Статистика

DS-поток

Лекция 13

7.2 Анализ остатков

Остатки

В качестве оценки шума ε_i рассмотрим остатки $e_i = Y_i - \widehat{Y}_i$

Проверка свойств

Нормальность $\mathsf{H}_0\colon e_i \sim \mathcal{N}$	$\Longrightarrow$	Критерий Шапиро-Уилка и др.
Несмещенность $\mathsf{H}_0\colon \mathsf{E}e_i = 0$	$\Longrightarrow$	Критерии монотонного отнош. правд. В непарам. случае позже
Гомоскедастичность $\mathsf{H}_0\colon \mathsf{D}\varepsilon = \sigma^2 I_n$	$\Longrightarrow$	Тут не все так просто... Часто это свойство более критично

Остатки

$\mathrm{D}\varepsilon = \sigma^2 I_n$ — гомоскедастичность. Обратное — гетероскедастичность.

В качестве оценки шума ε_i рассмотрим остатки $e_i = Y_i - \widehat{Y}_i$

Проблема: $\mathrm{D}e_i \neq \sigma^2$ при гомоскедастичности.

$$e = Y - \widehat{Y} = (I_n - H)Y, \qquad \text{где } H = X\left(X^T X\right)^{-1} X^T$$

$$\mathrm{D}e = (I_n - H)\mathrm{D}Y(I_n - H)^T = \sigma^2 (I_n - H)(I_n - H)^T = \sigma^2 (I_n - H)$$

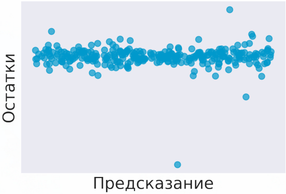
Проверять на однородность дисп. нужно **поправленные остатки**:

$$\widehat{e}_i = \frac{e_i}{\sqrt{\widehat{\mathrm{D}e_i}}} = \frac{e_i}{\sqrt{\frac{\left\|Y - X\widehat{\theta}\right\|^2}{n-d}(1 - H_{ii})}}$$ — стюдентизированные остатки

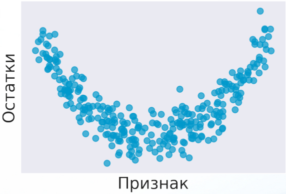
Визуальный анализ

Строятся графики зависимости $\widehat{e}_i$ от y, x, i

Все хорошо

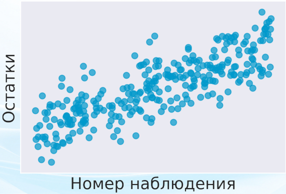


Есть выбросы



Нужно добавить x^2

Гетероскедастичность

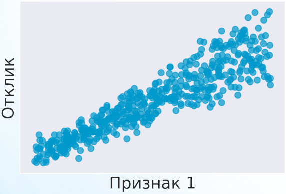


Тренд

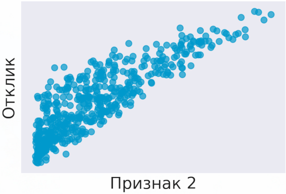
Неправильная модель

Визуальный анализ

Что будет если строить графики зависимостей таргета от признаков:

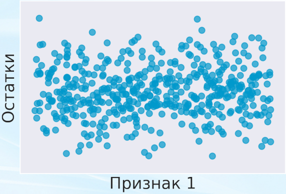


Гетероскедастичность?

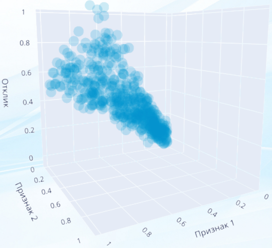


Гетероскедастичность?

Признаки зависимы



Нет, все хорошо!



Критерии проверки на гомоскедастичность

$H_0: D\varepsilon = \sigma^2 I_n$

Критерий Бройша-Пагана

$R^2_{\hat{e}^2}$ — коэф. детерминации для лин. регрессии предсказания $\hat{e}^2$ по X

$nR^2_{\hat{e}^2} \sim \chi^2_d$ — при справедливости H_0

Критерии проверки на гомоскедастичность

$\mathrm{H}_0 : \mathrm{D}\varepsilon = \sigma^2 I_n$

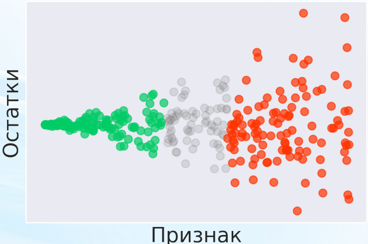
Критерий Голдфелда-Квандта

Упорядочим наблюдения по предполаг. возрастанию дисперсий.

