Конспект билетов по теории

10 декабря 2025 г.

Содержание

1 Корреляция Пирсона, доверительный интервал и тест; Спирмен, Кендалл

Пирсон — сила *линейной* связи для количественных признаков при приблизительной нормальности и эллиптическом облаке точек; чувствителен к выбросам.

$$r = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{S_X S_Y},$$

Где S_X^2 и S_Y^2 — смещённые оценки дисперсий.

Спирмен – позволяет уйти от условия линейности зависимости, заменив наблюдения X_i на их ранги R_i в ряду X, а Y_i на ранги T_i в ряду Y.

$$\rho_S = \frac{\overline{RT} - \overline{RT}}{\sqrt{S_R^2 S_T^2}},$$

Кендалл – мера монотонной зависимости между двумя переменными. Основана на подсчёте согласованных и несогласованных пар наблюдений.

$$\tau = \frac{C - D}{C + D} = \frac{2(C - D)}{n(n-1)},$$

где C — число согласованных пар, D — число несогласованных пар.

Доверительный интервал и тесты для корреляции. Для Пирсона при $H_0 : \rho = 0$ используют t-критерий

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2},$$

по которому строят двусторонний тест и доверительный интервал по схеме "оценка $\pm t_{1-\alpha/2} \cdot \mathrm{se}(r)$".

Удобно также применять преобразование Фишера

$$z = \tfrac{1}{2} \ln \frac{1+r}{1-r} = \mathrm{artanh}(r),$$

для которого при большом n выполняется

$$z \approx \mathcal{N}\Big(\mathrm{artanh}(\rho), \frac{1}{n-3}\Big),$$

а ДИ для ρ получают обратным преобразованием через tanh.

Для Спирмена и Кендалла используют либо перестановочные тесты, либо асимптотическое приближение

$$\frac{\hat{\rho}}{\mathrm{se}(\hat{\rho})} \approx \mathcal{N}(0, 1),$$

и дальше ту же стандартную схему "оценка $\pm$ квантиль $\times$ стандартная ошибка".

2 Корреляционная матрица; мультиколлинеарность; наведённая зависимость; частные корреляции; конкордация

Корреляционная матрица. Симметрична, положительно полуопределённа; диагональ равна 1. Большие по модулю внедиагональные элементы — индикатор взаимосвязи признаков.

Мультиколлинеарность.

- Строгая коллинеарность: существует $v \neq 0$ с $Xv = 0 \Rightarrow X^\top X$ вырождена, оценки МНК неопределимы и зависимые предикторы надо удалить.
- Почти коллинеарность: $X^\top X$ плохо обусловлена $\Rightarrow$ дисперсии $\hat{\beta}$ раздуваются, знаки/величины неустойчивы, интерпретация ломается.
- Мультиколлинеарность — наличие сильной (почти линейной) зависимости между несколькими предикторами, когда один или несколько столбцов X хорошо аппроксимируются линейной комбинацией остальных.

Наведённая зависимость (конфаундинг). Связь X и Y может объясняться общим фактором Z. Тогда маргинальная корреляция вводит в заблуждение; контролируем Z и используем частные меры.

Частная корреляция. Очистка влияния Z:

$$\rho_{XY\cdot Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{(1-\rho_{XZ}^2)(1-\rho_{YZ}^2)}}.$$

Конкордация (согласованность ранжировок). Коэффициент конкордации Кендалла W агрегирует согласие нескольких ранжировок ("экспертов"). Для m ранжировок n объектов одна из формул:

$$W = \frac{12}{m^2(n^3-n)} \sum_{i=1}^{n} \left(\sum_{j=1}^{m} R_{ij} - \frac{m(n+1)}{2} \right)^2,$$

где R_{ij} — ранг i-го объекта у j-го эксперта. $W \in [0,1]$; большие значения означают сильное согласие. Также полезны условные/частные корреляции для интерпретации причинных структур.

3 Повторные выборки; критерий знаков; Уилкоксона; ANOVA

Повторные (парные) выборки. Сравниваем одну и ту же единицу до/после: анализируем разности $d_i = y_i^{(2)} - y_i^{(1)}$.

Критерий знаков. Тест медианы разностей $\mathrm{Med}(d_i) = 0$. Статистика — число положительных знаков $B \sim \mathrm{Bin}(m, 1/2)$, где m — число ненулевых d_i. При больших m — нормальная аппроксимация.

