

# Эконометрика-1

**Филатов Александр Юрьевич**

(Главный научный сотрудник, доцент ШЭМ ДВФУ)

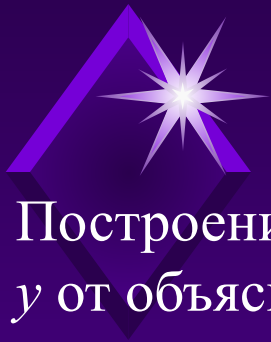
[alexander.filatov@gmail.com](mailto:alexander.filatov@gmail.com)

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

## Лекции 3.1-3.2

**Регрессионный анализ. МНК.**

**Мультиколлинеарность**



# Регрессионный анализ

Построение функциональной зависимости результирующей переменной  $y$  от объясняющих переменных  $x^{(1)}, \dots, x^{(n)}$ .

**Этимология (Фрэнсис Гальтон):** «регрессия» – отступление, возврат.

$x$  – рост отца                      Положительная связь, но тенденция возврата  
 $y$  – рост сына                      (отклонение сына < отклонения отца).

**Классическая линейная модель множественной регрессии (КЛММР):**

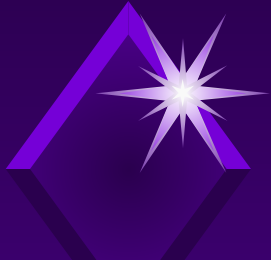
$$y_i = \theta_0 + \theta_1 x_i^{(1)} + \dots + \theta_p x_i^{(p)} + \varepsilon_i, \quad i = 1, \dots, n.$$

**Свойства:**

1.  $E\varepsilon_i = 0, \quad i = 1, \dots, n$  – остатки в среднем нулевые.
2.  $E(\varepsilon_i \varepsilon_j) = \begin{cases} \sigma^2, & i = j \\ 0, & i \neq j \end{cases}$  – гомоскедастичность.  
– взаимная некоррелированность.
3.  $\text{rank } X = p + 1 \leq n$  – линейная независимость регрессоров,  
существует матрица  $(X^T X)^{-1}$ ,  
если  $p + 1 > n$ , для выводов недостаточно данных.

# Линейная регрессия: матричная форма

3


$$Y = \begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_1^{(0)} = 1 & x_1^{(1)} & \dots & x_1^{(p)} \\ \dots & \dots & \dots & \dots \\ x_n^{(0)} = 1 & x_n^{(1)} & \dots & x_n^{(p)} \end{pmatrix}, \quad \Theta = \begin{pmatrix} \theta_0 \\ \dots \\ \theta_p \end{pmatrix},$$

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{pmatrix}, \quad \sum_{\varepsilon} = \begin{pmatrix} E(\varepsilon_1^2) & \dots & E(\varepsilon_1 \varepsilon_n) \\ \dots & \dots & \dots \\ E(\varepsilon_n \varepsilon_1) & \dots & E(\varepsilon_n^2) \end{pmatrix} \text{ – ковариационная матрица остатков.}$$

$$Y = X\Theta + \varepsilon, \quad E\varepsilon = \mathbf{0}_n, \quad \sum_{\varepsilon} = \sigma^2 E_n = \begin{pmatrix} \sigma^2 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & \sigma^2 \end{pmatrix}, \quad \text{rank} X = p+1 \leq n.$$

Если в дополнение к перечисленным 3 свойствам добавить распределение остатков по нормальному закону, получим нормальную КЛММР.



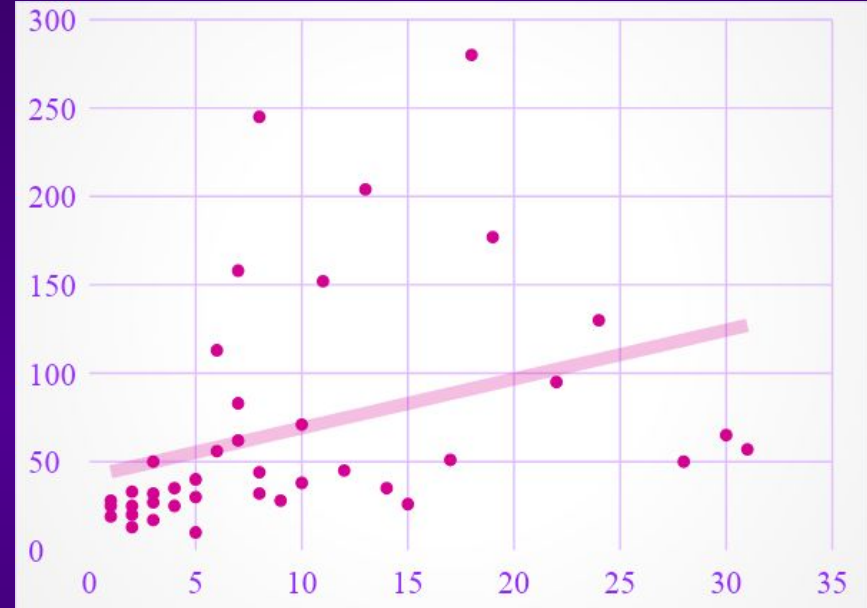
# Оценивание параметров. Метод наименьших квадратов