$\left\| Y - X\widehat{\theta}_1 \right\|^2$ — остатки регрессии по первым $\frac{n-r}{2}$ наблюдений, $r > 0$

$\left\| Y - X\widehat{\theta}_2 \right\|^2$ — остатки регрессии по последним $\frac{n-r}{2}$ наблюдений

$$\frac{\left\| Y - X\widehat{\theta}_2 \right\|^2}{\left\| Y - X\widehat{\theta}_1 \right\|^2} \sim F_{\frac{n-r}{2}-d, \frac{n-r}{2}-d} \quad \text{при } \mathrm{H}_0$$



Что делать при гетероскедастичности?

- Если нужна только оценка θ — ничего;
- Если есть предположения о природе гетероскедастичности, взвесить наблюдения:

$$Y_i/\widehat{\sigma}_i = \left(x_i/\widehat{\sigma}_i\right)^T \theta + \varepsilon_i,$$

  где $\widehat{\sigma}_i$ — предполагаемая дисперсия при i-м измерении;
- Преобразование признаков и отклика, напр., Бокса-Кокса:

$$Z_i = \begin{cases} \ln Y_i, & \lambda = 0 \\ (Y_i^\lambda - 1)/\lambda, & \lambda \neq 0 \end{cases}$$

  Величина λ подбирается по графику зависимости MSE от λ
- Использовать специальные оценки дисперсии, устойчивые к гетероскедастичности.

Устойчивые оценки дисперсии

Пусть $\mathsf{E}\varepsilon = 0$ и $\mathsf{D}\varepsilon = V$.

Тогда $\Sigma = \mathsf{D}\widehat{\theta} = \left(X^T X\right)^{-1} X^T V X \left(X^T X\right)^{-1}$.

1. $V = \sigma^2 I_n$ — гомоскедастичность:
   $\Sigma = \sigma^2 \left(X^T X\right)^{-1}$ — дисперсия оценки коэффициентов;
   $\widehat{\Sigma} = \widehat{\sigma}^2 \left(X^T X\right)^{-1}$ — оценка дисперсии оценки коэффициентов;
2. $V = \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_n^2\right)$ — отсутствие автокорреляций:
   $\Sigma = \left(X^T X\right)^{-1} X^T \cdot \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_n^2\right) \cdot X \left(X^T X\right)^{-1}$ — д.о.к.;
   $\widehat{\Sigma} = \left(X^T X\right)^{-1} X^T \cdot \mathrm{diag}\left(\widehat{\sigma}_1^2, \ldots, \widehat{\sigma}_n^2\right) \cdot X \left(X^T X\right)^{-1}$ — о.д.о.к..
3. Наличие автокорреляций — более сложный случай, при котором зависимы элементы выборки.
   Ипользуются кластерное представление ковариационной матрицы или модели временных рядов.

Оценки Уайта

Если автокорреляции отсутствуют, используются **оценка Уайта**

White's heteroscedasticity-consistent estimator (HCE)

$$\widehat{\Sigma} = \left(X^T X\right)^{-1} X^T \cdot \operatorname{diag}\left(\widehat{\sigma}_1^2, ..., \widehat{\sigma}_n^2\right) \cdot X \left(X^T X\right)^{-1}$$

Варианты определения $\widehat{\sigma}_i^2$

1. HC0: $\widehat{e}_i^2$ — оценка Уайта
2. Модификации МакКиннона-Уайта

$$\text{HC1:} \quad \frac{n}{n-d}\widehat{e}_i^2, \qquad \text{HC2:} \quad \frac{\widehat{e}_i^2}{1-H_{ii}}, \qquad \text{HC3:} \quad \frac{\widehat{e}_i^2}{(1-H_{ii})^2}$$

Точнее оценивают при малых выборках.

Как ее применять?

Если автокорреляции отсутствуют, то выполнена асимптотическая нормальность оценки коэффициентов

$$\sqrt{n}\left(\widehat{\theta} - \theta\right) \xrightarrow{d} \mathcal{N}(0, B).$$

НСЕ дает состоятельную оценку на матрицу B:

$$n\widehat{\Sigma} \xrightarrow{\mathsf{P}} B.$$

Данный факт позволяет строить асимптотические дов. интервалы для коэффициентов моделей и таргета, а также критерий Вальда для проверки линейных гипотез $\mathsf{H}_0\colon\ T\theta = \tau$.


ВСЁ!