Уилкоксона (ранговых знаков). Ранжируем $|d_i|$, присваиваем знак, суммируем знаковые ранги W. При $n \gtrsim 10$ используется нормальная аппроксимация с поправкой на связи, для малых n — точные таблицы. Более мощен, чем знаковый, при симметричных распределениях разностей.

Однофакторная ANOVA. Гипотеза $H_0 : \mu_1 = \cdots = \mu_g$. Разложение вариации: $TSS = SS_B + SS_W$. Статистика

$$F = \frac{SS_B/(g-1)}{SS_W/(n-g)} \sim F_{g-1,\,n-g} \text{ при } H_0.$$

Предпосылки: нормальность в группах, гомоскедастичность, независимость. Пост-хок сравнения: Tukey HSD, Bonferroni.

4 МНК из ММП; нормальные уравнения

Линейная модель. $y = X\beta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$.

МНК как ММП. Лог-правдоподобие:

$$\ell(\beta, \sigma^2) = -\tfrac{n}{2}\ln(2\pi\sigma^2) - \tfrac{1}{2\sigma^2}\|y - X\beta\|^2.$$

Максимизация по β эквивалентна $\min \|y - X\beta\|^2$.

Нормальные уравнения и оценки. Нормальные уравнения: $X^\top X\,\hat\beta = X^\top y$, решение при $\operatorname{rank}(X) = k+1$:

$$\hat\beta = (X^\top X)^{-1}X^\top y, \qquad \hat\sigma^2 = RSS/n, \qquad RSS = \|y - X\hat\beta\|^2.$$

Распределение: $\hat\beta \sim \mathcal{N}(\beta, \sigma^2(X^\top X)^{-1})$, а $\frac{RSS}{\sigma^2} \sim \chi^2_{n-k-1}$.

5 Простая линейная регрессия; оценки, ДИ; Гаусс–Марков

Оценки коэффициентов.

$$\hat b_1 = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - \overline{X}^{\,2}}, \qquad \hat b_0 = \overline{Y} - \hat b_1 \overline{X}.$$

Дисперсии и доверительные интервалы.

$$\widehat{\mathrm{Var}}(\hat b_1) = \hat\sigma^2/S_{xx}, \qquad \widehat{\mathrm{Var}}(\hat b_0) = \hat\sigma^2\Big(\tfrac{1}{n} + \tfrac{\overline{X}^2}{S_{xx}}\Big),$$

где $S_{xx} = \sum(x_i - \overline{X})^2$. ДИ для среднего отклика и предсказания в x_0:

$$\hat y_0 \pm t_{\alpha/2,\,n-2}\,\hat\sigma\sqrt{\tfrac{1}{n} + \tfrac{(x_0-\overline{X})^2}{S_{xx}}}, \qquad \hat y_0 \pm t_{\alpha/2,\,n-2}\,\hat\sigma\sqrt{1 + \tfrac{1}{n} + \tfrac{(x_0-\overline{X})^2}{S_{xx}}}.$$

Теорема Гаусса–Маркова. При линейности, экзогенности и гомоскедастичности МНК является BLUE (лучшей линейной несмещённой оценкой).

6 Значимость предиктора/группы факторов

Значимость одного предиктора. Один коэффициент: t-тест $t = \dfrac{\hat\beta_j - \beta_{j,0}}{\widehat{\mathrm{se}}(\hat\beta_j)}$.

Значимость группы факторов. Группа ограничений $R\beta = r$ (ранг q):

$$F = \frac{(RSS_R - RSS_F)/q}{RSS_F/(n-k-1)} \sim F_{q,\,n-k-1},$$

где RSS_R и RSS_F — остаточные суммы квадратов ограниченной и полной моделей. Эквивалентная форма через $R(\hat\beta - \beta)$ и ковариацию $\hat\beta$.

7 Состоятельность МНК и асимптотическая нормальность

Состоятельность. При $\frac{1}{n}X^\top X \to Q \succ 0$, $E(\varepsilon|X) = 0$, $E(\varepsilon\varepsilon^\top|X) = \sigma^2 I$ и ограниченных моментах

$$\hat{\beta} \xrightarrow{p} \beta.$$

Асимптотическая нормальность.

$$\sqrt{n}(\hat{\beta} - \beta) \Rightarrow \mathcal{N}(0,\, \sigma^2 Q^{-1}).$$

При гетероскедастичности асимптотическая нормальность сохраняется, но ковариацию заменяют на робастную (HC0–HC3).

