

# Эконометрика-1

**Филатов Александр Юрьевич**

(Главный научный сотрудник, доцент ШЭМ ДВФУ)

[alexander.filatov@gmail.com](mailto:alexander.filatov@gmail.com)

<http://vk.com/alexander.filatov>, <http://vk.com/baikalreadings>

## Лекции 4.1-4.2

**Взвешенный и обобщенный МНК.  
Неоднородность. Дамми-переменные**



# Обобщенная линейная модель множественной регрессии (ОЛММР)

2

Второе условие классической модели может не выполняться:

$$Y = X\Theta + \varepsilon, \quad E\varepsilon = \mathbf{0}_n, \quad \sum \varepsilon = \sigma^2 \Sigma_0, \quad \text{rank} X = p + 1 < n.$$

$\sigma^2$  – неизвестная положительная константа,

$\Sigma_0$  – известная, не обязательно единичная матрица.

$\sigma^2$  – уже не является, как в классической модели дисперсией остатков.

Например, можно умножить  $\Sigma_0$  на любую константу, тогда  $\sigma^2$  разделится на нее.

## Частные случаи:

1. Модель с гетероскедастичными остатками (например, постоянство не абсолютного, а относительного разброса остатков).
2. Модель с автокоррелированными остатками (данные регистрируются во времени, регрессионные остатки взаимосвязаны).



## Обобщенный метод наименьших квадратов

МНК-оценки – состоятельные и несмещенные, но не эффективные.

### Критерий ОМНК:

$$(Y - X\Theta)^T \Sigma_0^{-1} (Y - X\Theta) \rightarrow \min_{\Theta}.$$

### ОМНК-оценки:

$$\hat{\Theta}_{ОМНК} = \left( X^T \Sigma_0^{-1} X \right)^{-1} \left( X^T \Sigma_0^{-1} Y \right) - \text{обладают всеми тремя свойствами.}$$

### Ковариационная матрица оценок параметров:

$$\Sigma_{\hat{\Theta}} = \sigma^2 \left( X^T \Sigma_0^{-1} X \right)^{-1},$$

### Дисперсия остатков:

$$\hat{\sigma}^2 = \frac{1}{n - p - 1} \left( Y - X \hat{\Theta}_{ОМНК} \right)^T \Sigma_0^{-1} \left( Y - X \hat{\Theta}_{ОМНК} \right)$$

### Проблема практической реализации ОМНК:

Матрица  $\Sigma_0$  – неизвестна в подавляющем большинстве случаев.

Включить ее элементы в число параметров нельзя, т.к. их число  $n(n+1)/2$  превышает объем данных  $np$ . Необходимо наложить ограничения.



## Модель с

гетероскедастичными остатками.

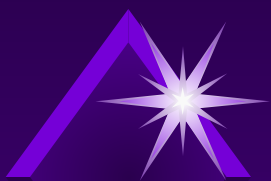
### Взвешенный метод наименьших квадратов

1. Остатки взаимно некоррелированы:  $r(\varepsilon_i, \varepsilon_j) = 0, i \neq j$ .
2. Остатки не обладают постоянной дисперсией:  $D\varepsilon_i \neq D\varepsilon_j, i \neq j$ .
3. По диагонали матрицы  $\Sigma_0$  стоят дисперсии:  $1/\lambda_i = D\varepsilon_i$ .

$$\Sigma_0 = \begin{pmatrix} 1/\lambda_1 & 0 & \dots & 0 \\ 0 & 1/\lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1/\lambda_n \end{pmatrix}$$

#### Критерий ВМНК:

$$\sum_{i=1}^n \lambda_i \left( y_i - \theta_0 - \theta_1 x_i^{(1)} - \dots - \theta_p x_i^{(p)} \right)^2 \rightarrow \min_{\theta_0, \theta_1, \dots, \theta_p} \quad \text{— чем больше разброс, тем меньше вес.}$$



# Проверка гетероскедастичности

5

Для проверки типично строится регрессия абсолютной величины остатков по некоторой функции от  $X$ :

$$|\hat{\varepsilon}_i| = \left| y_i - \hat{\theta}_{0.МНК} - \hat{\theta}_{1.МНК} x_i^{(1)} - \dots - \hat{\theta}_{p.МНК} x_i^{(p)} \right| = f\left(x_i^{(1)}, \dots, x_i^{(p)}\right)$$

Для подтверждения гетероскедастичности хотя бы один регрессор должен оказаться значимым.

