

## Анализ данных.

*Литература: Гайдыйшев И. Анализ и обработка данных: Специальный справочник. - СПб: Питер, 2001*

### 1 Эксперименты над моделями

В процессе имитационного эксперимента над моделями учетом случайных исходов каждого прогона необходимо:

1. классифицировать факторы на существенные и несущественные. (используется дисперсионный анализ)
2. разделить и оценить количественное влияние факторов и их комбинаций на целевую функцию. (используется регрессионный анализ)
3. найти наиболее выгодную комбинацию факторов. (используются методы статистического анализа)

Эксперимент начинается с выбора одной или нескольких целевых функций, которые количественно оценивают полезность системы. Типичной задачей системы массового обслуживания является обеспечение минимального времени системы при ограниченных затратах или минимизация затрат при среднем времени реакции не более заданного; обеспечить обслуживание максимального процента заявок за время не более заданного или минимизация времени, за которое обслуживается заданный процент заявок.

Измерением значений целевых функций является наблюдение, а все их множество результатом.

Основная задача – измерить влияние факторов на результат.

Каждое повторное моделирование – прогон.

Назначим значение (уровни) факторов, которые будут определять условия каждого прогона.

Комбинация уровней определяет условия прогона. Когда условия прогона неизменны мы называем мы называем эти прогоны репликами. Результаты в совокупности заполняют матрицу результатов. Протяженность матрицы по измерению должна быть равна числу

уровней этого фактора. Каждая размерность матрицы результатов сопоставляется каждому из факторов. Если использовать несколько целевых функций, о столько же матриц результатов.

	$\Pi_1$	$\Pi_2$	$\dots$	$\Pi_k$
$\Phi_1$	$m_{11}$	$m_{12}$	$\dots$	$m_{1k}$
$\Phi_2$	$m_{21}$	$m_{22}$	$\dots$	$m_{2k}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\Phi_n$	$m_{n1}$	$m_{n2}$	$\dots$	$m_{nk}$

## 2 Проверка гипотез

Закономерности в экспериментальных данных можно обнаружить путем статистической проверки результатов многих опытов, принимая или отклоняя определенные гипотезы.

Процедуры статистической проверки начинаются с формулировки нулевой гипотезы  $H_0$  (например: "нет статистически значимого различия") и альтернативной (конкурирующей) гипотезы  $H_1$  ("имеет место статистически значимое различие"). Затем проводится проверка нулевой гипотезы относительно альтернативной с помощью соответствующего теста (критерия). Результатом статистической проверки является вывод о том, в скольких случаях на каждые 100 проведенных испытаний в предположении определенной модели отклонения от модели можно считать случайными.

### 2.1 Общая методика

Статистическая величина лежит в доверительном интервале от  $t_0$  до  $t_1$  означает, что  $P(t_0 \leq \Theta \leq t_1)$  также обозначается  $p = 1 - a$ , где  $(1 - a)$  - коэффициент доверия,  $a$  - уровень значимости, выполняется в  $(1 - a)$  случаях.

Для проверки гипотез выборочно пространство  $W$  разделяют на две области:  $\omega$  и  $(W - \omega)$ , называемых критической и областью принятия.

Если известно распределение вероятностей, соответствующее  $H_0$ , можно определить так, чтобы при выполнении  $H_0$  вероятность отвергнуть эту гипотезу была равна уровню значимости

$$P\{x \in \omega | H_0\} = 1 - p$$

Доверительный уровень  $p = 1 - \alpha$ . Мощность критерия определяется как вероятность того, что критерий отклонит ложную нулевую гипотезу

$$P\{x \in (W - \omega) | H_0\} = 1 - p$$

Значение  $(1 - \beta)$  – мощность критерия.

Формулировки результатов применения критерия:

1.  $p < 0,05$  -различия недостоверны, нулевая гипотеза об отсутствии различий может быть принята.
2.  $p < 0,01$  - различия достоверны на уровне значимости 0,05.
3.  $p < 0,01$  - различия достоверны на уровне значимости 0,01.
4.  $p < 0,005$  - различия достоверны на уровне значимости 0,005.
5.  $p < 0,001$  - различия достоверны на уровне значимости 0,001.

Может быть принята альтернативная гипотеза о наличии различий на указанном уровне значимости.

### 3 Проверка нормальности распределения признака

Многие статистические методы требуют соответствия нормальному распределению, а значит предварительно необходимо проверить статистические данные на "нормальность".

3 метода:

1. Графический.  
Строим гистограмму распределения и сравниваем с кривой нормального распределения.
2. "Правило трех сигм".  
При нормальном распределении 97,7 – 97,8% всех значений признака лежат в интервале  $M_x \pm 3\sigma$ ,  $M_x$  - мат. ожидание,  $\sigma$  - среднеквадратичное отклонение.
3. Критерий К-С.  
Вычисляется величина  $k_s = \sqrt{N} \sup_x |F_N(x) - F(x)|$ ,  $F_N(x)$  – эмпирическая функция распределения,  $F(x)$  – теоретическая функция распределения, N – количество наблюдений. Чем меньше  $k_s$ , тем ближе признак  $x$  к теоретическому закону F(x).

## 4 Дисперсионный анализ

К большинству сложных систем применим принцип Парето, согласно которому 20% факторов определяют свойства системы на 80%. Поэтому первоочередной задачей исследования имитационной модели является отсеивание несущественных факторов, позволяющих оптимизировать задачу исходной модели.

*Анализ дисперсий* (Analysis of Variation – ANOVA) оценивает отклонение наблюдений от общего среднего. Затем вариация разбивается на две части, каждая из которых имеет свою причину. Остаточная часть вариации (residual), которую не удастся связать с условиями, считается его случайной ошибкой. Ошибка оценивается в средне-квадратическом отклонении, т.е. эффекты, которые превышают величину среднего-квадратич. отклонения считаются значимыми. Для значимости используется тест "F-статистика" (Критерий Фишера). Существует однофакторный (one-way) и многофакторный (multi-way) дисперсионный анализ.

### 4.1 Пример двухфакторного анализа

1. фактор  $i = \overline{1, I}$
2. фактор  $j = \overline{1, J}$

$$S = \sum_i \sum_j (y_{ij} - y_{..})^2 = \sum_i \sum_j (y_{0j} - y_{..})^2 + \sum_i \sum_j (y_{ij} - y_{.j})^2 = S_1 + S_2,$$

где  $y_{..}$  – усреднение по соответствующему индексу.

$$S_1 = \sum_i \sum_j (y_{.j} - y_{..})^2 = I \sum_j (y_{.j} - y_{..})^2$$

$S_1$  – рассеивание средних для уровня 2-го фактора относительно общего.  
 $S_2$  представляет рассеивание откликов внутри уровней.

$$y_{.j} = \frac{\sum_i y_{ij}}{I}, \quad y_{..} = \frac{\sum_i \sum_j y_{ij}}{IJ}.$$

Поправка на число связей позволяет получить несмещенную оценку СКО при отсутствии других эффектов. Если верна гипотеза  $H_0$  о независимости ожидаемых откликов от уровней факторов, то

нормированным числом степеней свободы компонент можно представить в виде:

$$\frac{S_1}{J-1} = \sigma_Y^2 \chi_{J-1}^2; \quad \frac{S_2}{J(J-1)} = \sigma_Y^2 \chi_{J(J-1)}^2.$$

$\sigma_Y^2$  – общая дисперсия откликов, по предположению не зависит от уровня факторов.

$\chi^2$  – случайная величина, распределенная по  $\chi^2$  с  $M$  степенями свободы. Отношение  $R$  нормированных компонент должно быть подчинено распределению Фишера с параметром  $F_{J-1, J(J-1)}$  с указанием пары степеней свободы.

Если  $R > r(p)$ ,  $r(p)$  выбрано в зависимости от доверительной вероятности  $p$ , то рассматриваемый фактор является значимым. В противном случае верна  $H_0$ .

Пример: (таблица 1)

Таблица 1: Таблица ANOVA

Источник вариации	Сумма квадратов	Степени свободы	Средн. квадрат	$F$	Крит. значение $F(p = 0.05)$
A	28.000	2	14.000	5.600	3.89
B	21.000	3	7.000	2.800	3.49
AB	11.000	6	1.833	0.733	3.00
Ошибка	30,000	12	2,500		
Всего	90,000	23			

$$F = \frac{\text{средний квадрат}}{\text{СКО}}, \quad \text{средний квадрат} = \frac{\text{сумма квадратов}}{\text{СКО}}$$

Если  $F < \text{критическое значение } F$ , то факторы ( $B$  и  $AB$ ) являются незначимыми. Необходимо либо увеличить  $F$ , либо уменьшить критическое значение  $F$ .