8 RSS, R^2, $R^2_{\mathbf{adj}}$, AIC, BIC

RSS и коэффициент детерминации.

$$RSS = \sum r_i^2, \quad R^2 = 1 - \frac{RSS}{TSS}, \quad R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-k-1}.$$

AIC и BIC. Для нормальной ошибки: $-2\ell(\hat{\theta}) = n\ln(RSS/n) + \text{const}$. Тогда

$$AIC = n\ln(RSS/n) + 2k, \qquad BIC = n\ln(RSS/n) + k\ln n$$

(с точностью до константы). **Критерий C_p Маллоуза.** $C_p = \frac{RSS}{\hat{\sigma}^2} - n + 2k$.

9 Ridge, Lasso, Elastic Net; байесовская мотивация

Ridge. Ridge: $\min RSS + \lambda\|\beta\|_2^2$, решение $\hat{\beta}_{ridge} = (X^\top X + \lambda I)^{-1}X^\top y$. Стабилизирует при коллинеарности; требует стандартизации признаков. **Lasso.** Lasso: $\min RSS + \lambda\|\beta\|_1$, даёт разреженные решения (координатный спуск/ISTA). **Elastic Net.** Elastic Net: $\lambda(\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2)$. Выбор λ по K-fold CV. **Байесовская интерпретация.** Нормальное априори для Ridge, лапласовское для Lasso.

10 Остатки: требования, стьюдентизированные остатки; нормальность

Требования к остаткам. Линейность, экзогенность, гомоскедастичность, независимость, нормальность.

Леверидж и стьюдентизированные остатки. Леверидж $h_i = x_i^\top(X^\top X)^{-1}x_i$. Стандартизированные/стьюдентизированные остатки: $r_i/(\hat{\sigma}\sqrt{1-h_i})$.

Диагностика и нормальность. Диагностика по графикам Residuals vs Fitted, QQ-plot, Scale–Location. Тесты нормальности: Джарка–Бера, Шапиро–Уилка.

Влияние наблюдений. Расстояние Кука $D_i = \frac{r_i^2}{p\hat{\sigma}^2}\frac{h_i}{(1-h_i)^2}$.

11 Автокорреляция остатков: признаки и тесты

Признаки автокорреляции. ACF/PACF остатков, график e_t vs e_{t-1}.
Тест Дарбина–Уотсона. $DW \approx 2(1-\hat{\rho}_1)$ для проверки первой автокорреляции.
Тест Бреуша–Годфри. Регрессия e_t на лаги $e_{t-1..p}$ и X; LM-статистика $\sim \chi^2_p$.
Тест Льюнга–Бокса.

$$Q = n(n+2)\sum_{h=1}^{H} \hat{\rho}_h^2/(n-h) \sim \chi^2_H.$$

12 Гетероскедастичность: признаки и тесты

Признаки гетероскедастичности. Графики Scale–Location и Residuals vs Fitted с веерообразным рисунком.

Тест Бреуша–Пагана. Регрессия e_i^2 на X, LM-статистика $\sim \chi^2_k$.

White-тест. Регрессия дисперсии с квадратичными и перекрёстными членами предикторов.

Робастные ковариации. Используют HC0–HC3 для корректировки ковариационной матрицы оценок.

13 Преобразования Бокса–Кокса и Йео–Джонсона

Преобразование Бокса–Кокса. Box–Cox (для $y > 0$): $g_\lambda(y) = \frac{y^\lambda - 1}{\lambda}$ при $\lambda \neq 0$, и $\ln y$ при $\lambda = 0$. **Преобразование Йео–Джонсона.** Yeo–Johnson допускает $y \leq 0$.
Практика. (i) подбор λ по максимуму $\ell(\lambda)$; (ii) трансформировать Y и/или X, переоценить модель и проверить гомоскедастичность; (iii) можно выбирать λ по минимуму $\log|\widehat{\Sigma}|$ по предикторам.

14 Подбор предикторов: forward/backward, ADD–Del

Forward. Старт с константы, добавляем лучший по AIC/BIC/R^2_{adj}/CV; стоп при отсутствии улучшения.