## Варианты:

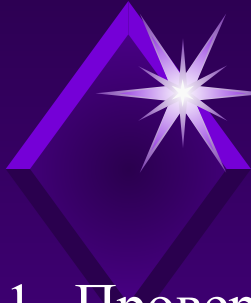
$$1. \quad |\hat{\varepsilon}| = a_0 + a_1 \left(x^{(j)}\right)^\gamma = \begin{cases} a_0 + a_1 x^{(j)} & \text{тест Глейсера,} \\ a_0 + a_1 / x^{(j)} & \text{— возможно обобщение} \\ \dots\dots\dots & \text{на несколько переменных.} \end{cases}$$

$$2. \quad |\hat{\varepsilon}| = a_0 \left(x^{(j)}\right)^{a_1} \text{ — тест Парка.}$$

$$3. \quad \hat{\varepsilon}^2 = a_0 + a^T x + x^T A x \text{ — тест Уайта.}$$

## Другие тесты:

1. Тест Голдфельда-Квандта (сравниваются дисперсии остатков по двум подвыборкам — при больших и малых значениях  $x^{(j)}$ ).
2. Тест Бартлетта (обобщение на произвольное число подвыборок).



# Практическое оценивание модели с гетероскедастичными остатками

6

1. Проверка гипотезы о наличии гетероскедастичности.
2. Переход от исходной модели к вспомогательной модели «с волной».

$$y_i = \theta_0 x_i^{(0)} + \theta_1 x_i^{(1)} + \dots + \theta_p x_i^{(p)} + \varepsilon_i \Rightarrow \tilde{y}_i = \theta_0 \tilde{x}_i^{(0)} + \theta_1 \tilde{x}_i^{(1)} + \dots + \theta_p \tilde{x}_i^{(p)} + \tilde{\varepsilon}_i$$
$$|\hat{\varepsilon}_i| = f(x_i^{(1)}, \dots, x_i^{(p)}) \Rightarrow \tilde{y}_i = \frac{y_i}{f(x_i^{(1)}, \dots, x_i^{(p)})}, \tilde{x}_i^{(0)} = \frac{1}{f(x_i^{(1)}, \dots, x_i^{(p)})}, \dots, \tilde{x}_i^{(p)} = \frac{x_i^{(p)}}{f(x_i^{(1)}, \dots, x_i^{(p)})}.$$

3. Оценивание коэффициентов  $\hat{\theta}_0, \hat{\theta}_1, \dots, \hat{\theta}_p$  вспомогательной модели с помощью обычного МНК, проверка значимости регрессоров.

## Замечание 1:

Оценивание в Excel происходит с учетом отсутствия свободного члена, т. к. он уже включен в модель. Используем **ЛИНЕЙН(y; X; 0; 1)**.

## Замечание 2:

Коэффициенты и их стандартные ошибки можно искать для вспомогательной модели, используя функцию ЛИНЕЙН. Для расчета  $R^2$  и ошибки прогноза, нужно вернуться в исходные координаты.



## Модель с

7

автокоррелированными остатками.

## Обобщенный метод наименьших квадратов

**Модель авторегрессии первого порядка:**

1. Данные регистрируются во времени.
2.  $|\rho| \in (0; 1)$  – коэффициент корреляции между соседними остатками.
3. Корреляция зависит только от разнесенности периодов во времени и ослабляется по мере ее роста:  $r(\varepsilon_i, \varepsilon_j) = \rho^{|i-j|}$ .

$$\sum_0 = \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \dots & \dots & \dots & \dots & \dots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{pmatrix}$$

**Формализация модели:**

$$\varepsilon_i = \rho \varepsilon_{i-1} + \delta_i, \quad |\rho| < 1, \quad E\delta_i \equiv 0, \quad E(\delta_i \delta_j) = \begin{cases} \sigma_0^2, & i = j \\ 0, & i \neq j \end{cases}$$



# Проверка автокорреляции. Критерий Дарбина-Уотсона

1. Выбираем уровень значимости  $\alpha$ .
2. Находим эмпирическое значение критерия

$$d = \sum_{i=2}^n (\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1})^2 / \sum_{i=1}^n \hat{\varepsilon}_i^2$$

В формуле  $\hat{\varepsilon}_i$  – остатки, вычисленные с помощью обычного МНК.

Если  $d \approx 2$ , то автокорреляции нет.

3. Вычисляем критические точки  $d_l(\alpha; n) < d_u(\alpha; n)$ .
4. Проверяем гипотезу о положительной/отрицательной автокорреляции.

## Случай $d < 2$ (наличие положительной автокорреляции):

$d < d_l \Rightarrow$  есть положительная автокорреляция,

$d \in [d_l; d_u] \Rightarrow$  неизвестно, есть ли положительная автокорреляция,

$d > d_u \Rightarrow$  положительной автокорреляции нет.

## Случай $d > 2$ (наличие отрицательной автокорреляции):

$4 - d < d_l \Rightarrow$  есть отрицательная автокорреляция,

$4 - d \in [d_l; d_u] \Rightarrow$  неизвестно, есть ли отрицательная автокорреляция,

$4 - d > d_u \Rightarrow$  отрицательной автокорреляции нет.