4

## Принцип:

Прогнозные значения должны минимально отличаться от наблюдаемых. Минимальность понимается в смысле суммы квадратов отклонений.

$$\sum_{i=1}^n \varepsilon_i^2 \rightarrow \min_{\theta_0, \dots, \theta_p},$$
$$\varepsilon_i = y_i - \theta_0 - \theta_1 x_i^{(1)} - \dots - \theta_p x_i^{(p)}$$



## Матричная форма:

$$\varepsilon = Y - X\Theta, \quad (Y - X\Theta)^T (Y - X\Theta) \rightarrow \min_{\Theta}, \quad (AB)^T = B^T A^T$$
$$Y^T Y - 2\Theta^T X^T Y + \Theta^T X^T X\Theta \rightarrow \min_{\Theta}, \quad -2X^T Y + 2X^T X\Theta = 0,$$
$$\hat{\Theta} = (X^T X)^{-1} X^T Y.$$

# Метод наименьших квадратов. Случай парной регрессии

$$X = \begin{pmatrix} 1 & x_1 \\ \dots & \dots \\ 1 & x_n \end{pmatrix}, \quad X^T X = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}, \quad X^T Y = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}.$$

$$(X^T X)\Theta = X^T Y, \quad \begin{cases} n\theta_0 + \theta_1 \sum x_i = \sum y_i, \\ \theta_0 \sum x_i + \theta_1 \sum x_i^2 = \sum x_i y_i. \end{cases} \quad \theta_0 = \frac{\sum y_i - \theta_1 \sum x_i}{n}.$$

$$\sum x_i \sum y_i - \theta_1 (\sum x_i)^2 + \theta_1 n \sum x_i^2 = n \sum x_i y_i.$$

**Формулы МНК для парной регрессии  $y = \theta_0 + \theta_1 x$ :**

$$\hat{\theta}_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}, \quad \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}.$$

# Численный пример

6

	объем	цена	рекл	празд
	$y$	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$
янв.16	91	1990	10	6
фев.16	93	1990	30	1
мар.16	84	1990	30	2
апр.16	77	1990	10	0
май.16	69	2190	10	3
июн.16	49	2190	0	1
июл.16	53	2190	0	0
авг.16	55	2190	20	0
сен.16	62	2190	20	0
окт.16	69	2190	20	0
ноя.16	68	2190	20	1
дек.16	109	2190	20	0
янв.17	70	2590	20	5
фев.17	87	2390	20	2

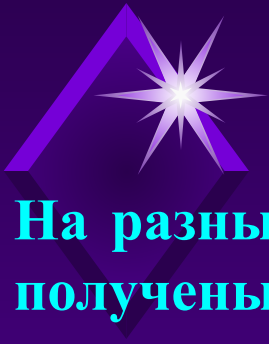
	объем	цена	рекл	празд
	$y$	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$
мар.17	66	2290	20	1
апр.17	61	2290	20	0
май.17	66	2290	20	3
июн.17	55	2090	50	1
июл.17	89	2090	50	0
авг.17	64	2090	10	0
сен.17	56	2090	0	0
окт.17	68	2090	0	0
ноя.17	109	2090	80	1
дек.17	115	1890	20	0
янв.18	95	2090	20	6
фев.18	88	2290	40	1
мар.18	82	2290	40	2
апр.18	72	2290	20	0

= ЛИНЕЙН ( $y_1, \dots, y_n; x_1^{(1)}, \dots, x_n^{(p)}; 1; 1$ ).

$3 \times (p+1) \Rightarrow$  формула  $\Rightarrow$  Ctrl-Shift-Enter

$$\hat{y}_i = 158,8 - 0,045x_i^{(1)} + 0,471x_i^{(2)} + 2,70x_i^{(3)}$$

<b>2,70</b>	<b>0,471</b>	<b>-0,045</b>	<b>158,8</b>
1,62	0,164	0,020	43,7
0,386	14,91	#Н/Д	#Н/Д



# Свойства оценок

7

На разных выборках за счет случайного характера остатков будут получены различные оценки!