## 4.2 Планирование эксперимента

Число уровней каждого фактора обычно ограничивают 2-мя. Тогда объём полного эксперимента с  $k$  факторами составит  $2_k$  прогонов.

Теория планирования эксперимента рассматривает проблему построения наиболее экономичных планов, позволяя получить оценивающие коэффициенты. Рациональный план содержит

значительно меньше экспериментов, чем полный факторный эксперимент.

Дробный эксперимент: каждый прогон кодируется последовательностью латинских букв, обозначающую принимаемых максимальные значения факторов. Компоненты, заведомо включенные в план, называются генератором эксперимента.

пример: все факторы принимают максимальные значения.

$$I = ABCDE$$

Затем назначается дробность эксперимента  $m$  и выбирается  $(m * 2^k - 1)$  факторов и их комбинации, эффект которых по результатам статистической обработки нужно различать.

При  $k = 5$  и  $m = \frac{1}{2}$  получается  $m = \frac{1}{2} * 2^5 - 1 = 15$  факторов.

$$A, B, C, D, E, AB, AC, AD, AE, BC, BD, BE, CD, CE, DE$$

Получаем план эксперимента: (таблица 2)

Таблица 2: план эксперимента

Фактор	ABCDE	A	B	C	D	E	AB	AC	AD	AE	BC	BD	BE	CD	CE	DE
A	1	1	-1	-1	-1	-1	1	1	1	1	-1	-1				
B	1	-1	1	-1	-1	-1	1	-1	-1	-1	1	1				
C	1	-1	-1	1	-1	-1	-1	1	-1	-1	1	-1				
D	1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	1				
E	1	-1	-1	-1	-1	1	-1	-1	-1	1	-1	-1				

Пример:  $CE * ABCDE = AB(C * C)D(E * E) = ABD \Rightarrow$  из  $ABCDE$  можно убрать несущественные факторы и получить также самое  $(A = BCDE)$  и т.д.

четверть  $2^{5-2} = 7$   $A, B, C, D, D, E, AD, AE$

$$I = ABC, CDE, ABDE$$

$$A = BC = ACDE = BDE$$

$$B = AC = BCDE = AE$$

$$C = AB = DE = ABCDE$$

$$D = ABCD = CE = ABE$$

$$E = ABCE = CD = ABD$$

$$AD = BCD = ACE = BE$$

$$AE = BCE = ACD = BD$$

Восьмая  $2^{5-3} = 3 : A, B, C$   
 $I = ABC, ADE, CD, BCDE, BAD$   
 $A = BC = DE = ACD = ABCDE = BD$   
 $B = AC = ABDE = BCD = CDE = AD$   
 $C = AB = ACDE = D = BDE = ABCD$

## 5 Кластерный анализ

*Цель* – построение алгоритмов классификации исследуемых объектов. Каждый объект характеризуется одинаковым числом  $n$  переменных, являющимися точками в  $k$ -мерном пространстве. В результате кластеризации "близкие" объекты разбиваются в группы. Близость понимается как расстояние  $d$  в  $k$ -мерном пространстве.

### 5.1 Расстояние между объектами

Расстояние (метрика) на множестве объектов – это число  $d(x, y)$ , поставленное в соответствие любым двум объектам  $x$  и  $y$  и обладающее свойствами:

- а)  $d(x, y) = 0$
- б)  $d(x, y) = d(y, x)$
- в)  $d(x, y) \leq d(x, z) + d(z, y)$  (неравенство треугольника)

Примеры:

- а) Линейное расстояние  $d(x, y) = \sum_{i=1}^n |x_i - y_i|$

- б) Евклидово  $d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$

- в) Расстояние Минковского степени  $p \geq 1$ ,

$$d(x, y) = \left( \sum_{i=1}^n |x_i - y_i|^p \right)^{\frac{1}{p}}$$

- г) Расстояние Махалонобиса между объектами  $x_i$  и  $y_i$ , заданными таблицей:

Номер объекта k	Объекты	Признаки
1	$x_1$	$x_{11}x_{12} \dots x_{1n}$
2	$x_2$	$x_{21}x_{22} \dots x_{2n}$
$\vdots$	$\vdots$	$\dots$
p	$x_p$	$x_{p1}x_{p2} \dots x_{pn}$

$$d(x_i, x_j) = (x_i - x_j)A_x^{-1}(x_i - x_j)^T$$

$A_x = ||a_{ij}||$  - выборочная ковариационная матрица  $x_i = (x_{i1}, \dots, x_{ip})$

## 5.2 Близость

Близость на множестве объектов - число  $\beta(x, y)$ , поставленное в соответствие 2 объектам  $x, y$ , обладающее свойствами:

1.  $\beta(x, y)$  – функция двух переменных  $x$  и  $y$ , т.е.

$$\lim_{\substack{x \rightarrow x_0 \\ y \rightarrow y_0}} \beta(x, y) = \beta(x_0, y_0)$$

2.  $\beta(x, y) = \beta(y, x)$
3.  $0 \leq \beta(x, y) \leq 1$  и  $\beta(x, y) = 1 \Leftrightarrow x = y$

Примеры:

1. Косинус

$$\beta(x, y) = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2}}$$

2. Коэффициент корреляции

$$\beta(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x}) \sum_{i=1}^n (y_i - \bar{y})}{(n-1)\sigma_x \sigma_y}$$

где  $\sigma_x \sigma_y$  - соответственно среднее и среднеквадратичное отклонение  $x, y$ .



### 5.3 Расстояние между кластерами

Методы вычисления расстояния между кластерами, основанные на понятии расстояния между объектами:

- а) Среднее расстояние между всеми объектами пары кластеров с учетом расстояния внутри кластеров.
- б) Расстояние между ближайшими соседями - ближайшими объектами кластеров.
- в) Расстояние между самыми далекими соседями.
- г) Расстояние между центрами кластеров.
- д) Метод медиан.
- г) Метод медиан, но центр объединенного кластера вычисляется как среднее всех объектов.
- е) Среднее расстояние между кластерами.
- ж) Метод Варда. В качестве расстояния между кластерами берется прирост суммы квадратов расстояний объектов до центров кластеров, получаемый в результате их объединения.

### 5.4 Иерархический алгоритм

III1 Каждый объект объединяется в кластер (0-уровень)

III2 Два ближайших кластера объединяются в новый кластер. В качестве расстояния между кластерами используют расстояние между объектами (1-уровень).

...

III<sub>n</sub> Пусть даны кластеры (n-1)-уровня. Два ближайших кластера объединяются в новый кластер. в качестве расстояния между кластерами используют заранее выбранное расстояние между кластерами.

...

IIIN Все объекты объединяются в единый кластер

## 6 Регрессионный анализ

Дана независимая переменная  $x$  и ее  $n$  значений  $x_1, \dots, x_n$ , полученных в эксперименте. Одновременно известны значения  $y_1, \dots, y_n$  другой переменной  $y$ , зависимой от  $x$ . Задача линейного регрессивного анализа заключается в определении переменной  $y$  как линейной функции переменной  $x$

$$y = ax + b$$

### 6.1 Метод наименьших квадратов

$$\begin{aligned} U(a, b) &= \sum_{i=1}^n [y_i - (a + bx_i)]^2 \rightarrow \min \\ \frac{\delta U}{\delta a} &= \frac{\delta U}{\delta b} = 0 \\ a &= \frac{\sum y_i \sum x_i^2 - \sum x_i y_i \sum x_i}{n \sum x_i^2 - (\sum x_i)^2} \\ b &= \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} \end{aligned}$$

### 6.2 Статистическая значимость модели линейной регрессии.

**F- критерий Фишера.**

$$\begin{aligned} \overline{S_y^2} &= \frac{\sum y_i^2 - \frac{1}{n}(\sum y_i)^2}{n - 1} && \text{— дисперсия среднего} \\ \overline{S_{y,oc}^2} &= \frac{\sum [y_i - (a + bx_i)]^2}{n - 2} && \text{— остаточная дисперсия} \\ F &= \frac{\overline{S_y^2}}{\overline{S_{y,oc}^2}} \end{aligned}$$

Чем меньше  $\overline{S_{y,oc}^2}$ , тем лучше уравнение регрессии описывает переменную  $y$ .