# Практическое оценивание модели с автокоррелированными остатками

9

1. Проверка гипотезы о наличии автокорреляции.
2. Переход от исходной модели к вспомогательной модели «с волной».  
$$y_i = \theta_0 x_i^{(0)} + \theta_1 x_i^{(1)} + \dots + \theta_p x_i^{(p)} + \varepsilon_i \Rightarrow \tilde{y}_i = \theta_0 \tilde{x}_i^{(0)} + \theta_1 \tilde{x}_i^{(1)} + \dots + \theta_p \tilde{x}_i^{(p)} + \tilde{\varepsilon}_i$$
$$\tilde{y}_1 = y_1 \sqrt{1 - \rho^2}, \quad \tilde{x}_1^{(0)} = \sqrt{1 - \rho^2}, \quad \tilde{x}_1^{(1)} = x_1^{(1)} \sqrt{1 - \rho^2}, \quad \dots, \quad \tilde{x}_1^{(p)} = x_1^{(p)} \sqrt{1 - \rho^2}.$$
$$\tilde{y}_i = y_i - \rho y_{i-1}, \quad \tilde{x}_i^{(0)} = 1 - \rho, \quad \tilde{x}_i^{(1)} = x_i^{(1)} - \rho x_{i-1}^{(1)}, \quad \dots, \quad \tilde{x}_i^{(p)} = x_i^{(p)} - \rho x_{i-1}^{(p)}, \quad i = 2, \dots, n.$$
3. Оценивание коэффициентов  $\hat{\theta}_0, \hat{\theta}_1, \dots, \hat{\theta}_p$  вспомогательной модели с помощью обычного МНК, проверка значимости регрессоров.

## Замечание 1:

Оценивание в Excel происходит с учетом отсутствия свободного члена, т. к. он уже включен в модель. Используем **ЛИНЕЙН(y; X; 0; 1)**.

## Замечание 2:

Коэффициенты и их стандартные ошибки можно искать для вспомогательной модели, используя функцию ЛИНЕЙН. Для расчета  $R^2$  и ошибки прогноза, нужно вернуться в исходные координаты.



# Итеративная процедура Кохрейна-Оркатта

10

1. Вычисляем МНК-оценки 1-итерации  $\hat{\theta}_0^{(1)}, \hat{\theta}_1^{(1)}, \dots, \hat{\theta}_p^{(1)}$ .
2. Подсчитываем остатки 1-итерации  $\varepsilon_i^{(1)} = y_i - \hat{\theta}_0^{(1)} - \hat{\theta}_1^{(1)}x_i^{(1)} - \hat{\theta}_p^{(1)}x_i^{(p)}$ .
3. С помощью МНК оцениваем параметры  $a_1, \dots, a_m$  1-итерации.  
$$\begin{aligned} |\hat{\varepsilon}_i^{(1)}| = a_0 + a_1x_i^{(j)} + \delta_i &\Rightarrow \hat{a}_0^{(1)}, \hat{a}_1^{(1)}, \\ \varepsilon_i^{(1)} = \rho\varepsilon_{i-1}^{(1)} + \delta_i &\Rightarrow \rho^{(1)}. \end{aligned}$$
4. Осуществляем переход к переменным  $\tilde{y}_i, \tilde{x}_i^{(0)}, \tilde{x}_i^{(1)}, \dots, \tilde{x}_i^{(p)}$ .
5. Вычисляем МНК-оценки 2-итерации  $\hat{\theta}_0^{(2)}, \hat{\theta}_1^{(2)}, \dots, \hat{\theta}_p^{(2)}$ .
6. Подсчитываем остатки 2-итерации  $\varepsilon_i^{(2)} = y_i - \hat{\theta}_0^{(2)} - \hat{\theta}_1^{(2)}x_i^{(1)} - \hat{\theta}_p^{(2)}x_i^{(p)}$ .
7. С помощью МНК оцениваем параметры  $a_1, \dots, a_m$  2-итерации.  
$$\begin{aligned} |\hat{\varepsilon}_i^{(2)}| = a_0 + a_1x_i^{(j)} + \delta_i &\Rightarrow \hat{a}_0^{(2)}, \hat{a}_1^{(2)}, \\ \varepsilon_i^{(2)} = \rho\varepsilon_{i-1}^{(2)} + \delta_i &\Rightarrow \rho^{(2)}. \end{aligned}$$
8. Осуществляем переход к переменным  $\tilde{y}_i, \tilde{x}_i^{(0)}, \tilde{x}_i^{(1)}, \dots, \tilde{x}_i^{(p)}$ .

.....



