Прикладная статистика

и анализ данных

Съезд II

Анализ остатков

Остатки

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

Проверка свойств

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

Остатки

$\mathrm{D}\varepsilon = \sigma^2 I_n$ — гомоскедастичность. Обратное — гетероскедастичность.
В качестве оценки ошибки ε_i рассмотрим остатки $e_i = Y_i - \widehat{Y}_i$

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

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

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

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

$$\widehat{e}_i = \frac{e_i}{\sqrt{\widehat{\mathrm{D}e_i}}} = \frac{e_i}{\sqrt{\frac{RSS}{n-d}(1 - h_{ii})}}$$ — стюдентизированные остатки

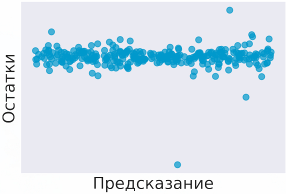
Т.к. $\mathrm{tr}\, H = d$ [*упр.*], то при $d \ll n$: $h_{ii} \approx 0$. Тогда

$$\widehat{e}_i = \frac{e_i}{\sqrt{RSS/(n-d)}}$$ — стандартизированные остатки

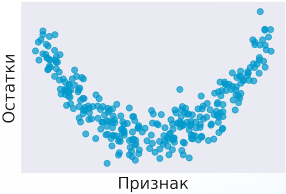
Визуальный анализ

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

Все хорошо

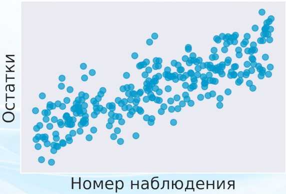


Есть выбросы



Нужно добавить x^2

Гетероскедастичность

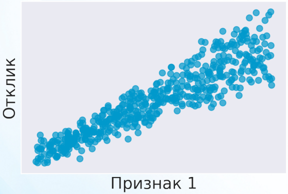


Тренд

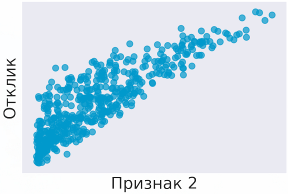
Неправильная модель

Визуальный анализ

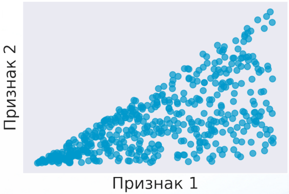
Что будет если строить графики зависимостей остатков от признаков:



Гетероскедастичность?

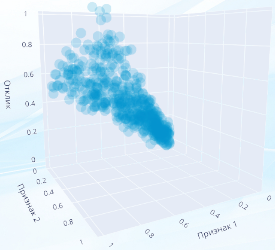


Гетероскедастичность?



Признаки зависимы

Нет, все хорошо!



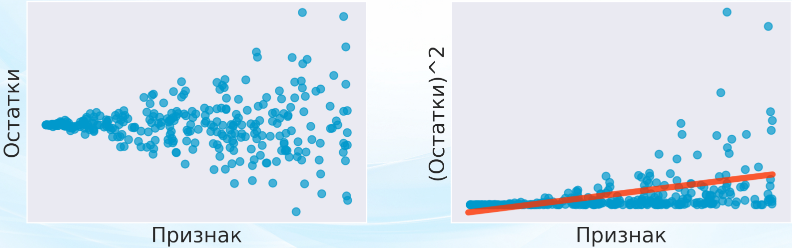
Критерии проверки на гомоскедастичность

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

Критерий Бройша-Пагана

$R^2_{\hat{e}^2}$ — коэф. детерминации при регрессии $\hat{e}^2 \sim X$

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



Критерии проверки на гомоскедастичность

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

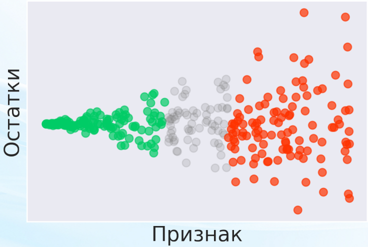
Критерий Голдфелда-Квандта

Упорядочим наблюдения по предполаг. возрастанию дисперсий.

RSS_1 — регрессия по первым $\frac{n-r}{2}$ наблюдений, $r > 0$

RSS_2 — регрессия по последним $\frac{n-r}{2}$ наблюдений

$$\frac{RSS_2}{RSS_1} \sim F_{\frac{n-r}{2}-d,\,\frac{n-r}{2}-d} \quad \text{при } \mathrm{H}_0$$



Что делать при гетероскедастичности?

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

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

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 $RSS(\lambda)$ от λ
- Использовать специальные оценки дисперсии, устойчивые к гетероскедастичности.

Устойчивые оценки дисперсии

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

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

1. $V = \sigma^2 I_n$ — гомоскедастичность:
   $\mathrm{D}\widehat{\theta} = \sigma^2 \left(X^T X\right)^{-1}$ — дисперсия оценки коэффициентов;
   $\widehat{\mathrm{D}\widehat{\theta}} = \widehat{\sigma}^2 \left(X^T X\right)^{-1}$ — оценка дисперсии оценки коэффициентов;

2. $V = \mathrm{diag}\left(\sigma_1^2, \dots, \sigma_n^2\right)$ — отсутствие автокорреляций:
   $\mathrm{D}\widehat{\theta} = \left(X^T X\right)^{-1} X^T \cdot \mathrm{diag}\left(\sigma_1^2, \dots, \sigma_n^2\right) \cdot X \left(X^T X\right)^{-1}$ — д.о.к.;
   $\widehat{\mathrm{D}\widehat{\theta}} = \left(X^T X\right)^{-1} X^T \cdot \mathrm{diag}\left(\widehat{\sigma}_1^2, \dots, \widehat{\sigma}_n^2\right) \cdot X \left(X^T X\right)^{-1}$ — о.д.о.к..

3. Наличие автокорреляций — см. временные ряды.

Оценки Уайта

Если автокорреляции отсутствуют, используются оценка Уайта White's heteroscedasticity-consistent estimator (HCE):

$$\widehat{\mathrm{D}\hat{\theta}} = \left(X^T X\right)^{-1} X^T \cdot \mathrm{diag}\left(\hat{\sigma}_1^2, ..., \hat{\sigma}_n^2\right) \cdot X \left(X^T X\right)^{-1}$$

Варианты определения $\hat{\sigma}_i^2$:

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

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

(точнее оценивают при малых выборках)

Асимптотическая нормальность при гетероскедаст.

Если автокорреляции отсутствуют, то

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

НСЕ дает состоятельную оценку на матрицу Σ:

$$n\widehat{\mathrm{D}\widehat{\theta}}\xrightarrow{\mathrm{P}}\Sigma$$

Данный факт позволяет строить асимптотические дов. интервалы и критерий Вальда для проверки линейных гипотез $\mathrm{H}_0 \colon T\theta=\tau$.


ВСЁ!