Величина  $F$  распределяется по Фишеру. Для проверки значимости уравнения регрессии зададим уровень значимости  $\alpha (= 0.05, = 0.01, = 0.001)$  и рассмотрим критерий

$$P(F < F_{(n-1, n-2, \alpha)}^{tab}) = \alpha$$

$F_{(n-1, n-2, \alpha)}^{tab}$  берется из таблиц распределения Фишера. Следовательно, если  $F < F^{tab}$ , то при уровне значимости  $\alpha * 100\%$  уравнение регрессии статистически значимо, т.е. адекватно описывает результаты эксперимента.

### 6.3 Проверка наличия корреляции

*Корреляция* - наличие линейной зависимости  $x_1, \dots, x_n$  и  $y_1, \dots, y_n$ .

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad \text{коэффициент корреляции}$$

при  $r \neq 0$  можно говорить о наличии корреляции. Для статистической значимости отличия коэффициента корреляции от 0, проведем проверку статистической гипотезы:  $H_0 : r = 0$  Используем критерий Стьюдента:

$$P(|t| > t_{n-2, \alpha}^{tab}) = \alpha$$

где

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Если  $|t| < t_{n-2, \alpha}^{tab}$ , то нет оснований отвергнуть гипотезу  $H_0$ . Но если  $|t| > t_{n-2, \alpha}^{tab}$ , то гипотезу отвергают и принимают конкурирующую гипотезу  $H_1 : r \neq 0$

## Анализ данных(продолжение).

### 7 Компонентный анализ

Пусть дано  $n$  наблюдаемых величин  $x_1, \dots, x_n$ , описывающих исследуемое явление.

Цель компонентного анализа состоит в том, чтобы преобразовать величины  $x_1, \dots, x_n$  в набор главных компонент  $f_1, \dots, f_n$  так, чтобы учет каждой следующей компоненты  $f_{j+1}$  соответствовал более точному приближению суммарной дисперсии:

$$\sum_{i=1}^n Dx_i$$

переменных  $x_1, \dots, x_n$ . Обычно выделение первой главной компоненты  $f_1$  должно соответствовать самой большой доле  $\lambda_1$  вклада этой компоненты в суммарную дисперсию. Вторая компонента  $f_2$  объясняет уже меньшую долю  $\lambda_2 < \lambda_1$  суммарной дисперсии и т.д.

## 7.1 Методы компонентного анализа

Поиск компонент  $f_1, \dots, f_n$  подобен вращению "осей"  $x_1, \dots, x_n$  поскольку используются следующие уравнения, связывающие  $x_i$  с  $f_j$ :

$$x_i = \sum_{j=1}^m \omega_{ij} f_j, \quad i = \overline{1, n}$$

$\omega_{ij}$  - нагрузки  $j$ -й компоненты в  $i$ -й переменной.

В матричной записи:

$$X = W * F$$

Предполагаем, что величины  $x_i, y_i, f_i$  центрированные, т.е.

$$Mx_i = My_i = Mf_i = 0, \quad i = \overline{1, n}$$

"Вращение осей" задается в виде

$$Y = UX, \quad X = U'Y$$

где  $U$  - ортогональная матрица.

На первом этапе поиска главных компонент рассматриваем вместо осей  $f_1, \dots, f_m$  вспомогательные "оси"  $y_1, \dots, y_m$ . Если обозначить через  $u_k$  -  $k$ -ую строку матрицы  $U$ , то

$$y_i = u_i X = \sum_{k=1}^n U_{ik} x_k$$

Положим  $\lambda_i = My_i^2$ . Это дисперсия переменной  $y_i$ . Естественно считать, что величины  $y_i$  некоррелированные, т.е.  $M(y_i y_j) = 0$ ,  $(i \neq j)$ , то

$$My_i y_j = M * \sum_{k=1, s=1}^n u_{ik} x_k u_{js} x_s = \sum_{k=1, s=1}^n u_{ik} M(x_k x_s)$$

$$(u)'_{sj} = \sum_{k=1, s=1}^n u_{ik} k_{ks} (u)'_{sj} = u_i K_x u'_j = \lambda_i \delta_{ij},$$

где  $K_x = ||k_{kj}||$ ,  $k_{kj} = M(x_k x_j)$  - ковариационная матрица для  $x$ -ов. Следовательно, матрица  $\Lambda = UK_x U'$  является диагональной. Иначе говоря, благодаря вращению мы привели ее к диагональному виду. Элементы  $\lambda_1, \dots, \lambda_n$  матрицы  $\Lambda$  расположены в порядке убывания по величине, т.к.  $u_i K_x u_i' = \lambda_i$ , а матрица  $U$  - ортогональная, т.е.  $u_i u_i' = 1$ , то  $K_x u_i' = \lambda_i u_i'$ .

Иначе говоря,  $\lambda_i$  - собственное число матрицы  $K_x$ ,  $u_i$  - собственный вектор.

Теперь вычислим главные компоненты, нормируя  $y_i$ :

$$\delta_i = \lambda^{-\frac{1}{2}} y_i = \lambda^{-\frac{1}{2}} u_i X$$

или в матричном виде

$$F = \Lambda^{-\frac{1}{2}} Y = \Lambda^{-\frac{1}{2}} U X \quad \Rightarrow \quad X = U' \Lambda^{\frac{1}{2}} F$$

и для матрицы весов имеем

$$W = U' \Lambda^{\frac{1}{2}}$$

Поскольку

$$\sum_{i=1}^n D x_i = Sp[K_x] = Sp[\Lambda] = \sum_{i=1}^n D y_i, \quad Sp - \text{след, т.е. сумма собств. значений}$$

т.е. суммарная дисперсия переменных  $x_i$  равна суммарной дисперсии главных компонент  $y_i$ .

Таким образом, можно легко найти процент, вносимый каждой главной компонентой в суммарную дисперсию переменных  $x_i$ .

Основная идея метода главных компонент заключается в выделении первых компонент  $f_1, \dots, f_m$  ( $m < n$ ), дающих высокий процентный вклад в суммарную дисперсию.

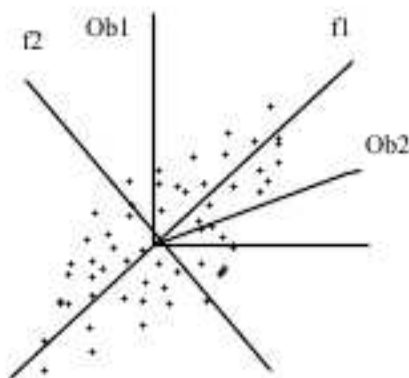
## 7.2 Геометрическая иллюстрация метода главных компонент

Каждая переменная  $x_i$  ( $i = \overline{1, n}$ ) при измерении получает значение  $x_{ij}$  ( $j = \overline{1, p}$ ). Назовем вектор  $Ob_j = (x_{1j}, \dots, x_{nj})$  ( $j = \overline{1, p}$ ) -  $j$ -тым объектом.

Результаты наблюдений представляются в виде матрицы "объекты-признаки":

Номер объекта $k$	Объекты	Признаки
1	$Ob_1$	$x_{11}x_{21} \dots x_{n1}$
2	$Ob_1$	$x_{12}x_{22} \dots x_{n2}$
$\vdots$	$\vdots$	$\dots$
$p$	$Ob_p$	$x_{1p}x_{2p} \dots x_{np}$

Каждой переменной  $x_i$  сопоставим в  $N$ -мерном  $\mathbb{R}^n$  пространстве ось  $x_i$ . Тогда объекты – это точки в пространстве  $\mathbb{R}^n$ . Геометрически



поиск компонент означает, что для построения первой компоненты берется в  $R^n$  прямая, проходящая через центр координат и облако рассеяния объектов (данных).

В основном процедура выделения главных компонент подобна вращению, максимизирующему дисперсию (варимакс) исходного пространства переменных. Цель вращения заключается в максимизации дисперсии "новой" переменной  $f_1$  и минимизации разброса вокруг нее.

### 7.3 Сколько главных компонент выделять?

На этот вопрос отвечает:

- Критерий Кайзера:  $\lambda_i > 1$

- Критерий "каменстой осыпи" Кэттелла:  $\lambda_i$ , где график не похож на горизонтальную ось.