**1. Состоятельность:**  $\lim_{n \rightarrow \infty} \hat{\theta} = \theta$ .

При росте выборки оценка стремится к истинному значению параметра (асимптотическое свойство проявляющееся при больших  $n$ ).

**Замечание 1:** Состоятельные оценки бывают разного качества.

## В случае симметрично распределенной случайной величины

$$\hat{\theta}_1 = \bar{x} = \frac{1}{n}(x_1 + \dots + x_n), \quad \hat{\theta}_2 = \frac{1}{2}(x_{\min} + x_{\max}) - \text{состоятельные оценки.}$$

**Замечание 2:** Состоятельная оценка может быть сколь угодно далекой от истинного значения.

## Средняя зарплата в отрасли, где работают  $n$  человек

$$\hat{\theta} = \begin{cases} \theta_0, & n < N \\ \bar{x}, & n = N \end{cases} \quad \text{при любом объеме выборки, кроме сплошного обследования, получаем сколь угодно завышенный результат.}$$



# Свойства оценок

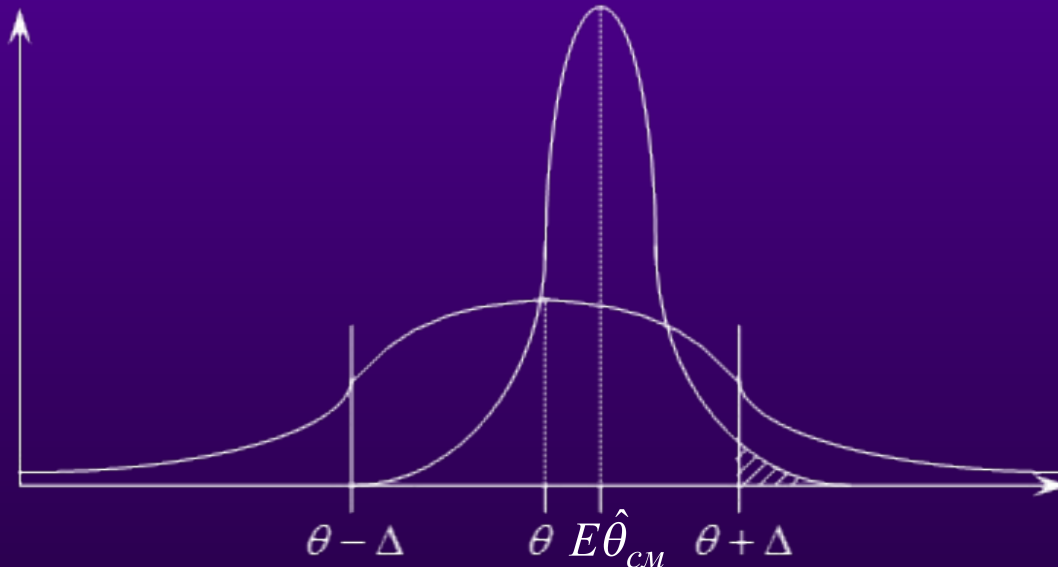
**2. Несмещенность:**  $E\hat{\theta} = \theta$  при любом объеме выборки.

Усреднение полученных оценок по всем выборкам данного объема дает истинное значение параметра (свойство «хороших свойств» оценки при каждом конечном объеме выборки).

**3. Эффективность:**  $E(\hat{\theta}_{eff} - \theta)^2 = \min_{\hat{\theta} \in M} (\hat{\theta} - \theta)^2$

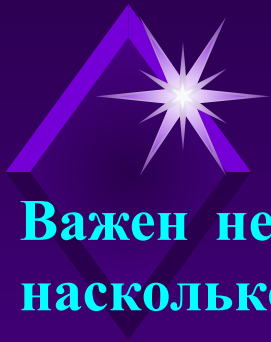
Эффективная оценка обладает наименьшим случайным разбросом в изучаемом классе  $M$ .