# Точечный прогноз в моделях линейной регрессии

# 11

Наиболее распространенная задача: предсказывать  $y$  по известным  $X$ .

$$\begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix} \begin{pmatrix} x_1^{(0)} = 1 & x_1^{(1)} & \dots & x_1^{(p)} \\ \dots & \dots & \dots & \dots \\ x_n^{(0)} = 1 & x_n^{(1)} & \dots & x_n^{(p)} \end{pmatrix} - \text{известные данные}$$

$$\begin{pmatrix} y_{n+1} \end{pmatrix} \begin{pmatrix} x_{n+1}^{(0)} = 1 & x_{n+1}^{(1)} & \dots & x_{n+1}^{(p)} \end{pmatrix} - \text{неизвестное значение}$$

Также известен характер ковариационных связей остатка  $\varepsilon_{n+1}$ :

$$\sigma_{\varepsilon}^{(n+1)} = (E(\varepsilon_1 \varepsilon_{n+1}), \dots, E(\varepsilon_n \varepsilon_{n+1}))^T, \quad E\varepsilon_{n+1} = 0, \quad E\varepsilon_{n+1}^2 = \Delta^2.$$

**Наилучший несмещенный прогноз для  $y_{n+1}$ :**

$$\hat{y}_{n+1} = X_{n+1}^T \hat{\Theta}_{ОМНК} + \left( \sigma_{\varepsilon}^{(n+1)} \right)^T \Sigma_{\varepsilon}^{-1} \hat{\varepsilon}.$$

Только если остаток  $\varepsilon_{n+1}$  не коррелирует ни с каким другим ( $\Sigma_0$  – диагональная матрица), прогноз совпадает со значением функции регрессии.

Для автокоррелированных остатков  $\hat{y}_{n+1} = X_{n+1}^T \hat{\Theta}_{ОМНК} + \rho \hat{\varepsilon}_n$ .



# Интервальный прогноз в моделях линейной регрессии

# 12

Для построения доверительного интервала необходима оценка точности точечного прогноза:

**Классическая модель:**

$$\sigma_{n+1}^2 = E(\hat{y}_{n+1} - y_{n+1})^2 = \hat{\sigma}^2 \left( X_{n+1}^T (X^T X)^{-1} X_{n+1} + 1 \right)$$
$$y_{n+1} \in \left[ \hat{y}_{n+1} - u_{(1+\gamma)/2} \sigma_{n+1}; \quad \hat{y}_{n+1} + u_{(1+\gamma)/2} \sigma_{n+1} \right].$$

**Частный случай парной регрессии:**

$$y_{n+1} \in \left[ \hat{y}_{n+1} - u_{\frac{1+\gamma}{2}} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}}; \quad \hat{y}_{n+1} + u_{\frac{1+\gamma}{2}} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \right].$$

**Обобщенная модель – отличия от классической:**

1.  $\sigma_{n+1}^2 = \hat{\sigma}^2 \left( X_{n+1}^T (X^T \Sigma_0^{-1} X)^{-1} X_{n+1} + 1 \right)$
2.  $\hat{\sigma}^2$  найдены на последней итерации практически реализуемого ОМНК.
3.  $y_{n+1} \in \left[ \hat{y}_{n+1} - t_{1-\gamma}(n-p-1) \sigma_{n+1}; \quad \hat{y}_{n+1} + t_{1-\gamma}(n-p-1) \sigma_{n+1} \right]$



# Неоднородность данных

13

Результирующий показатель  $y$  зависит не только от регрессоров  $X$ , но и от уровня сопутствующих переменных  $Z$  (как правило, не являющихся количественными).

## Сезонность, часы, пол, социальная страта, регион, кризис, санкции...

## Способы оценивания моделей с переменной структурой:

1. Разбиение имеющихся статистических данных на однородные порции (внутри каждой подвыборки значения переменных  $Z$  постоянны).

Для каждой подвыборки своя функция регрессии

$$\hat{f}(X, Z^*) = \hat{\theta}_0(Z^*) + \hat{\theta}_1(Z^*)x^{(1)} + \dots + \hat{\theta}_p(Z^*)x^{(p)}$$

При этом  $\hat{f}(X, Z^*)$  и  $\hat{f}(X, Z^{**})$  могут значительно отличаться.

### Проблемы:

- 1) сопутствующие переменные  $Z$  ненаблюдаемы, либо эти значения не были зарегистрированы при сборе исходных данных, прямое разбиение выборки невозможно.
  - 2) прямое разбиение возможно, но приводит к малым подвыборкам.
2. Метод дамми-переменных.



# Метод дамми-переменных

# 14

Если категоризованная переменная  $z^{(j)}$  имеет  $k_j$  градаций, вводим  $(k_j - 1)$  бинарных дамми-переменных, принимающих значения 0 или 1.

## Преимущества:

1. Сильно повышается статистическая надежность оценок.
2. Одновременно появляется возможность проверки гипотез о значимом влиянии сопутствующих переменных.