## 7.4 Статистическая значимость компонентного анализа

При рассмотрении предполагалась известной корреляционная матрица:

$$R_x = \left\| \frac{k_{ij}}{\sqrt{Dx_i} \sqrt{Dx_j}} \right\|, \quad k_{ij} = M(x_i x_j)$$

На практике главные компоненты оцениваются по выборочным ковариационной и корреляционной матрице. Для проверки значимости корреляционной матрицы  $R_x$  применяют критерий Бартлета:

$$\chi^2 = -\left[p - \frac{1}{6}(2n + 5)\right] * \ln |R_x|, \quad |R_x| = \prod_{i=1}^n \lambda_i$$

который распределен по  $\chi^2$  с  $\nu = (n - 1)/2$  степенями свободы. Нулевая гипотеза  $H_0$  о том, что корреляционная матрица является незначимой, отвергается, и принимается альтернативная  $H_1$ , если  $\chi^2 > (\chi^2)_{\gamma, \alpha}^{tab}$ . Если при проверке гипотезы отвергается значимость всей корреляционной матрицы, то нахождение главных компонент не имеет смысла.

После выделения  $m$ -компонент возникает вопрос: значимо ли различие между оставшимися главными компонентами: Критерий Бартлета

$$\chi^2 = -\left(p - \frac{1}{6}(2n + 5) - \frac{2}{3}\right) * \ln R_{n-m}$$

Эта случайная величина приближенно имеет  $\chi^2$  распределение. Причем:

$$R_{n-m} = |R_x| \prod_{i=1}^m \lambda_i \left( \frac{n - \sum_{i=1}^n \lambda_i}{n - m} \right)^{-(n-m)}$$

Но о различии между оставшимися главными компонентами, являющемся незначимым, отвергается и  $H_1$ , говорящая о значимости различия, принимается, если  $\chi^2 > (\chi^2)_{\gamma, \alpha}^{tab}$ , где  $\alpha$  – уровень значимости.

## 7.5 Пример компонентного анализа

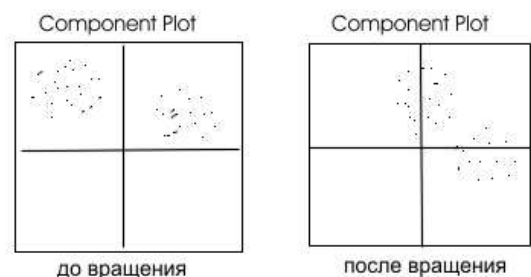
В результате опроса использовали 15 признаков  $x_1 - x_{15}$ , каждый по 7-балльной шкале. Результаты анализа в таблицах: (Таблица 3), (Таблица 4)

Таблица 3: Объединенная суммарная дисперия

Номер объекта $k$	Объекты	Доля
34.31%	$x_1$	$\lambda_1 = 5.146$
12.97	$x_2$	$\lambda_2 = 1.945$
9.433%	$x_3$	$\lambda_3 = 1.415$
6.60%	$x_4$	$\lambda_4 = 0.990$
$\vdots$	$\vdots$	$\dots$
1.20%	$x_{15}$	$\lambda_{15} = 0.181$

Таблица 4: Повернутая матрица факторных нагрузок

	1	2	3
$x_1$	-4.666	0.628	-1.191
$x_2$	-1.141	0.657	-.215
$x_3$	0.327	-0.153	0.711
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_{15}$	0.725	-4.882E-02	0.144



## Анализ данных(продолжение)

### 8 Факторный анализ

Пусть дано  $n$  наблюдаемых случайных величин  $x_1, \dots, x_n$ , описываемых исследуемое явление. Предположим, что пытаясь более полно описать явление, мы переборщили с переменными и в действительности они сводятся к меньшему числу величин  $f_1, \dots, f_m$ ,  $m \ll n$ .

Поставим задачу выявить величины  $f_i$  по  $x_1, \dots, x_n$ . Это задача факторного анализа. При этом  $f_1, \dots, f_m$  называются



факторами.

Факторы  $f_1, \dots, f_m$  выделяются таким образом, чтобы наилучшим образом описать корреляционную матрицу для  $x_1, \dots, x_n$ .

## 8.1 Основная модель факторного анализа

Факторы  $f_1, \dots, f_m$  ищутся из уравнений:

$$x_i = \sum_{k=1}^m l_{ik} F_k \quad (1)$$

Величины  $l_{ik}$  называются переменными,  $l_{ik}$  – нагрузками  $k$ -го фактора на  $i$ -ой переменной,  $e_i$  – остатками, представляющими источники отклонений, действующие только на  $x_i$ .

Предполагаем, что  $x_i, e_i, f_i$  центрированы, т.е.

$$Mx_i = Me_i = 0, \quad i = \overline{1, n}, \quad Mf_k = 0, \quad k = \overline{1, m}$$

и для ковариационных матриц

$$M[(f_i - Mf_i)(f_j - Mf_j)] = M[f_i f_j] = \delta_{ij}$$

$$M[(e_i - Me_i)(e_j - Me_j)] = M[e_i e_j] = v_i \delta_{ij}$$

$$M[(f_i - Mf_i)(e_j - Me_j)] = M[f_i e_j] = 0$$

где  $\delta_{ij} = 1$  при  $i = j$ , и  $\delta_{ij} = 0$  при  $i \neq j$ , т.е. факторы и остатки считаем некоррелированными. Из уравнения (1) следует, что

$$l_{ij} = M[x_i f_j]$$

т.е. нагрузки – это корреляции между переменными и факторами, и

$$k_{ii} = \sum_{k=1}^m l_{ik}^2 + v_i, \quad k_{ij} = \sum_{k=1}^m l_{ik} l_{jk}, \quad i \neq j \quad (2)$$

где  $k_{ij} = M[(x_i - Mx_i)(x_j - Mx_j)] = M[x_i x_j]$

Величины  $l_{ik}$  называются  $i$ -ми общностями, а  $v_i$  - специфичностями.

Общность представляет собой часть дисперсии  $k_{ij}$  переменных, объясненную факторами, специфичность - часть, не объясненную факторами дисперсию.

В матричной записи эти уравнения вычисляются так:

$$K_x = LL' + V \quad (3)$$

Уравнение (3) называется основной моделью факторного анализа.  
Т.к. ковариационная матрица  $K_x$  связана с корреляционной

$$R_x = \left\| \frac{k_{ij}}{\sqrt{Dx_i}\sqrt{Dy_i}} \right\|, \quad K_x = \|k_{ij}\| \quad (4)$$

то уравнения (3)-(4) дают представление корреляционной матрицы через матрицу нагрузок  $L = \|l_{ik}\|$  и диагональную матрицу  $V = \|v_i\delta_{ij}\|$  с положительными собственными элементами  $v_i$ . Отметим, что матрица  $LL' = L'IL$ , где  $I$ -единичная. Иначе говоря, матрица  $LL'$  - это результат преобразования матрицы  $I$  к новому базису. Поэтому его собственные числа положительны.

## 8.2 Решение уравнений основной модели

Нахождение факторов в основном сводится к нахождению нагрузок  $l_{ik}$  и ошибок  $e_i$  по заданным  $x_1, \dots, x_n$ . Решаем уравнение (3) вместо (1). Слева  $\frac{n(n+1)}{2}$  известных величин – элементы ковариационной матрицы, справа  $n(m+1)$  неизвестных  $l_{ik}, v_i$ .

- при  $m+1 > \frac{n+1}{2}$  нельзя найти единственное решение;
- при  $m+1 \leq \frac{n+1}{2}$  число неизвестных меньше числа уравнений и задача становится неразрешима.

Следовательно, число факторов не может быть небольшим, но их и недолжно быть слишком много, иначе теряется смысл задачи по отысканию факторов.

Факторное решение не единственно. Перепишем (1) в виде

$$X = LF + E$$

Пусть  $U$ -ортогональная матрица. Для нее  $UU' = I$ , т.к.

$$X = LUU'F + E = (LU)(U'F) = \hat{L}\hat{F} + E \quad (5)$$

, где  $\hat{F} = U'F$  – новые факторы, а  $\hat{L} = LU$  – новые нагрузки. При этом основная модель (3) инвариантна относительно (5)

$$K_x = LL' + V = LUU'L' + V = (LU)(LU)' + V = \hat{L}\hat{L}' + V$$

Таким образом, факторы определяются с точностью до вращения  $F \rightarrow U'F$  в  $m$ -мерном факторном пространстве.

### 8.3 Метод максимального правдоподобия

Функция максимального правдоподобия берется в виде:

$$F = -\frac{1}{2} \ln |K_x| - \frac{1}{2}(p-1)Sp(A_x K_x^{-1}), \quad K_x = LL' + V$$

,где  $A_x = |||a_{ij}||$  – выборочная ковариационная матрица. Объем выборки каждой переменной  $x_i$  равен  $p$ . Для оценки неизвестных переменных  $l_{ij}, v_i$  полагается, что они максимизируют функцию

$$F(l_{ij}, v_i) \Rightarrow \frac{\delta F}{\delta l_{ij}} = \frac{\delta F}{\delta v_i} = 0, \quad i = \overline{1, n}, \quad j = \overline{1, m} \quad (6)$$

Матрица  $L$  ищется так, чтобы  $(n \times m)$  матрица  $J = L'V(-1)L$  была диагональной. Предполагаем также, что диагональные элементы  $J$  расположены в порядке убывания (по величине).

