Статистика

DS-поток

Лекция 11

6.1 Гауссовская линейная модель

Доска

Доверительные интервалы

Величина	**Интервал**
σ	$\left(\sqrt{RSS(\widehat{\theta}) \,/\, \chi^2_{n-d,1-\alpha/2}},\quad \sqrt{RSS(\widehat{\theta}) \,/\, \chi^2_{n-d,\alpha/2}}\right)$
Дов. интервал для размера шума в отклике	
θ_j	$\left(\widehat{\theta}_j \pm T_{n-d,1-\alpha/2} \cdot \widehat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}\right)$
Дов. интервал для коэффициента перед j-м признаком	
$x_0^T\theta$	$\left(x_0^T\widehat{\theta} \pm T_{n-d,1-\alpha/2} \cdot \widehat{\sigma}\sqrt{x_0^T\,(X^TX)^{-1}\,x_0}\right)$
Дов. интервал для **среднего** отклика на объекте x_0	
$x_0^T\theta + \varepsilon$	$\left(x_0^T\widehat{\theta} \pm T_{n-d,1-\alpha/2} \cdot \widehat{\sigma}\sqrt{1 + x_0^T\,(X^TX)^{-1}\,x_0}\right)$
Предск. интервал для **наблюдаемого** отклика на объекте x_0	

Значим ли признак x_j?

Гипотеза о **не**значимости коэффициента θ_j

$\mathrm{H}_0\colon \theta_j = 0$ *vs.* $\mathrm{H}_1\colon \theta_j \ \{<, \neq, >\}\ 0$

Критерий Стьюдента

$$T_j^0(X, Y) = \frac{\widehat{\theta}_j}{\widehat{\sigma}\sqrt{(X^T X)^{-1}_{jj}}} \overset{\mathrm{H}_0}{\sim} T_{n-d}$$

$T_j^0(X, Y)$ — T-статистика критерия

Для $\mathrm{H}_1\colon \theta_j \ \neq \ 0$ критерий $\left\{ \left| T_j^0(X, y) \right| > T_{n-d, 1-\alpha/2} \right\}$

Если H_0 не отвергается, то можно считать, что θ_j отклоняется от нуля статистически незначимо. Возможно, признак j стоит убрать.

Значима ли группа признаков?

Пусть $\underset{n\times d}{X} = \left(\underset{n\times(d-k)}{X_1}, \quad \underset{n\times k}{X_2} \right), \quad \underset{d\times 1}{\theta} = \left(\underset{(d-k)\times 1}{\theta_1^T}, \quad \underset{k\times 1}{\theta_2^T} \right)^T$

$\mathsf{H}_0\colon \theta_2 = 0$ — гипотеза о незначимости второй группы

Критерий Фишера

$$\frac{\left(RSS(\widehat{\theta}_1) - RSS(\widehat{\theta})\right)/k}{RSS(\widehat{\theta})/(n-d)} \overset{\mathsf{H}_0}{\sim} F_{k,n-d},$$

где $\quad RSS(\widehat{\theta}) = \left\| Y - X\widehat{\theta} \right\|_2^2, \quad RSS(\widehat{\theta}_1) = \left\| Y - X_1\widehat{\theta}_1 \right\|_2^2$

Значима ли регрессия вообще?

Берем $k = d - 1$:

$$\frac{R^2/(d-1)}{(1-R^2)/(n-d)} \overset{\mathsf{H}_0}{\sim} F_{d-1,n-d}, \qquad \text{где } R^2 = 1 - \frac{RSS(\widehat{\theta})}{\left\| Y - \overline{Y} \right\|_2^2}$$

Таблица в statsmodels

```
# Load data
In [4]: dat = sm.datasets.get_rdataset("Guerry", "HistData").data

# Fit regression model (using the natural log of one of the regressors)
In [5]: results = smf.ols('Lottery ~ Literacy + np.log(Pop1831)', data=dat).fit()

# Inspect the results
In [6]: print(results.summary())
                            OLS Regression Results
==============================================================================
Dep. Variable:                Lottery   R-squared:                       0.348
Model:                            OLS   Adj. R-squared:                  0.333
Method:                 Least Squares   F-statistic:                     22.20
Date:                Mon, 14 May 2018   Prob (F-statistic):           1.90e-08
Time:                        21:48:09   Log-Likelihood:                -379.82
No. Observations:                  86   AIC:                             765.6
Df Residuals:                      83   BIC:                             773.0
Df Model:                           2
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
Intercept         246.4341     35.233      6.995      0.000     176.358     316.510
Literacy           -0.4889      0.128     -3.832      0.000      -0.743      -0.235
np.log(Pop1831)   -31.3114      5.977     -5.239      0.000     -43.199     -19.424
```