**Замечание:** Смещенная оценка может быть точнее несмещенной.



значения оценок  
на разных выборках





# Свойства оценок КЛММР

**Важен не только полученный по выборке вид регрессии, но и то, насколько мы можем ему доверять!**

**Несмещенная оценка ошибки прогноза:**

$$\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^n \left( y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i^{(1)} - \dots - \hat{\theta}_p x_i^{(p)} \right)^2.$$

$$\hat{\sigma} = 14,91.$$

2,70	0,471	-0,045	158,8
1,62	0,164	0,020	43,7
0,386	<b>14,91</b>	#Н/Д	#Н/Д

**Ковариационная матрица оценок параметров:**

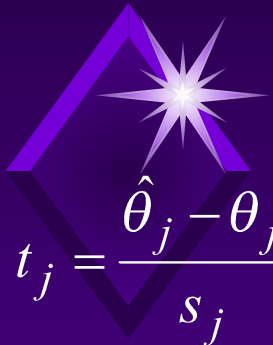
$$\hat{\Sigma}_{\hat{\Theta}} = E \left( (\hat{\Theta} - \Theta) (\hat{\Theta} - \Theta)^T \right) = \hat{\sigma}^2 (X^T X)^{-1}$$

Наиболее важными являются диагональные элементы – квадраты среднеквадратических ошибок  $s_j$  оценок коэффициентов  $\theta_j$ .

2,70	0,471	-0,045	158,8
<b>1,62</b>	<b>0,164</b>	<b>0,020</b>	<b>43,7</b>
0,386	14,91	#Н/Д	#Н/Д

$$\hat{y}_i = 158,8 - 0,045 x_i^{(1)} + 0,471 x_i^{(2)} + 2,70 x_i^{(3)}.$$

(43,7)
(0,020)
(0,164)
(1,62)



# Значимость регрессоров

# 10

$t_j = \frac{\hat{\theta}_j - \theta_j}{s_j} \sim t(n - p - 1)$  – распределена по закону Стьюдента.

**Проверка гипотезы о значимости регрессоров:  $H_0: \theta_j = 0$**

1. Задаем уровень значимости  $\alpha$ .
2. Находим эмпирическую точку  $t_j = \hat{\theta}_j / s_j$ .
3. Находим критическую точку  $t_{\text{крит}} = \text{СТЪЮДРАСПОБР}(\alpha; n - p - 1)$ .
4. Если  $|t_j| > t_{\text{крит}}$ , то  $H_0$  отвергается и делается вывод о наличии связи.

$$\hat{y}_i = 158,8 - 0,045 x_i^{(1)} + 0,471 x_i^{(2)} + 2,70 x_i^{(3)}.$$

(43,7)    (0,020)            (0,164)            (1,62)

$$t_0 = \frac{158,8}{43,7} = 3,64, \quad t_1 = \frac{-0,045}{0,020} = -2,22, \quad t_2 = \frac{0,471}{0,164} = 2,87, \quad t_3 = \frac{2,70}{1,62} = 1,67,$$

$$t_{\text{крит}} = \text{СТЪЮДРАСПОБР}(0,05; 28 - 3 - 1) = 2,06.$$

Гипотеза  $H_0$  принимается для  $\theta_3$  и отвергается для  $\theta_0, \theta_1, \theta_2$  при  $\alpha = 0,05$ .  
**Регрессор  $x^{(3)}$  незначим, коэффициент  $\theta_3$  не отличается значимо от 0, регрессоры  $x^{(1)}$  и  $x^{(2)}$  значимо влияют на  $y$ .**



# Построение

# 11

## доверительного интервала

$$\hat{y}_i = 158,8 - 0,045 x_i^{(1)} + 0,471 x_i^{(2)} + 2,70 x_i^{(3)}.$$

(43,7)      (0,020)      (0,164)      (1,62)

При уровне значимости 1% ( $t_{\text{крит}} = 2,80$ ) незначимой становится цена, при 0,1% ( $t_{\text{крит}} = 3,75$ ) – реклама.

При уровне значимости 10% ( $t_{\text{крит}} = 1,71$ ) число праздников по-прежнему незначимо, но если бы число наблюдений составило  $n=100$  ( $t_{\text{крит}} = 1,66$ ), то выводы сменились на противоположные.

### Построение доверительного интервала для $\theta_j$ :

1. Задаем доверительную вероятность  $\gamma$ .