Решение системы ( 6) имеет вид:

$$v_i = a_{ii} - \sum_{j=1}^m l_{ij}^2 \quad (7)$$

$$L' = J^{-1}L'V^{-1}(A_x - V) \quad (8)$$

Из ( 8) следует, что матрица

$$H = L'V^{-1}(A_x - V)V^{-1}L$$

равна  $J^2$  и поэтому диагональные уравнения ( 7) и ( 8) решаем методом итераций.

Зададим  $L_{(1)}, V_{(1)}$  и приближаем по итерационной схеме:

$$\begin{cases} L'_{(N+1)} = J^{-1}(L'_{(N)}V_{(N)}^{-1}A_x - L'_{(N)}) \\ v_{(N+1)} = a_{ii} - \sum_{j=1}^m l_{ij(N)}^2 \end{cases} \quad (9)$$

В действительности итерации производятся по более сложной схеме, чем ( 9). Усложнение связано с включением в расчетную схему значений, полученных в предыдущей строке.

Обозначим  $l_i, l'_i$   $i$ -е строки  $L, L'$ .

Вычислим

$$\begin{aligned} \omega'_1 &= l_{1(1)}V_{(1)}^{-1} \\ v'_1 &= \omega'_1 A_x - l'_{1(1)} \end{aligned}$$

и положительное число  $h_1 = v'_1 \omega_1$  – 1-й элемент матрицы  $H$ .  
Тогда по схеме (9) 1-я строка следующего приближения  $L'_{(2)}$  матрицы  $L'$  равна

$$l'_{1(2)} = \frac{1}{\sqrt{h_1}} v'_1 = \frac{1}{\sqrt{h_1}} (l'_{1(1)} V_{(1)}^{-1} A_x - l'_{1(1)})$$

Если имеется второй фактор,  $m \geq 2$ , то

$$l'_{2(2)} = \frac{1}{\sqrt{h_2}} v'_2 = \frac{1}{\sqrt{h_2}} (l'_{2(1)} V_{(1)}^{-1} A_x - l'_{2(1)} - j_{21} l'_{1(2)})$$

где

$$j_{21} = \omega'_2 l_{1(2)}, \quad \omega'_2 = l'_{2(1)} V_{(1)}^{-1} \\ h_2 = v'_2 \omega_2, \quad v'_2 = \omega'_2 A_x - l'_{2(1)} - j_{21} l'_{1(2)}$$

Для третьего фактора

$$l'_{3(2)} = \frac{1}{\sqrt{h_3}} v'_3 = \frac{1}{\sqrt{h_3}} (l'_{3(1)} V_{(1)}^{-1} A_x - l'_{3(1)} - j_{31} l'_{1(2)} - j_{32} l'_{2(2)})$$

где

$$j_{31} = \omega'_3 l_{1(2)}, \quad j_{32} = \omega'_3 l_{2(2)}, \quad \omega'_3 = l'_{3(1)} V_{(1)}^{-1} \\ h_3 = v'_3 \omega_3, \quad v'_3 = \omega'_3 A_x - l'_{3(1)} - j_{31} l'_{1(2)} - j_{32} l'_{2(2)}$$

и т.д. до  $m$ .

Найденные  $L_{(2)}, V_{(2)}$  используются для нахождения 3-го приближения.

После остановки итерационного процесса получают оценки  $\widehat{L}, \widehat{V}$  – оценки  $L, V$ ,  $\widehat{K}_x$  – оценка ковариационной матрицы  $K_x$ .

На практике метод сходится.

## 8.4 Оценка числа факторов

Сколько нужно факторов?

$H_0$ : требуется  $M$  факторов.

Предлагается приближенный  $\chi^2$  - критерий:

$$\chi^2 = \left[ (p-1) - \frac{1}{6}(2n+5) - \frac{2}{3}m \right] * \sum_{i < j} \frac{(a_{ij} - \widehat{c}_i)^2}{\widehat{v}_i \widehat{v}_j} \quad (10)$$

Здесь величины  $\widehat{c}_i, \widehat{v}$  – результат оценки ковариационной матрицы, матрицы нагрузок  $L$  и матрицы  $V$ , т.е. результат решения основной

факторной модели.

Зададим уровень значимости  $\alpha$  и вычислим  $\chi^2_{\text{выч}}$  по (10). Сравниваем с  $\chi^2_{\text{табл}}$  с числом тепеней свободы  $\frac{(n-m)^2-(n+m)}{2}$ . Если  $\chi^2_{\text{выч}} < \chi^2_{\text{табл}}$ , то гипотеза  $H_0$  принимается. В противном случае гипотеза  $H_0$  отвергается и в факторной модели требуется как минимум  $m + 1$  фактор.

## 8.5 Сравнение с компонентным анализом

Также строится таблица наблюдений "объекты-признаки".

Номер объекта $k$	Объекты	Признаки
1	$Ob_1$	$x_{11}x_{21} \dots x_{n1}$
2	$Ob_1$	$x_{12}x_{22} \dots x_{n2}$
$\vdots$	$\vdots$	$\dots$
p	$Ob_p$	$x_{1p}x_{2p} \dots x_{np}$

Главные компоненты являются линейными функциями от наблюдаемых переменных, а факторы не выражаются через комбинацию исходных признаков. Главные компоненты не объясняют корреляцию между переменными, а факторы для этого и отыскиваются.

## Анализ данных(продолжение). Регрессионный анализ.

## 9 Регрессионный анализ.

*Литература: Малхорта Н.К. Маркетинговые исследования. Практическое руководство. -М. Вильямс, 2002*

*Регрессионный анализ* - метод установления формы и изучения связей между метрической зависимой переменной и одной или несколькими независимыми переменными. Используют в случаях:

1. Действительно ли независимые переменные обуславливают значимую вариацию зависимой переменной: действительно ли эти переменные взаимосвязаны?
2. В какой степени вариацию зависимой переменной можно объяснить независимыми переменными. Теснота связи?

3. Определить форму связи: математическое уравнение, описывающее зависимость между зависимой и независимой переменными.
4. Предсказать значения зависимой переменной.
5. Контролировать другие независимые переменные при определении вкладов конкретной переменной.

Парная регрессия (bivariate regression) - это метод установления математической зависимости между одной метрической зависимой переменной и одной метрической независимой переменной.

Статистики, связанные с парным регрессионным анализом:

- Модель парной регрессии  $Y_i = \beta_0 + \beta_1 X_i + e_i$ , где  $Y$  - зависимая переменная,  $X$  - независимая переменная,  $\beta_0$  - точка пересечения прямой регрессии с осью  $OY$ ,  $\beta_1$  - тангенс угла наклона прямой,  $e_i$  - остаточный член, связанный с  $i$ -м наблюдением.
- Коэффициент детерминации. Тесноту связи измеряют коэффициентом детерминации  $r^2$ , который  $\in [0, 1]$  и указывает долю полной вариации  $Y$ , которая обуславливается вариацией  $X$ .
- Вычисляемое значение  $\hat{Y}$ .  $\hat{Y} = a + bx$ ,  $\hat{Y}$  - вычисляемое значение  $Y_i$ ,  $a, b$  - вычисленные оценки  $\beta_0$  и  $\beta_1$ .  $b$  - ненормированный коэффициент регрессии.
- Поле корреляции - графическое представление точек с координатами  $X$  и  $Y$ .
- Стандартная ошибка уравнения регрессии.  $S_{EE}$  - стандартное отклонение фактического значения  $Y$  от теоретического значения  $\hat{Y}$ .
- Стандартная ошибка коэффициента регрессии. Стандартное отклонение в обозначении  $SE_b$  называют стандартной ошибкой.
- Нормированный коэффициент регрессии. Обозначение  $\beta_1$
- Сумма квадратов ошибок  $\sum e_i^2$
- t-статистика с  $n - 2$  степенями свободы можно использовать для проверки  $H_0$ : между  $X$  и  $Y$  не существует линейной зависимости, т.е.  $\beta_1 = 0$ , где  $t = \frac{b}{SE_b}$

### Алгоритм парного регрессионного анализа.

1. Построение поля корреляции.
2. Формулирование общей модели.
3. Вычисление параметров.
4. Вычисление нормированного коэффициента регрессии.
5. Проверка значимости.
6. Определение тесноты и значимости связи.
7. Проверка точности предсказания.
8. Анализ остаточных членов.
9. Перекрестная проверка модели.

