) Имеем новый набор признаков $x = (x_1, \dots, x_p)$

а) относим его к $w^{(1)}$, если

$$\hat{Z} = \sum_{i=1}^{p} \hat{\alpha}_i x_i > \frac{\hat{m}_1 + \hat{m}_2}{2} + \ln\left(\frac{q_2 \cdot C(1|2)}{q_1 \cdot C(2|1)}\right).$$

Если n_1 и n_2 больше, то "все хорошо" = алгоритм приближается к оптимальному решению

Глава 11. Кластерный анализ (не знаем число классов)

§1. Описание данных

Из некоторого множества (генеральной совокупности) отобрано n объектов $I = (I_1, \dots, I_n)$.

У каждого объекта измерен некоторый характеристик

$C = (c_1, \dots, c_p)^T$ – это количественные характеристики.

$\Rightarrow x_{ij}$ – измерение i-ой характеристики у j-ого объекта.

$X_j = (x_{1j}, \dots, x_{pj})^T$ – все характеристики у объекта I_j

$\Rightarrow X = (X_1, \dots, X_n)$ – матрица измерений.

§2. Задача кластерного анализа.

Пусть $m < n$. (Даже скорее $m << n$)

Требуется на основе измерений X разбить множество объектов I на m классов (кластеров) $\pi_1, \dots, \pi_m$ так, чтобы:

1) каждый объект I_j принадлежит только одному классу
2) объекты внутри одного класса в некотором смысле близки (сходные),
3) объекты из разных классов далеки (несходные).

§3. Меры сходства.

Предположим, что в $\mathbb{R}^p$ задано некоторое расстояние (метрика ?). часто метрика в соответствии с нашим определением метрики, но не всегда!

1) Евклидово расстояние

$$d_2(X_k, X_j) := \left[\sum_{i=1}^{p} (X_{ik} - X_{ij})^2 \right]^{1/2}$$

2) l_1 – норма

$$d_1(X_k, X_j) = \sum_{i=1}^{p} |X_{ik} - X_{ij}|$$

3) максимальная норма (получается при $p \to \infty$ из l_p – норм)

$$d_\infty(X_k, X_j) = \sup_i |X_{ik} - X_{ij}|$$

4) Расстояние Махаланобиса

$$D^2(X_k, X_j) = (X_k - X_j)^T \cdot W^{-1} \cdot (X_k - X_j),$$

где W – матрица рассеяния (обычно задана в задаче)

Замечание. Пусть $Y = B \cdot X$. Тогда

$D^2(Y_k, Y_j) = D^2(X_k, X_j)$, т.е. расст. Махал. не меняется при однородных линейных преобразованиях.

11.12.

$D = (d_{kj})$ - матрица попарных расстояний

Опр. Неотрицательная функция $S(X_k, X_j) := s_{kj}$ называется мерой сходства, если

1) $0 \le S(X_k, X_j) < 1$, если $X_k \neq X_j$

2) $S(X_k, X_k) = 1$

3) $S(X_k, X_j) = S(X_j, X_k)$

Замечания

1) s_{kj} - мера сходства признаков X_k и X_j

2) Если X состоит из 0 и 1, то s_{kj} наз-ся "коэфф. ассоциации" или "парный коэфф. сопряжённости".

3) В статистике очень часто в качестве s_{kj} используют модули коэфф-та корреляции

подразумевается, что задано какое-то расстояние

Опр. Пусть X - матрица признаков. Величина

$$Sd := \frac{1}{2} \sum_{k,j=1}^{n} d(X_k, X_j)$$ - общее расстояние м-ва объектов I.

Опр. $\overline{Sd} := \frac{Sd}{Nd}$ - среднее расстояние,

$Nd = \frac{n(n-1)}{2}$ - число всех пар, которые мы сравнивали

Опр. $Sx := \sum_{j=1}^{n} (X_j - \overline{X})(X_j - \overline{X})^T$ - матрица рассеяния, где (1)

$\overline{X} = \frac{1}{n} \sum_{j=1}^{n} X_j$.