$$2. \theta_j \in \left[ \hat{\theta}_j - t_{1-\gamma/2} (n-p-1) s_j; \hat{\theta}_j + t_{1-\gamma/2} (n-p-1) s_j \right].$$

$$\theta_0 \in [68,7; 249,0],$$

$$\theta_1 \in [-0,086; -0,003],$$

$$\theta_2 \in [0,132; 0,809],$$

$$\theta_3 \in [-0,64; 6,04] \text{ с вероятностью } \gamma = 0,95.$$



# Проверка гипотезы о значимости модели

# 12

**Проверка гипотезы о значимости модели:  $H_0: R^2 = 0$**

1. Задаем уровень значимости  $\alpha$ .
2. Находим эмпирическую точку  $F_{\text{эмп}} = \frac{\hat{R}_{y.X}^2}{1 - \hat{R}_{y.X}^2} \cdot \frac{n - p - 1}{p}$ .
3. Находим критическую точку  $F_{\text{крит}} = F_{\text{РАСПОБР}}(\alpha; p; n - p - 1)$ .
4. Если  $F_{\text{эмп}} > F_{\text{крит}}$ , то  $H_0$  отвергается и делается вывод о наличии связи, иначе гипотеза принимается, линейная модель неадекватна.

**В случае линейной модели квадрат множественного коэффициента корреляции  $R^2$  равен коэффициенту детерминации!**

$$\hat{R}^2 = 0,386, \quad F_{\text{эмп}} = \frac{0,386}{1 - 0,386} \cdot \frac{28 - 3 - 1}{3} = 5,03,$$

$$F_{\text{крит}} = F_{0,05}(3; 24) = 3,01.$$

Гипотеза  $H_0$  отвергается, линейная модель значима при  $\alpha = 0,05$ .



# Ошибки спецификации модели: исключение значащих переменных

# 13

## Неправомерное исключение значащих объясняющих переменных

- 1) Смещены оценки коэффициентов регрессии;
- 2) Еще сильнее смещена оценка дисперсии остатков.

Всё это приводит к неверным выводам!

## В примере не учтена дополнительная переменная – цена конкурента.  
Цена конкурента  $x^{(4)}$  в течение 24 месяцев из 28 совпадает с нашей.

### Но есть 4 отличающихся месяца:

Декабрь 2016:  $x_{12}^{(4)} = 2390$  конкурент раньше поднял цены.

Февраль 2017:  $x_{14}^{(4)} = 2590$  конкурент позже опустил цены.

Июнь 2017:  $x_{18}^{(4)} = 1690$  конкурент организовал летнюю распродажу.

Январь 2018:  $x_{25}^{(4)} = 1890$  конкурент продолжил зимнюю распродажу.

## Старая модель:

$$\hat{y}_i = 158,8 - 0,045 x_i^{(1)} + 0,471 x_i^{(2)} + 2,70 x_i^{(3)}, \quad \hat{R}^2 = 0,386.$$

(43,7)    (0,020)            (0,164)            (1,62)

## Новая модель:

$$\hat{y}_i = 201,3 - 0,177 x_i^{(1)} + 0,623 x_i^{(2)} + 4,22 x_i^{(3)} + 0,111 x_i^{(4)}, \quad \hat{R}^2 = 0,713.$$

(31,6)    (0,029)            (0,118)            (1,17)            (0,022)

## Можно учесть влияние предпраздничного месяца:

$$\hat{y}_i = 173,3 - 0,142 x_i^{(1)} + 0,641 x_i^{(2)} + 4,31 x_i^{(3)} + 0,085 x_i^{(4)} + 5,29 x_{i+1}^{(3)}, \quad \hat{R}^2 = 0,908.$$

(18,7)    (0,018)            (0,068)            (0,68)            (0,013)            (0,77)

## Есть риск введения в модель лишних несущественных переменных:

Меньшее из зол, однако при увеличении числа переменных

- 1) Ослабевают точность выводов, зависящая от  $n / (p+1)$ ;
- 2) Возможно появление **мультиколлинеарности** – взаимозависимости объясняющих переменных.



**Полная мультиколлинеарность** – линейная функциональная связь между объясняющими переменными, одна из них линейно выражается через остальные.

$\text{rank } X < p+1$ ,  $X^T X$  – вырожденная,  $(X^T X)^{-1}$  – не существует.

Избежать легко – на этапе отбора объясняющих переменных.

**Частичная мультиколлинеарность** – тесная, однако не функциональная связь между объясняющими переменными, выявляется сложнее.

## Эвристические рекомендации для выявления частичной мультиколлинеарности

1. Анализ корреляционной матрицы  $R$ :  $|r_{ij}| > 0,8$ .
2. Анализ обусловленности матрицы  $X^T X$ ,  $|X^T X| \approx 0$ .
3. Анализ собственных чисел матрицы  $X^T X$ ,  $\lambda_{\min} \approx 0$ .
4. Анализ коэффициентов детерминации каждой объясняющей переменной  $x^{(j)}$  по всем остальным:  $R^2_j > 0,9$ .