**## Уровень доходов (низкий / средний / высокий),  $k_1 = 3 - 1 = 2$ .**

$$z_i^{(1.1)} = \begin{cases} 1, & \text{если } i\text{-наблюдение за среднедоходным домашним хозяйством,} \\ 0, & \text{иначе;} \end{cases}$$

$$z_i^{(1.2)} = \begin{cases} 1, & \text{если } i\text{-наблюдение за высокодоходным домашним хозяйством,} \\ 0, & \text{иначе;} \end{cases}$$

**## Сезонность (зима / весна / лето / осень),  $k_2 = 4 - 1 = 3$ .**

$$z_i^{(2.1)} = \begin{cases} 1, & \text{если } i\text{-наблюдение осуществлено весной,} \\ 0, & \text{иначе;} \end{cases}$$

$$z_i^{(2.2)} = \begin{cases} 1, & \text{если } i\text{-наблюдение осуществлено летом,} \\ 0, & \text{иначе;} \end{cases}$$

$$z_i^{(2.3)} = \begin{cases} 1, & \text{если } i\text{-наблюдение осуществлено осенью,} \\ 0, & \text{иначе.} \end{cases}$$



# Модификации метода. Варианты зависимостей

# 15

**Пример.** Продажи мороженого в зависимости от цены, сезона и принадлежности к определенному уровню богатства.

**Вариант 1.** Спрос зависит от сезона, происходит параллельный сдвиг, меняется свободный член прогрессии  $\theta_0$  (абсолютное потребление).

Базовый зимний спрос составляет  $\hat{y} = \theta_0 + \theta_1 x$ ,

Весной, летом и осенью он соответственно растет на  $\theta_{2.1}$ ,  $\theta_{2.2}$  и  $\theta_{2.3}$ .

$$\hat{y} = \theta_0 + \theta_1 x + \theta_{2.1} z^{(2.1)} + \theta_{2.2} z^{(2.2)} + \theta_{2.3} z^{(2.3)}.$$

**Вариант 2.** При переходе из группы в группу меняется не абсолютное потребление, а отношение к цене, склонность к потреблению.

Для низкодоходной страты склонность к потреблению равна  $\theta_1$ .

Для среднедоходной и высокодоходной страты она соответственно увеличивается до уровня  $\theta_1 + \theta_{1.1}$  и  $\theta_1 + \theta_{1.2}$ .

$$\hat{y} = \theta_0 + \theta_1 x + \theta_{1.1} \left( z^{(1.1)} x \right) + \theta_{1.2} \left( z^{(1.2)} x \right) + \theta_{2.1} z^{(2.1)} + \theta_{2.2} z^{(2.2)} + \theta_{2.3} z^{(2.3)}.$$



## Несколько замечаний

# 16

### Замечание 1. Статистическая надежность:

Точность модели зависит от соотношения  $n / (p+1)$  — чем оно больше, тем точнее оценки.

## Помесячный спрос на мороженое за 5 лет, линейный тренд + зависимость от цены, числа торговых точек и цены конкурентов + сезонность.

1. Изолированная оценка по сезонам:  $n / (p+1) = (12 \cdot 5 / 4) / 5 = 3$

2. Оценка по дамми-переменным:  $n / (p+1) = (12 \cdot 5) / (3+5) = 7,5$ .

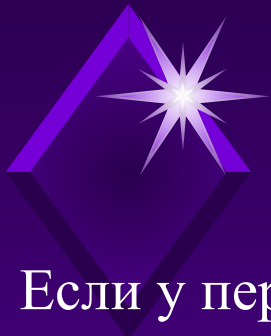
**Точность выросла в 2,5 раза. При большем числе подвыборок разница еще сильнее!**

### Замечание 2. Проверка неоднородности:

Дамми, как и обычные переменные, можно проверять на значимость. Если ни одна из них не является значимой, неоднородности нет!

### Замечание 3. Мультиколлинеарность:

При правильном использовании дамми мультиколлинеарность не возникает, даже если вводим 11 дамми для месяцев или 23 дамми для часов.



# Ловушка, связанная с введением дамми-переменных

# 17

Если у переменной  $z^{(j)}$  есть  $k$  градаций, то есть риск ввести  $k$  дамми.

$$z_i^{(2.4)} = \begin{cases} 1, & \text{если } i\text{-наблюдение осуществлено зимой,} \\ 0, & \text{иначе.} \end{cases}$$

| месяц    | $z^{(2.1)}$ | $z^{(2.2)}$ | $z^{(2.3)}$ | $z^{(2.4)}$ |
|----------|-------------|-------------|-------------|-------------|
| январь   | 0           | 0           | 0           | 1           |
| февраль  | 0           | 0           | 0           | 1           |
| март     | 1           | 0           | 0           | 0           |
| апрель   | 1           | 0           | 0           | 0           |
| май      | 1           | 0           | 0           | 0           |
| июнь     | 0           | 1           | 0           | 0           |
| июль     | 0           | 1           | 0           | 0           |
| август   | 0           | 1           | 0           | 0           |
| сентябрь | 0           | 0           | 1           | 0           |
| октябрь  | 0           | 0           | 1           | 0           |
| ноябрь   | 0           | 0           | 1           | 0           |
| декабрь  | 0           | 0           | 0           | 1           |

В данной модели присутствует линейная зависимость переменных (полная мульти-коллинеарность):

$$z^{(2.1)} + z^{(2.2)} + z^{(2.3)} + z^{(2.4)} = x^{(0)} \equiv 1.$$

Матрица  $X^T X$  – вырожденная, обратной матрицы  $(X^T X)^{-1}$  не существует, формулы МНК не работают.