Опр. Величина

$S_t := tr\, S_x = \sum_{j=1}^{n} \sum_{i=1}^{p} (X_{ij} - \overline{X}_i)^2$ - статистическое рассеяние

(trace)

Опр. Величина

$\det(S_x)$ - статистическое рассеяние, соответствующее определителю

§4. Расстояние между кластерами

Пусть имеем два набора объектов $I = (I_1, \dots, I_{n_1})$ и $J = (J_1, \dots, J_{n_2})$ $\Rightarrow$ $X = (X_1, \dots, X_{n_1})$ и $Y = (Y_1, \dots, Y_{n_2})$.

Опр. Величина

$$D_1(I, J) := \min_{k,j} d(X_k, Y_j)$$ – минимальное локальное расстояние.

Опр. Величина

$$D_2(I, J) := \max_{k,j} d(X_k, Y_j)$$ – максимальное локальное расстояние

Опр. $D_3(I, J) := \sum_k \sum_j d(X_k, X_j) \Big/ n_1 \cdot n_2$ – среднее расстояние между кластерами

Опр. $D_4(I, J) := (\bar{X} - \bar{Y})^T (\bar{X} - \bar{Y}) \Big/ \frac{n_1 \cdot n_2}{n_1 + n_2}$ – статистическое расстояние между кластерами

$\frac{1}{h} = \frac{1}{n_1} + \frac{1}{n_2}$ $\Rightarrow$ ср. гармонич.

квадрат расстояния между усл. нами, по сути

Рассмотрим новый кластер $K = I \cup J$.

Образуем

$$S_K = \sum_{k=1}^{n_1} (X_k - M)(X_k - M)^T + \sum_{j=1}^{n_2} (Y_j - M)(Y_j - M)^T,$$

где $M = \frac{\sum_k X_k + \sum_j Y_j}{n_1 + n_2}$.

Можно показать, что

$$S_K = S_I + S_J + \frac{n_1 \cdot n_2}{n_1 + n_2} (\bar{X} - \bar{Y})(\bar{X} - \bar{Y})^T. \quad (*)$$

Опр. Величина $\frac{n_1 \cdot n_2}{n_1 + n_2} (\bar{X} - \bar{Y})(\bar{X} - \bar{Y})^T$ – матрица межгруппового рассеяния.

$\Rightarrow$ $tr \frac{n_1 \cdot n_2}{n_1 + n_2} (\bar{X} - \bar{Y})(\bar{X} - \bar{Y})^T = \frac{n_1 \cdot n_2}{n_1 + n_2} (\bar{X} - \bar{Y})^T \cdot (\bar{X} - \bar{Y})$ – статистическое

матрица число

расстояние между I и J = межгрупповая сумма квадратов

§5. Последовательное построение кластеров.

Описание общего алгоритма:

1) Сначала все точки рассматриваем как отдельные кластеры.
2) Выбираем два порога: $0 < t < s < \infty$.
3) Если все расстояния меньше s, то всё объединяем в один кластер и останавливаемся.
4) Если есть два кластера, расстояние между которыми меньше s, то объединяем их в один.
5) Пересчитываем расстояния внутри кластеров и между ними.
6) Процедура продолжается до тех пор, пока расстояния внутри классов не более t, а между классами не более s.

Досрок 25.12. в 10.30 (у кафедры, аудиторию ищем сами)

Это основная попытка.

Но не помню про 15-е число, а то нам нельзя проводить досрок!

Основная учебно-методическая литература

1. Андерсон Т. Введение в многомерный статистический анализ. – М.: Наука, 1963.
2. Johnson R.A. and Winchern D.W. Applied Multivariate Statistical Analysis. – Pearson Prentice Hall, 2007.
3. Хохлов Ю.С. Эконометрика. Вводный курс: Учебное пособие. М.: Издательский отдел факультета ВМиК МГУ, 2006. – 100 с.

Дополнительная учебно-методическая литература

1) Факторный, дискриминантный и кластерный анализы. – М.: Финансы и статистика, 1989.
2) Шеффе Г. Дисперсионный анализ. – М.: Наука, 1980.
3) Болч К., Хуань К.Дж. Многомерные статистические методы для экономики. – М.: Статистика, 1979.
4) Hardle W., Simar L. Applied Multivariate Statistical Analysis. - http://www.xplorestat.de/ebooks/ebooks.html
5) Кулаичев А.П. Методы и средства комплексного анализа данных. – М.: Форум-Инфра-М, 2006.