# Эвристические рекомендации для выявления частичной мультиколлинеарности

16

## 5. Анализ экономической сущности модели.

## Некоторые оценки коэффициентов имеют неверные с точки зрения экономической теории значения (неверные знаки, слишком большие или слишком малые значения).

## 6. Анализ чувствительности модели.

## Небольшое изменение данных (добавление или изъятие небольшой порции наблюдений) существенно изменяет оценки коэффициентов модели (вплоть до изменения знаков).

## 7. Анализ значимости модели.

## Большинство (или даже все) оценки коэффициентов модели статистически неотличимы от нуля, в то время как модель в целом является значимой.





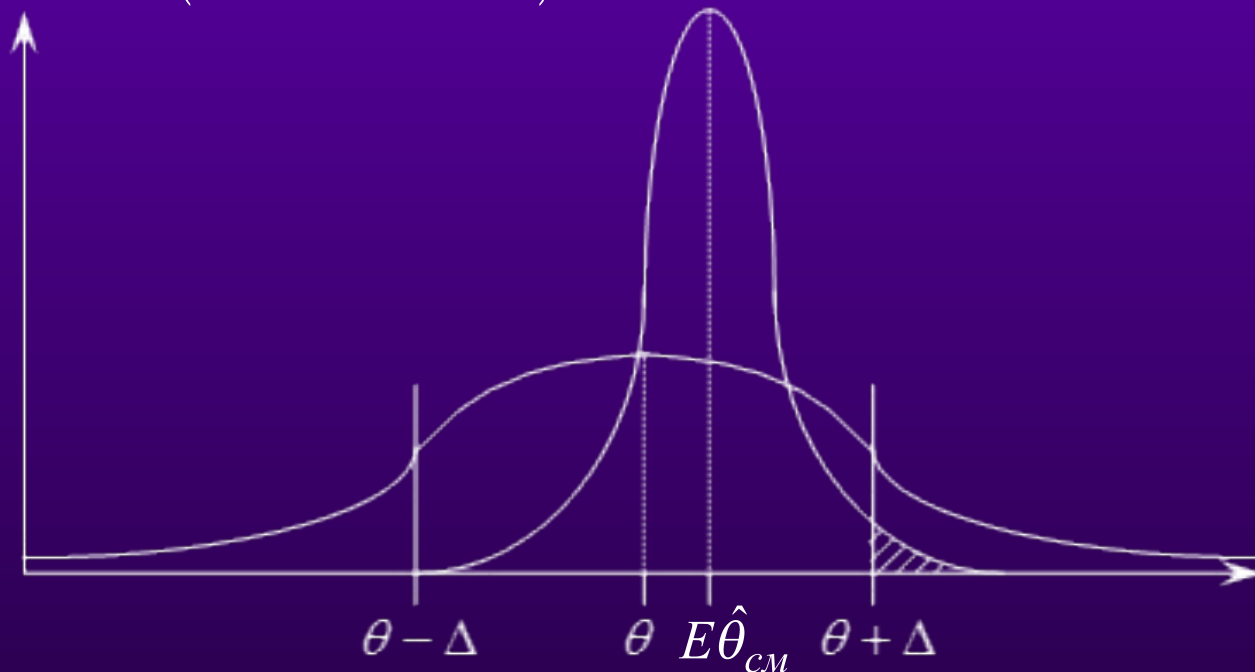
# Переход к смещенным методам оценивания

# 17

**Смещенная оценка может быть более точно, чем несмещенная!**

Один из методов – «ридж-регрессия» (ridge – гребень): добавляем к диагональным элементам матрицы  $X^T X$  «гребень»  $\tau \in (0, 1; 0, 4)$ , матрица становится хорошо обусловленной:

$$\hat{\Theta} = (X^T X + \tau E_{p+1})^{-1} X^T Y$$



значения оценок  
на разных выборках



# Отбор наиболее существенных объясняющих переменных

18

## 1. Версия всех возможных регрессий.

Для заданного  $k = 1, \dots, p - 1$  находится набор переменных  $x^{(j_1)}, \dots, x^{(j_k)}$ , дающих максимальное значение коэффициента детерминации  $R^2(k)$ .

Увеличиваем число переменных  $k$ , пока растет нижняя граница  $\sim 95\%$ -доверительного интервала для коэффициента детерминации.

$$R_{\min}^2(k) = \hat{R}_{\text{несм}}^2(k) - 2 \sqrt{\frac{2k(n-k-1)}{(n-1)(n^2-1)}} (1 - \hat{R}^2(k)), \quad \hat{R}_{\text{несм}}^2(k) = 1 - \left(1 - \hat{R}^2(k)\right) \frac{n-k}{n-k-1}.$$

**Проблема:** огромное количество переборов (для 20 переменных – более 1 млн).

## 2. Версия пошагового отбора переменных.

При переходе от  $k$  переменных к  $(k+1)$  учитываются результаты предыдущего шага – все отобранные переменные остаются навсегда.