### Поле корреляции

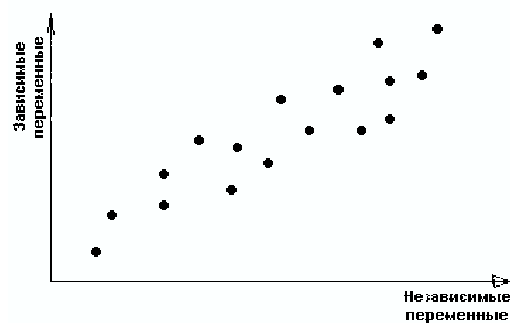


Рис. 1: Поле корреляции

*Метод наименьших квадратов* - метод расчета коэффициентов уравнения линейной регрессии на основе минимизации расстояния всех точек поля корреляции(рис.1 ) (по вертикали) от графика регрессии(рис. 2).

Определение параметров уравнения регрессии:

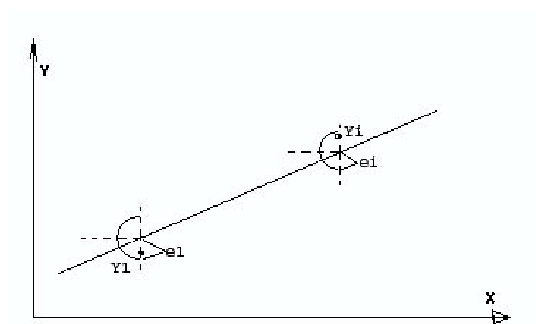


Рис. 2: График линейной регрессии

$\hat{Y}_i = a + b\hat{x}_i$  – теоретическое значение  $Y_i$ ,  $a, b$  – и вычисленные значения  $\beta_0$  и  $\beta_1$ .

$$b = \frac{COV_{xy}}{S_x^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}} = \frac{\sum_{i=1}^n X_i Y_i - \bar{X}\bar{Y}}{\sum_{i=1}^n X_i^2 - n\bar{X}^2}$$

$$a = \bar{Y} - b\bar{X}$$

### Нормированный коэффициент регрессии

*Нормирование* – процедура преобразования координат после которой значение среднего равно 0, дисперсии равно 1.

После нормирования отрезок отсечения от  $OY = 0$ .

Коэффициент регрессии обозначается "бетакоэффициент  $B_{xy}$  и он равен коэффициенту корреляции  $B_{xy} = B_{yx} = r_{xy}$ .

$$B_{xy} = b_{yx} \left( \frac{S_x}{S_y} \right), \quad b_{yx} - \text{ненормированный коэффициент регрессии.}$$

### Проверка значимости

Гипотезы:  $H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0$

$$t = \frac{b}{SE_b} \quad \text{с } n-1 \text{ степенями свободы}$$



Теснота и значимость связи(рис. 3)

Тесноту связи определяют коэффициентом детерминации  $r^2$ ,  $r^2 \in [0, 1]$

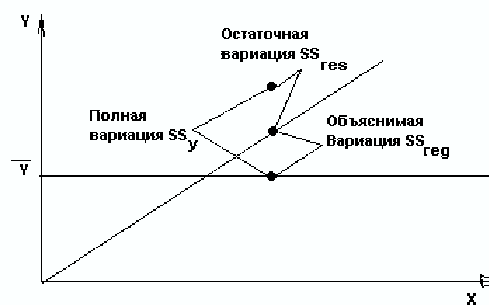


Рис. 3: Различные вариции Y

$$SS_y = SS_{reg} + SS_{res}, \quad \text{где} \quad SS_y = \sum_{i=1}^n (Y_i - \bar{Y})^2,$$

$$SS_{reg} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2, \quad SS_{res} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Тесноту вычисляют:

$$r^2 = \frac{SS_{reg}}{SS_y} = \frac{SS_y - SS_{res}}{SS_y}$$

Проверка значимости коэффициента детерминации

$$H_0 : R^2_{\text{совокупность}} = 0$$

$$H_1 : R^2_{\text{совокупность}} > 0$$

Критерий F-статистики:

$$F = \frac{SS_{reg}}{SS_{res}/(n-2)},$$

которая подчиняется F-распределению с 1 и  $n - 2$  степенями свободы.

#### Точность предсказания

Для оценки точности вычисления  $\hat{Y}$  полезно вычислить стандартную ошибку оценки уравнения регрессии  $SEE$ .

$$SEE = \sqrt{\frac{SS_{res}}{n - 2}}$$

Для  $k$  независимых переменных

$$SEE = \sqrt{\frac{SS_{res}}{n - k - 1}}$$

$SEE$  можно интерпретировать как вид среднего значения остатка или среднюю ошибку предсказания  $Y$ , исходя из уравнения регрессии.

Могут иметь место два случая предсказания: для заданного значения  $X_0$  для одного измерения или для всех измерений, где  $X = X_0$ . В обоих случаях  $Y = \hat{Y} = a + bX_0$

Однако стандартная ошибка для этих ситуаций разная, хотя в обоих случаях это функция  $SEE$ . Для больших выборок стандартная ошибка предсказания среднего значения  $Y$  равна  $\frac{SEE}{\sqrt{n}}$ , а ошибка предсказания отдельного значения  $Y$  равна  $SEE$ .

Следовательно, построение доверительного интервала варьируется в зависимости от того, необходимо построить оценку для одного значения или для среднего значения.

#### Предпосылки регрессионного анализа

Допущения:

1. Ошибочный член уравнения регрессии подчиняется закону нормального распределения. Для каждого определенного значения  $X$  распределение  $Y$  нормальное.
2. Среднее значение всех этих нормальных распределений  $Y$ , при заданном  $X$ , лежит на прямой линии с угловым коэффициентом  $b$ .
3. Среднее значение ошибочного члена равно 0.
4. Дисперсия ошибочного члена постоянна. Эта дисперсия не зависит от значений, принятых  $X$ .

5. Между ошибочными членами автокорреляция отсутствует. Значения ошибочных величин независимы между собой.

## 10 Множественная регрессия.

*Multiple regression* – статистический метод для построения зависимости между двумя и более независимыми переменными и зависимой переменной, выраженной с помощью интервальной или относительной шкалы. Общая форма модели множественной регрессии имеет вид:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + e$$

Модель оценивается уравнением

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + \dots + b_k X_k$$

Используем метод наименьших квадратов, который оценивает параметры таким образом, чтобы минимизировать суммарную ошибку  $SS_{res}$ . Этот процесс также максимизирует корреляцию между фактическим значением  $Y$  и предсказанным значением  $\hat{Y}$ .

Все предположения, которые используются для парной регрессии, применимы и для множественной.

### 10.1 Статистики, связанные с множественной регрессией.

- Скорректированный коэффициент множественной детерминации. Он корректируется с учетом числа независимых переменных и размера выборки, чтобы снизить влияние зависимости коэффициента детерминации от количества переменных.
- $\underline{R^2}$  - Теснота связи между переменными, измеряют, вводя квадратичный коэффициент множественной корреляции.
- F - критерий. Используется для проверки  $H_0$ : Коэффициент множественной детерминации в совокупности  $R_{сов}^2 = 0$  или  $H_0 : \beta_0 = \beta_1 = \dots = \beta_k = 0$ . Статистика для проверки гипотезы, подчиняется  $F$ -распределению с  $k$  и  $(n - k - 1)$  степенями свободы.
- Частный коэффициент регрессии  $b_i$  обозначает изменение в предсказанном значении  $\hat{Y}$  при изменении  $X_1$  на единицу, когда другие независимые переменные от  $X_2$  до  $X_k$  остаются неизменными.

## 10.2 Выполнение множественного регрессионного анализа.

Рассмотрим на примере двумерного случая:

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 \quad (*)$$

Где  $b_1$  - прирост  $Y$  при увеличении  $X_1$  на 1;  $b_2$  - прирост  $Y$  при увеличении  $X_2$  на 1.

Нормированные коэффициенты, имеющие среднее = 0 и дисперсию = 1, получаем:

$$B_1 = b_1 * \left( \frac{Sx_1}{S_y} \right), \quad \dots \quad B_k = b_k * \left( \frac{Sx_k}{S_y} \right)$$

Для вычисления  $a, b_1, \dots, b_k$  необходимо уравнение (\*) продифференцировать по всем  $X$  и приравнять их 0. Затем решить систему линейных уравнений.