Backward. Старт с полной модели, удаляем наихудший предиктор.

ADD–Del. После каждого добавления пытаемся удалить любой уже включённый признак по тому же критерию.

Избегать утечки (фиксировать валидацию).

15 Бинарные переменные, one-hot; взаимодействия

One-hot кодирование. k уровней $\Rightarrow k-1$ дамми плюс базовая категория (иначе ловушка фиктивных переменных). Коэффициент при дамми — сдвиг относительно базы.

Взаимодействия. Взаимодействие $X \times D$ задаёт разные наклоны между группами; интерпретируем через комбинации базовых эффектов.

16 Нефиксированные потери: LAD, Huber, Tukey, LMS; LAD из ММП

LAD (L1). Минимизирует $\sum |e_i|$, ММП при лапласовских ошибках.

Huber. $\rho_c(e) = \begin{cases} e^2/2, & |e| \le c \\ c|e| - c^2/2, & |e| > c \end{cases}$, даёт линейную ψ-функцию.

Потеря Тьюки (biweight). Ограничивает влияние больших $|e|$.

LMS. Минимум медианы e_i^2; крайне робастно, но неэффективно.

Оценивание: IRLS/координатный спуск.

17 Сравнение LS/LAD/LMS; Пуассон-регрессия

LS. Оптимален при нормальных ошибках, чувствителен к выбросам.

LAD. Более устойчив к выбросам в y, даёт медианный фит.

LMS. Обеспечивает максимальную устойчивость, но имеет высокую дисперсию и дорог в вычислениях.

Пуассон-регрессия. Пуассон GLM: $Y \sim \text{Poisson}(\mu)$, линк $\log \mu = X\beta$, $\text{Var}(Y) = \mu$. Девианс

$$D = 2 \sum \big[y_i \ln(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i) \big].$$

18 Латентные переменные; EM-алгоритм

Латентные переменные. Ненаблюдаемые Z, которые влияют на распределение наблюдаемых Y и входят в полное правдоподобие.

EM-алгоритм. EM максимизирует $\ell(\theta)$ при латентных Z. E-шаг:

$$Q(\theta|\theta^{(t)}) = E_{Z|Y,\theta^{(t)}}[\ell_c(\theta)].$$

M-шаг: $\theta^{(t+1)} = \arg\max Q(\theta|\theta^{(t)})$. Для смеси норм: веса γ_{ik} и обновления π_k, μ_k, Σ_k. Свойство: монотонный рост $\ell(\theta^{(t)})$.

19 Временные ряды: определения; сезонность; стационарность; ADF

Белый шум. e_t некоррелирован, $E(e_t) = 0$, $\text{Var}(e_t) = \sigma^2$.

Случайное блуждание. $y_t = y_{t-1} + e_t$.

MA(q). $y_t = \sum_{i=0}^{q} \theta_i e_{t-i}$.

AR(p). $y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + e_t$.

Сезонность. Аддитивная/мультипликативная структура, повторяющаяся с периодом.

ADF-тест. Регрессия

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p} \psi_i \Delta y_{t-i} + u_t,$$

тест $H_0 : \gamma = 0$ на наличие единичного корня.

20 AR, MA, ARMA, ARIMA: уравнения, оценивание, проверка

ARIMA(p,d,q). $\nabla^d y_t$ моделируется ARMA(p,q).
Идентификация порядка. Анализ ACF/PACF, информкритерии AIC/BIC.
Оценивание. (Квази-)ММП или уравнения Йюла–Уокера для AR-части.
Проверка модели. Анализ остатков по ACF/PACF и тесту Льюнга–Бокса.

21 Регрессия с временными рядами; детрендирование; Гаусс–Марков для рядов

Детрендирование и сезонность. Удаляем тренд и вводим сезонные дамми-переменные перед оцениванием регрессии.

ARIMAX/GLS. При авторегрессии ошибок используют ARIMAX или GLS вместо обычного МНК.

Робастная ковариация HAC. Оценка Newey–West даёт корректные t/F-статистики при условной гетероскедастичности и слабой зависимости.

Гаусс–Марков для рядов. Классические BLUE не работают без некоррелированности ошибок, поэтому применяют GLS/Cohrane–Orcutt.