**Количество дамми-переменных должно быть на единицу меньше числа градаций соответствующей категоризованной переменной!**

# Численный пример

18

## на использование дамми-переменных

|         | $y$ | $\tilde{x}$ | $I_p$ | $x$  | $z^{(1)}$ | $z^{(2)}$ | $z^{(3)}$ |
|---------|-----|-------------|-------|------|-----------|-----------|-----------|
| весна13 | 1,5 | 22          | 1     | 22,0 | 1         | 0         | 0         |
| лето13  | 2,6 | 22          | 1,019 | 21,6 | 0         | 1         | 0         |
| осень13 | 1,7 | 22          | 1,029 | 21,4 | 0         | 0         | 1         |
| зима13  | 0,9 | 22          | 1,046 | 21,0 | 0         | 0         | 0         |
| весна14 | 1,4 | 25          | 1,073 | 23,3 | 1         | 0         | 0         |
| лето14  | 3   | 25          | 1,095 | 22,8 | 0         | 1         | 0         |
| осень14 | 2,8 | 22          | 1,114 | 19,7 | 0         | 0         | 1         |
| зима14  | 1,6 | 22          | 1,202 | 18,3 | 0         | 0         | 0         |
| весна15 | 1,9 | 25          | 1,25  | 20,0 | 1         | 0         | 0         |
| лето15  | 3,2 | 25          | 1,266 | 19,7 | 0         | 1         | 0         |
| осень15 | 2,7 | 25          | 1,287 | 19,4 | 0         | 0         | 1         |
| зима15  | 2   | 25          | 1,32  | 18,9 | 0         | 0         | 0         |
| весна16 | 2,2 | 28          | 1,34  | 20,9 | 1         | 0         | 0         |
| лето16  | 3,4 | 28          | 1,358 | 20,6 | 0         | 1         | 0         |
| осень16 | 2,6 | 25          | 1,366 | 18,3 | 0         | 0         | 1         |
| зима16  | 2,1 | 25          | 1,386 | 18,0 | 0         | 0         | 0         |
| весна17 | 2,9 | 25          | 1,395 | 17,9 | 1         | 0         | 0         |
| лето17  | 3,3 | 30          | 1,41  | 21,3 | 0         | 1         | 0         |
| осень17 | 2,5 | 27          | 1,403 | 19,2 | 0         | 0         | 1         |
| зима17  | 2,2 | 27          | 1,416 | 19,1 | 0         | 0         | 0         |

Собраны данные по продажам мороженого ( $y$ , млн шт.) за 5 лет в зависимости от цены ( $\tilde{x}$ , руб.)

### Индексирование:

Поскольку за 5 лет инфляция превысила 40%, необходимо все цены привести к одному уровню, разделив на индекс цен:  $x = \tilde{x} / I_p$ .

### Исходная модель:

$$\hat{y} = 3,65 - 0,065 x, \quad \hat{R}^2 = 0,022.$$

(2,06)    (0,102)

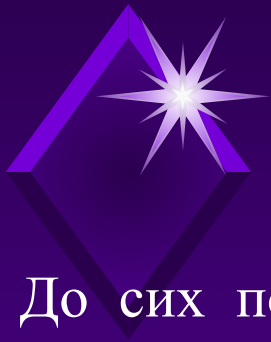
### Модель с дамми-переменными:

$$\hat{y} = 6,90 - 0,269 x + 0,691 z^{(1)} +$$

(1,04)    (0,054)    (0,216)

$$+ 1,916 z^{(2)} + 0,847 z^{(3)}, \quad \hat{R}^2 = 0,844.$$

(0,226)    (0,197)



# Учет эффекта взаимодействия сопутствующих факторов

19

До сих пор сопутствующие переменные влияли на результирующий показатель независимо, теперь рассмотрим случай их взаимодействия.

**Категоризованная переменная  $z^{(i)}$ :**

Соответствующие дамми-переменные:  $z^{(i.1)}, z^{(i.2)}, \dots, z^{(i.k_i-1)}$ .

**Категоризованная переменная  $z^{(j)}$ :**

Соответствующие дамми-переменные:  $z^{(j.1)}, z^{(j.2)}, \dots, z^{(j.k_j-1)}$ .