## 10.3 Теснота связи.

Полную вариацию можно представить

$$SS_y = SS_{reg} + SS_{res}, \quad \text{где} \quad SS_y = \sum_{i=1}^n (Y_i - \bar{Y})^2,$$

$$SS_{reg} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2, \quad SS_{res} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Коэффициент множественной корреляции

$$R^2 = \frac{SS_{reg}}{SS_y}$$

$R^2$  не может быть меньше, чем самое высокое значение  $r^2$  любой отдельной независимой переменной с зависимой. Если независимые переменные не коррелированы, то значение  $R^2$  = сумме коэффициентов парной детерминации каждой независимой переменной с зависимой.

Но  $R^2$  зависит от количества переменных, следовательно, скорректированный

$$\overline{R^2} = R^2 - \frac{k(1 - R^2)}{n - k - 1}$$

## 10.4 Проверка значимости общего уравнения и частных коэффициентов регрессии.

$$H_0 : R^2_{\text{совокупности}} = 0 \sim H_0 : \beta_0 = \beta_1 = \dots = \beta_k = 0.$$

Общую проверку можно провести, используя  $F$ -статистику.

$$F = \frac{SS_{reg}/k}{SS_{res}/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)},$$

которая имеет  $F$ -распределение с  $k$  и  $(n-k-1)$  степенями свободы.

Если общую  $H_0$  отклонили, то то один или несколько частных коэффициентов регрессии в совокупности имеют значение, отличное от 0.

Проверка каждого проводится также как и в случае парной регрессии, используя  $t$ -статистику с  $(n-k-1)$  степенями свободы. Зачастую  $F$ -критерий проводят на основе вычисления частных  $F$ -критериев. Расчет общей  $SS_{reg}$  разлагается на компоненты по независимым переменным  $X_i : SSx_i$ .

Значимость частных коэффициентов регрессии для переменной  $\beta_i$  проверяют  $F$ -статистикой:

$$F = \frac{SSx_i/1}{SS_{res}/(n-k-1)}, \quad F - \text{расп. с 1 и } (n-k-1) \text{ степенями свободы}$$

## 11 Анализ остатков.

Остаток – это разность между наблюдаемым значением  $Y_i$  и теоретическим значением, предсказанным уравнением  $\hat{Y}_i$ . Значения остатков используют для вычисления некоторых статистик, связанных с регрессией.

Проведение анализа сделанных допущений. Допущение нормальности распределения ошибочного числа, проанализируем, построив гистограмму остатков. Визуальный осмотр можно подкрепить проверкой правила трех сигм:

в интервале  $\pm 1E$  должно лежать 68%; в  $\pm 2SE$  – 95%. Либо проведем анализ критерия Колмогорова-Смирнова.

Предположение о постоянном значении дисперсии ошибочного члена проанализируем, нанеся на график значения остатков в зависимости от вычисленных значений независимой переменной  $\hat{Y}_i$ . Если точки нанесены на график неупорядоченно, то дисперсия ошибочного члена – величина постоянная.



Рис. 4: пример

Пример:(рис. 4) Дисперсия – величина непостоянная, зависит от  $Y$ .

График зависимости остатков от времени или последовательности наблюдений прольет свет на допущение, что ошибочные члены некоррелированы.

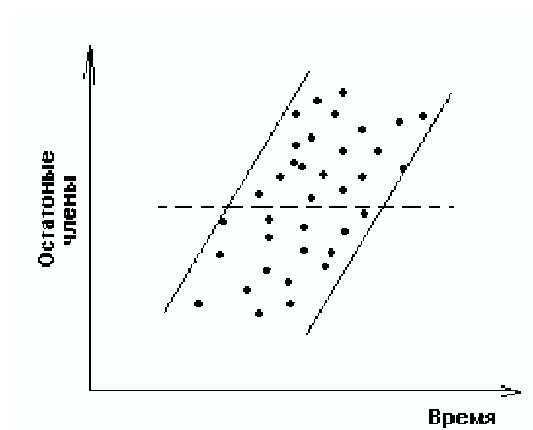


Рис. 5: пример

Пример:(рис. 5) Есть линейная зависимость значений остаточных членов от времени. Более формальную процедуру проверки корреляции между ошибочными членами дает критерий Дарбина-Уотсона.

Графическое изображение зависимости значений остатков от независимых переменных предоставляет доказательство о том, насколько подходит теоретическая модель регрессии. График должен показывать случайную форму расположения остаточных членов. Значения остатков располагаются случайным образом относительно одинаково вокруг 0.

Таким образом, анализ остатков позволяет глубже понять, как соответствие лежащим в основной регрессионной модели допущениям, так и соответствие регрессионной модели.

Пример:(рис. 6)Регрессионная модель удовлетворительного описания

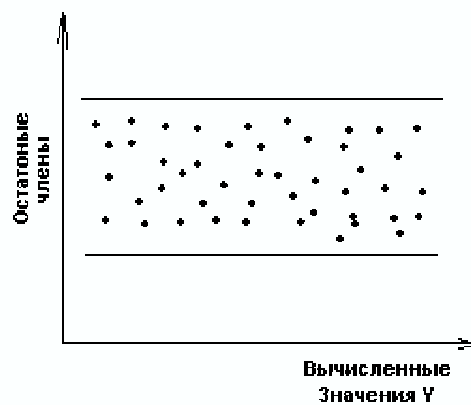


Рис. 6: пример

данных.

## 12 Пошаговая регрессия.

Регрессионная процедура, в которой предикторы по очереди вводят или выводят из уравнения регрессии. Подходы:

1. Прямое включение. Вводят предикторы по одному, если они удовлетворяют определенному F- критерию. Порядок определяется вкладом в общую вариацию.
2. Обратная пошаговая регрессия. Исключают по одной.
3. Пошаговый подход. С включением одновременно исключают переменные, не удовлетворяющие критерию.

### 13 Мультиколлинеарность.

Состояние очень высокой степени корреляции между независимыми переменными. Проблемы:

1. Частные коэффициенты нельзя точно определить.
2. Величины частных коэффициентов меняются от выборки.
3. Трудно оценить относительную важность независимой переменной.
4. Предикторы могут быть некорректно введены или выведены.

Относительная важность предикторов:

1. Статистическая значимость. Если частные коэффициенты не являются значимыми, то переменную не считают важной.
2. Квадрат линейного коэффициента корреляции  $r^2$ .
3. Квадрат частного коэффициента корреляции  $R^2_{yx_i x_j x_k}$ .
4. Квадрат частичного коэффициента корреляции. Измеряя  $R^2$  при вводе переменной.
5. Показатели, основанные на нормированных коэффициентах и "бетакоэффициентах.



## 14 Вырезка из лекций.

### 14.1 Оценка корреляционного момента системы двух случайных величин $X$ и $Y$ .

$$\widehat{K}_{xy} = \frac{1}{N-1} \sum_{i=1}^N (X_i - \widehat{x})(Y_i - \widehat{y})$$

а их коэффициент корреляции:

$$\widehat{\rho}_{xy} = \frac{\widehat{K}_{xy}}{\widehat{S}_x \widehat{S}_y}$$

Реккурентная формула:

$$\widehat{K}_{xy} = \frac{1}{N-1} \left[ \sum_{i=1}^N X_i Y_i - N \widehat{y}_N \widehat{x}_N \right]$$

Оценкой вероятности состояния  $r$  служит частота его появления:

$$\widehat{p}_r = \frac{N_r}{N}$$

где  $N_r$  – количество реализаций (из общего числа  $N$ ), в которых наблюдалось интересующее нас событие. В непрерывных моделях  $N$  заменяют на общее время  $T$ , а  $N_r$  на время наблюдения  $T_r$  состояния  $r$ .

$$\widehat{p}_r = \frac{T_r}{T}$$

$\widehat{p}_r$  – это выборочное среднее из-за влияния начальных условий будет смещенной оценкой истинного значения. Традиционным методом борьбы является отбрасывание данных за период вхождения в стационарный режим, однако не ясно, сколько длится переходный период.

Все перечисленные оценки состоятельные, при неограниченном возрастании  $N$  (длительности  $T$ ) сходятся по вероятности к исходным параметрам. Однако при любом конечном  $N$  или  $T$  оценки случайны. Вопрос: какова прогрешность этой оценки?