Разрыв мозга

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.991
Model:                            OLS   Adj. R-squared:                  0.991
Method:                 Least Squares   F-statistic:                     2651.
Date:                Mon, 11 Feb 2019   Prob (F-statistic):           2.68e-97
Time:                        03:50:00   Log-Likelihood:                 101.97
No. Observations:                 100   AIC:                            -195.9
Df Residuals:                      96   BIC:                            -185.5
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1            87.9447     93.569      0.940      0.350     -97.788     273.677
x2           -68.1838     88.958     -0.766      0.445    -244.764     108.396
x3           -86.4095     93.573     -0.923      0.358    -272.150      99.331
x4            68.3874     88.954      0.769      0.444    -108.184     244.959
```

Совместное применение крит. Фишера и Стьюдента

Возможны случаи разрыва мозга...

1. **Фишер:** группа признаков значима

   Стьюдент: ни один признак из группы не значим

   - отдельные признаки плохо объясняют y, но в совокупности OK
     один в поле не воин
   - признаки в группе мультиколлинеарны
     каждой твари по паре

2. **Фишер:** группа признаков не значима

   Стьюдент: в группе есть значимые признаки

   - незначимые признаки маскируют влияние значимых
     иголка в стоге сена

6.2 Анализ остатков

Остатки

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

Проверка свойств

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

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

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

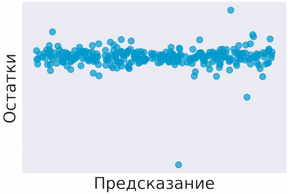
Проверять на однородность дисп. нужно **поправленные остатки**:

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

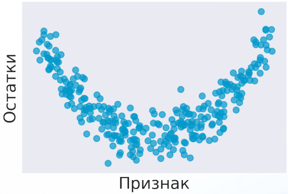
Визуальный анализ

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

Все хорошо

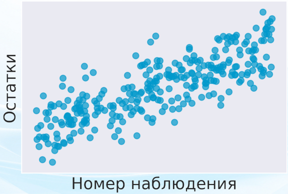


Есть выбросы



Нужно добавить x^2

Гетероскедастичность

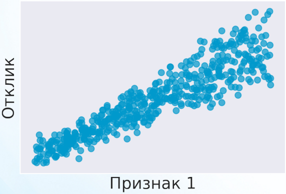


Тренд

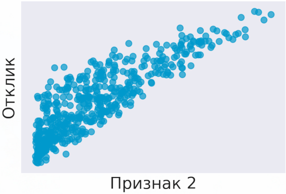
Неправильная модель

Визуальный анализ

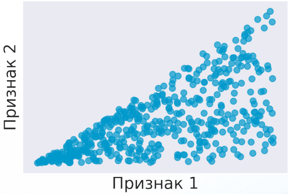
Что будет если строить графики зависимостей таргета от признаков:



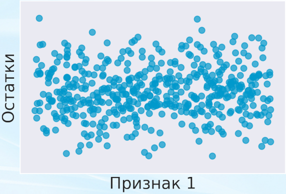
Гетероскедастичность?



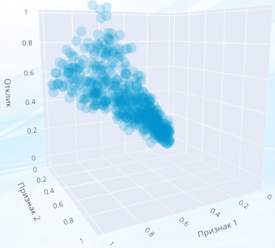
Гетероскедастичность?



Признаки зависимы



Нет, все хорошо!



Критерии проверки на гомоскедастичность

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

Критерий Бройша-Пагана

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

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

Критерии проверки на гомоскедастичность

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

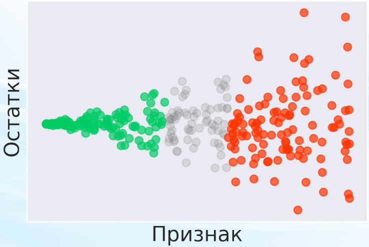
Критерий Голдфелда-Квандта

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

$$\frac{RSS_2}{RSS_1} \sim F_{\frac{n-r}{2}-d,\frac{n-r}{2}-d} \quad \text{при } H_0$$



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

1. $V = \sigma^2 I_n$ — гомоскедастичность:

   $\Sigma = \sigma^2 \left(X^T X\right)^{-1}$ — дисперсия оценки коэффициентов;

   $\hat{\Sigma} = \hat{\sigma}^2 \left(X^T X\right)^{-1}$ — оценка дисперсии оценки коэффициентов;

2. $V = \mathrm{diag}\left(\sigma_1^2, \dots, \sigma_n^2\right)$ — отсутствие автокорреляций:

   $\Sigma = \left(X^T X\right)^{-1} X^T \cdot \mathrm{diag}\left(\sigma_1^2, \dots, \sigma_n^2\right) \cdot X \left(X^T X\right)^{-1}$ — д.о.к.;

   $\hat{\Sigma} = \left(X^T X\right)^{-1} X^T \cdot \mathrm{diag}\left(\hat{\sigma}_1^2, \dots, \hat{\sigma}_n^2\right) \cdot X \left(X^T X\right)^{-1}$ — о.д.о.к..

3. Наличие автокорреляций — см. временные ряды.

Оценки Уайта

Если автокорреляции отсутствуют, используются **оценка Уайта**

White's heteroscedasticity-consistent estimator (HCE)

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

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

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

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

Точнее оценивают при малых выборках.

Асимптотическая нормальность при гетероскедаст.

Если автокорреляции отсутствуют, то

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

НСЕ дает состоятельную оценку на матрицу B:

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

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


ВСЁ!