**Вводим  $N = (k_i - 1)(k_j - 1)$  новых дамми, образуемых всевозможными попарными произведениями  $z^{(qs)} = z^{(i.q)}z^{(j.s)}$ .**

| образование | пол     | $z^{(1.1)}$ | $z^{(1.2)}$ | $z^{(2.1)}$ | $z^{(3.1)} = z^{(1.1)} z^{(2.1)}$ | $z^{(3.2)} = z^{(1.2)} z^{(2.1)}$ |
|-------------|---------|-------------|-------------|-------------|-----------------------------------|-----------------------------------|
| начальное   | мужской | 0           | 0           | 0           | 0                                 | 0                                 |
| начальное   | женский | 0           | 0           | 1           | 0                                 | 0                                 |
| среднее     | мужской | 1           | 0           | 0           | 0                                 | 0                                 |
| среднее     | женский | 1           | 0           | 1           | 1                                 | 0                                 |
| высшее      | мужской | 0           | 1           | 0           | 0                                 | 0                                 |
| высшее      | женский | 0           | 1           | 1           | 0                                 | 1                                 |

$$\hat{y} = \theta_0 + \theta_1 x^{(1)} + \dots + \theta_p x^{(p)} + \theta_{1.1} z^{(1.1)} + \theta_{1.2} z^{(1.2)} + \theta_{2.1} z^{(2.1)} + \theta_{3.1} z^{(3.1)} + \theta_{3.2} z^{(3.2)}.$$

# Проверка регрессионной однородности двух групп наблюдений

20

## Случай 1. Большая выборка В1 + большая выборка В2

Статистическая проверка  $\theta_0^{(1)} = \theta_0^{(2)}, \theta_1^{(1)} = \theta_1^{(2)}, \dots, \theta_p^{(1)} = \theta_p^{(2)}$ .

Например, построить доверительные интервалы для коэффициентов из одной выборки, и проверять, входят ли в них коэффициенты из другой.

## Случай 2. Большая выборка В1 + малая выборка В2. Критерий Чоу.

1. Выбираем уровень значимости  $\alpha$ .

2. По В1 строим МНК-оценки и вычисляем невязки  $\hat{\varepsilon}^{(1)} = Y^{(1)} - X^{(1)} \hat{\Theta}^{(1)}$ .

3. По В2 строим МНК-оценки и вычисляем невязки  $\hat{\varepsilon}^{(2)} = Y^{(2)} - X^{(2)} \hat{\Theta}^{(2)}$ .

4. По В1+В2 строим МНК-оценки и вычисляем невязки  $\hat{\varepsilon} = Y - X \hat{\Theta}$ .

$$5. F_{\text{ЭМП}} = \frac{\left( \sum_{i=1}^{n_1+n_2} (\hat{\varepsilon}_i)^2 - \sum_{i=1}^{n_1} (\hat{\varepsilon}_i^{(1)})^2 - \sum_{i=1}^{n_2} (\hat{\varepsilon}_i^{(2)})^2 \right) / (p+1)}{\left( \sum_{i=1}^{n_1} (\hat{\varepsilon}_i^{(1)})^2 + \sum_{i=1}^{n_2} (\hat{\varepsilon}_i^{(2)})^2 \right) / (n_1 + n_2 - 2p - 2)}.$$

6.  $F_{\text{ЭМП}} > F_{\text{РАСПОБР}}(\alpha; p+1; n_1+n_2-2p-2) \Rightarrow$  В1 и В2 неоднородны.



# Проверка регрессионной однородности двух групп наблюдений

# 21

## Случай 3. Большая выборка В1 + сверхмалая выборка В2

Вторая выборка В2 настолько мала, что по ней нельзя получить значимые оценки коэффициентов регрессии (например, при  $n_2 < p+1$ ).

В частности, ситуация возникает при добавлении к исходной выборке В1 малой порции дополнительных данных – можно ли их объединять?

### Модифицированный критерий Чоу.

1. Выбираем уровень значимости  $\alpha$ .

2. По В1 строим МНК-оценки и вычисляем невязки  $\hat{\varepsilon}^{(1)} = Y^{(1)} - X^{(1)} \hat{\Theta}^{(1)}$ .

3. По В1+В2 строим МНК-оценки и вычисляем невязки  $\hat{\varepsilon} = Y - X \hat{\Theta}$ .

$$4. F_{\text{ЭМП}} = \frac{\left( \sum_{i=1}^{n_1+n_2} (\hat{\varepsilon}_i)^2 - \sum_{i=1}^{n_1} (\hat{\varepsilon}_i^{(1)})^2 \right) / n_2}{\sum_{i=1}^{n_1} (\hat{\varepsilon}_i^{(1)})^2 / (n_1 - p - 1)}$$

5.  $F_{\text{ЭМП}} > F_{\text{РАСПОБР}}(\alpha; n_2; n_1 - p - 1) \Rightarrow$  В1 и В2 неоднородны.