**Проблема:** нет гарантии получения оптимума.



## 1.1. Подготовительный этап

- 1) Центрирование и нормирование переменных:  $(x_i^{(j)} - \bar{x}^{(j)}) / \sqrt{\sigma_j}$
- 2) Вычисление матрицы ковариаций

$$\Sigma = \begin{pmatrix} \hat{\sigma}_{11} & \dots & \hat{\sigma}_{1p} \\ \dots & \dots & \dots \\ \hat{\sigma}_{p1} & \dots & \hat{\sigma}_{pp} \end{pmatrix}, \quad \hat{\sigma}_{kj} = \frac{1}{n} \sum_{i=1}^n (x_i^{(k)} - \bar{x}^{(k)}) (x_i^{(j)} - \bar{x}^{(j)}) =$$

$$= \text{КОВАР}(x_1^{(k)}, \dots, x_n^{(k)}; x_1^{(j)}, \dots, x_n^{(j)}).$$

## 1.2. Решение характеристического уравнения $|\Sigma - \lambda E| = 0$

- 1) Нахождение собственных чисел  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{p'} > 0$ .
- 2) Нахождение собственного вектора  $l^{(k)}$  для каждого корня  $\lambda_k$ .

$$(\Sigma - \lambda_k E) l^{(k)} = 0, \quad \|l^{(k)}\| = 1.$$

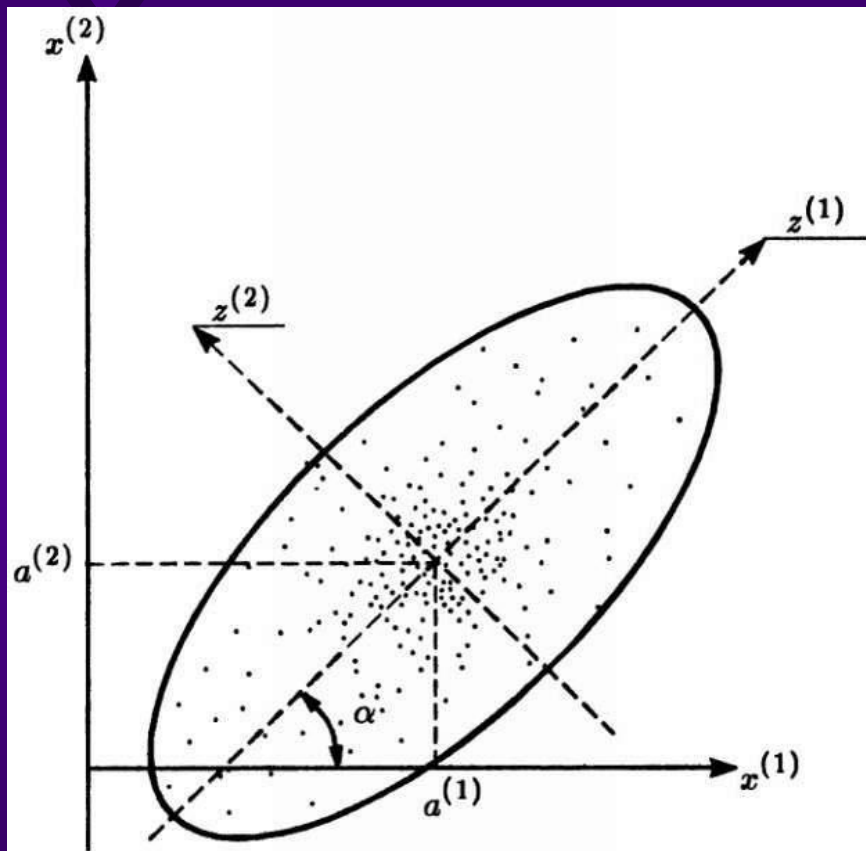
## 1.3. Переход к новым переменным $Z = XL$

$z^{(k)} = X l^{(k)}$ ,  $k = 1, \dots, p'$  – новые переменные, «главные компоненты»