# Численный пример

# 22

## на проверку однородности выборок

Зависимость зарплаты от стажа и образования (пример из практики 2):

| $y$ | $x^{(1)}$ | $x^{(2)}$ |
|-----|-----------|-----------|
| 10  | 5         | 1         |
| 13  | 2         | 1         |
| 17  | 3         | 2         |
| 19  | 1         | 4         |
| 20  | 2         | 2         |
| 25  | 1         | 4         |
| 25  | 2         | 3         |
| 25  | 4         | 2         |
| 26  | 15        | 1         |
| 27  | 3         | 2         |
| ... | ...       | ...       |
| 280 | 18        | 5         |

### Основная выборка:

$$\hat{y}^{(1)} = -57,1 + 2,90x^{(1)} + 28,58x^{(2)}, \quad \sum (\hat{\varepsilon}_i^{(1)})^2 = 97066.$$

### Дополнительная выборка 1:

| $y$ | $x^{(1)}$ | $x^{(2)}$ |
|-----|-----------|-----------|
| 90  | 5         | 2         |
| 40  | 25        | 5         |

$$\hat{y}^{(1)} = -36,0 + 2,19x^{(1)} + 24,18x^{(2)}, \quad \sum \hat{\varepsilon}_i^2 = 113900.$$

$$F_{\text{эмп}} = \frac{113900 - 97066}{97066} \frac{38}{2} = 3,30, \quad F_{\text{крит}} = 3,24.$$

$3,30 > 3,24 \Rightarrow$  гипотеза об однородности отвергается.

### Дополнительная выборка 2:

| $y$ | $x^{(1)}$ | $x^{(2)}$ |
|-----|-----------|-----------|
| 180 | 16        | 5         |
| 160 | 9         | 4         |

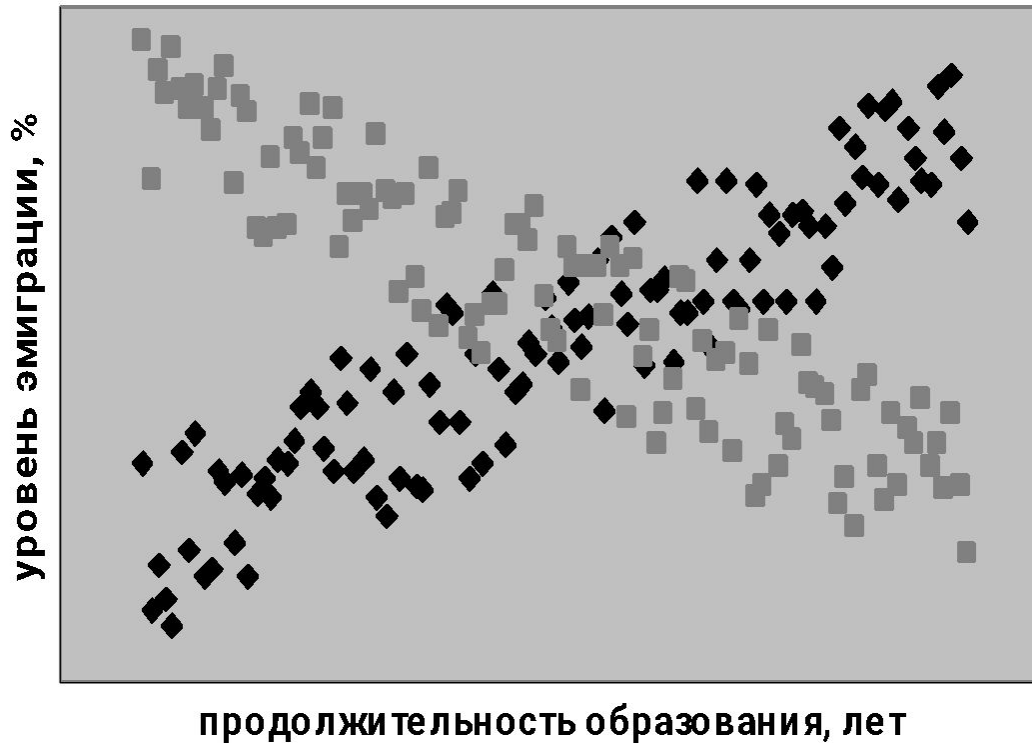
$$\hat{y}^{(1)} = -61,4 + 2,99x^{(1)} + 30,44x^{(2)}, \quad \sum \hat{\varepsilon}_i^2 = 104600.$$

$$F_{\text{эмп}} = \frac{104600 - 97066}{97066} \frac{38}{2} = 1,47, \quad F_{\text{крит}} = 3,24.$$

$1,47 < 3,24 \Rightarrow$  гипотеза об однородности принимается.

# Пример неоднородности данных при неизвестных сопутствующих факторах

23



Исследование проблемы «утечки мозгов» в 1990-е. Регрессионный анализ показывает отсутствия связи. **Геометрически** данные — две пересекающиеся крестом подвыборки.

**Вывод:** имеется скрытый сопутствующий признак — тип образования (гуманитарное / естественно-техническое).

**Проблема:** при  $p = 3$  визуальный анализ затруднен, а при  $p > 3$  практически невозможен.



*Спасибо  
за внимание!*

[alexander.filatov@gmail.com](mailto:alexander.filatov@gmail.com)

<http://vk.com/alexander.filatov>, <http://vk.com/baikalreadings>