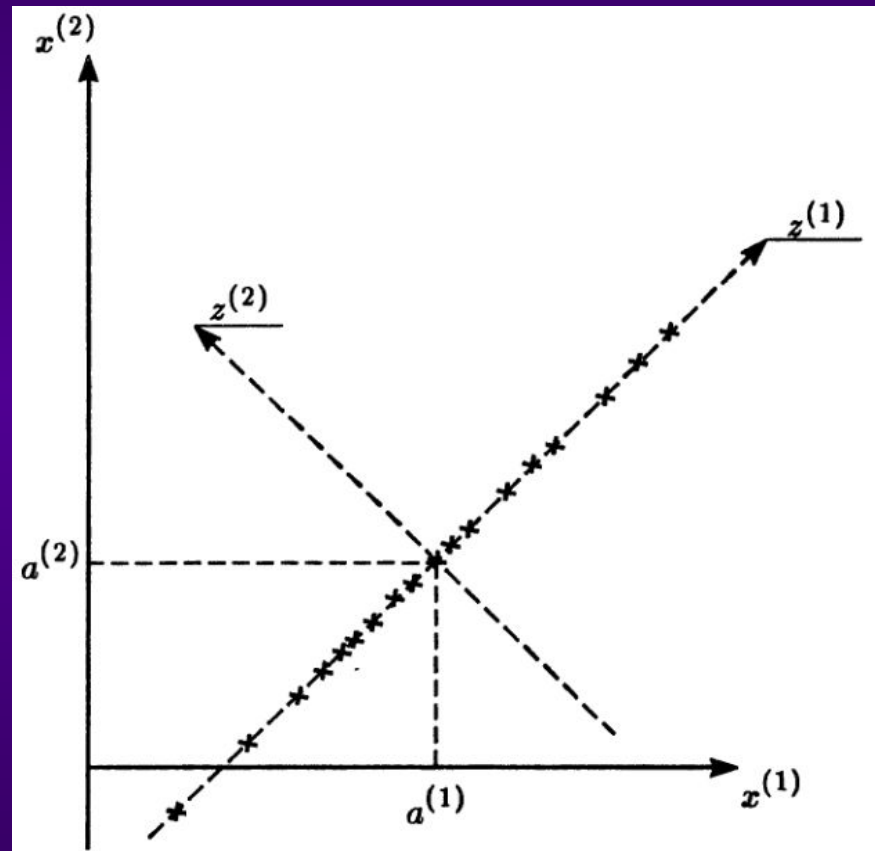
$I_{p'} = \frac{\lambda_1 + \dots + \lambda_{p'}}{\lambda_1 + \dots + \lambda_p}$  – доля дисперсии, вносимая первыми  $p'$  главными компонентами.

# Геометрическая интерпретация метода главных компонент

# 20



**Рис.1.** Умеренный разброс точек вдоль  $z^{(2)}$



**Рис.2.** Вырожденный случай: отсутствие разброса вдоль  $z^{(2)}$



# Проблема интерпретации метода главных компонент

# 21

**Матрица нагрузок главных компонент на исходные переменные:**

$$A \in R^{p \times p'}, \quad A = L\Lambda^{1/2}, \quad \Lambda^{1/2} = \text{diag}\{\sqrt{\lambda_j}\}, \quad a_{ij} = r(x^{(i)}, z^{(j)})$$

## Наблюдения – помесечные данные

$x^{(1)}$  – число торговых точек, где распространяется продукция, шт.

$x^{(2)}$  – расходы на рекламу, руб.

$x^{(3)}$  – доля новинок в ассортименте, %

$x^{(4)}$  – средний месячный доход на душу населения, руб.

$x^{(5)}$  – количество праздников, шт.

$$A = \begin{pmatrix} z^{(1)} & z^{(2)} \\ 0,95* & -0,19 \\ 0,97* & -0,17 \\ 0,94* & -0,28 \\ 0,24 & 0,88* \\ 0,56 & 0,67* \end{pmatrix} \begin{pmatrix} x^{(1)} \\ x^{(2)} \\ x^{(3)} \\ x^{(4)} \\ x^{(5)} \end{pmatrix}$$

$$\sum_{i=1}^p a_{ij}^2 = a_{1j}^2 + a_{2j}^2 + \dots + a_{pj}^2 = \lambda_j$$
$$\sum_{j=1}^p a_{ij}^2 = a_{i1}^2 + a_{i2}^2 + \dots + a_{ip}^2 = 1$$

$z^{(1)}$  тесно связана с  $x^{(1)}, x^{(2)}, x^{(3)}$

$z^{(2)}$  тесно связана с  $x^{(4)}, x^{(5)}$ .



22

*Спасибо  
за внимание!*

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

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