Дисперсия оценки  $\widehat{\nu}_r$ -го начального момента:

$$D[\widehat{\nu}_r] = \frac{\widehat{\nu}_{2r} - \widehat{\nu}_r^2}{N}$$

(в правую часть входят теоретические значения соответствующих моментов). В частности

$$D[\hat{\nu}_x] = \frac{D_x}{N}$$

## 14.2 Классический подход к интервальным оценкам.

Имея оценку и ее дисперсию, можно перейти к построению доверительного интервала. Практические требования к оценкам формулируются в терминах их точности и надежности. Под точностью понимается половина  $\delta$  длины доверительного интервала, а под надежностью – вероятность  $P$  того, истинное значение параметра окажется принадлежащим упомянутому интервалу (доверительная вероятность). Увеличение  $\delta$  ведет к уменьшению  $P$  и наоборот.

Пусть определяется среднее время  $\omega$  ожидания начала обслуживания, причем дисперсия времени ожидания равна  $\sigma_W^2$ . При характерной для имитационного моделирования кратности наблюдений  $N$  разность  $(\hat{\omega} - \omega)$  можно на основании теоремы А.М. Ляпунова считать распределенной нормально с дисперсией  $\frac{\sigma_W^2}{N}$ . Следовательно доверительная вероятность:

$$P\{|\hat{\omega} - \omega| \leq \delta\} = \Phi\left(\frac{\delta\sqrt{n}}{\sigma_W\sqrt{2}}\right).$$

где  $\Phi(\cdot)$  - функция Лапласа.

Переходя к обратной функции, имеем

$$\Phi^{-1}P = \frac{\delta\sqrt{n}}{\sigma_W\sqrt{2}},$$

откуда требуемое число наблюдений

$$N \geq 2[\Phi^{-1}(P)]^2 \left(\frac{\sigma_W}{\delta}\right)^2 = k(P) \left(\frac{\sigma_W}{\delta}\right)^2 \quad (11)$$

Коэффициент  $k(P)$  выбирается из таблицы(таблица 5)

Таблица 5: Коэффициент  $k(P)$

$P$	0.900	0.950	0.970	0.980	0.990	0.995	0.999
$k(P)$	2.69	3.84	4.71	5.43	6.66	7.90	10.82

Следовательно, необходимое число испытаний обратно пропорционально квадрату допустимой погрешности и резко возрастает с повышением доверительной вероятности.

Во всех формулах вместо теоретических значений среднеквадратичного отклонения  $\sigma_W$  практически приходится пользоваться его статистической оценкой. Поэтому важны формулы рекуррентного вычисления дисперсии. Возможен двухэтапный эксперимент:

1. определяем грубо  $\widehat{\sigma}_w$ , после чего согласно формуле 11 определяем  $N$ .
2. Проводим необходимое количество дополнительных испытаний.

В случае малого количества наблюдений ( $\sim$  несколько десятков) доверительные интервалы для нормального распределения величин строят с помощью распределения Стьюдента. Введем следующую величину

$$T = \frac{\widehat{\omega} - \omega}{\widehat{s}},$$

где

$$\widehat{s} = \sqrt{\frac{1}{N(N-1)} \sum_{i=1}^N (W_i - \widehat{\omega})^2}.$$

Можно доказать, что следующая величина  $T$  имеет плотность распределения

$$s_N(t) = \frac{\Gamma\left(\frac{N}{2}\right)}{\sqrt{(n-1)\pi}\Gamma\left(\frac{N-1}{2}\right)} * \left(1 + \frac{t^2}{N-1}\right)^{-\frac{N}{2}},$$

зависящую только от аргумента  $t$  и числа наблюдений  $N$ . Следовательно:

$$P\{|\widehat{\omega} - \omega| \leq t\widehat{s}\} = 2 \int_0^t s_N(t) dt.$$

Функция в правой части и ей обратная табулированы. С их помощью можно вычислить доверительные вероятности и доверительные интервалы. При оценке вероятностей в формулу (11) вместо  $\sigma$  следует подставить  $\sqrt{p(1-p)} \Rightarrow$  можно (11) переписать:

$$N \geq k(P) \cdot \frac{p(1-p)}{\delta^2}.$$

При определении вероятностей обычно задаются их относительной точностью, то есть считают  $\delta = \varepsilon p \Rightarrow$  для малых вероятностей  $1 - p \approx 1$  можно переписать:

$$N \geq \frac{k(P)}{p\varepsilon^2} .$$

Итак, необходимое число испытаний обратно пропорционально искомой вероятности и квадрату допустимой относительной погрешности. Например, для определения вероятности порядка с относительной погрешностью в 1% и доверительной вероятностью 0,98 должно быть проведено:

$$N > 5.43 \cdot 10^6 \cdot 10^4 \approx 5 \cdot 10^{10}$$

испытаний. Это обстоятельство значительно затрудняет непосредственную оценку вероятностей разных событий методом имитационного моделирования.

## **Динамические системы.**

### **15 Одномерные динамические системы с одним оператором и их бифуркации.**

$$\frac{\delta f}{\delta t} = f(x, \mu)$$

$f(x, \mu) = 0$ ,  $x$  – переменная,  $\mu$  – параметр,  $t$  – время.

$x = x(\mu)$  находится однозначно, как решения уравнения  $f(x(\mu), \mu)$

Его можно решить, если выполняется теорема "о неявных функциях".

Если она выполняется то  $f(x, \mu)$  – одна кривая (см. рис. 7).

Однако если условие теоремы не выполняется, то

$$f(x_0, \mu_0) = \frac{\delta f}{\delta x}(x_0, \mu_0) = \frac{\delta f}{\delta \mu}(x_0, \mu_0) = 0 \quad - \text{лежит как минимум на двух кривых. (12)}$$

И говорят, что в точке  $(x_0, \mu_0)$  происходит бифуркация, т.е. раздвоение кривых, равновесия (см. рис. 8).

Условие (см. формулу 12) – условие существования точки бифуркации.

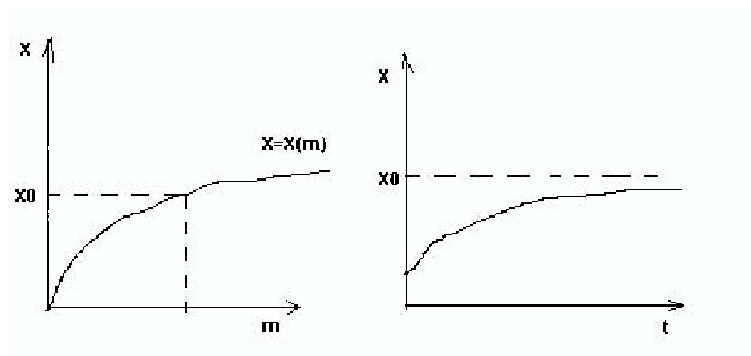


Рис. 7: одна кривая

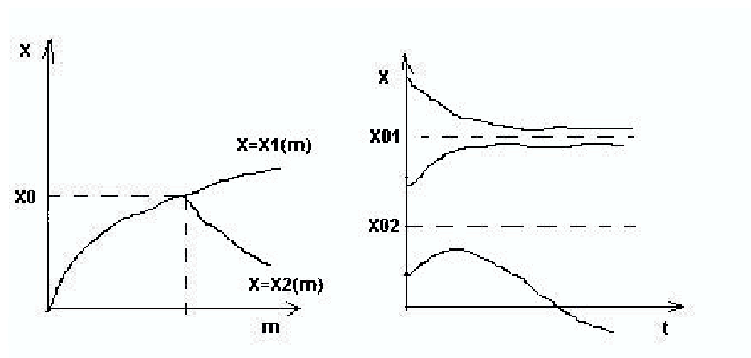


Рис. 8: несколько кривых

## 16 Типы точек бифуркации.

### 16.1 Двойная точка бифуркации.

Двойная точка бифуркации – это точка бифуркации, через которую проходят две гладкие кривые.

$$x = x_1(\mu), \quad x = x_2(\mu) \quad \text{или} \quad \mu = \mu_1(x), \quad \mu = \mu_2(x)$$

Для нее выполняется неравенство:

$$D = \left[ f_{x\mu}'' \right]^2 (x_0, \mu_0) - f_{xx}''(x_0, \mu_0) \cdot f_{\mu\mu}''(x_0, \mu_0) > 0$$

## 16.2 Экстремальная (двойная) точка бифуркации.

Экстремальная (двойная) точка бифуркации – это точка двойной бифуркации, у которой одна ветвь кривой равновесия имеет касательную параллельную оси  $X$ , и лежит по одну сторону этой касательной (см. рис. 9).

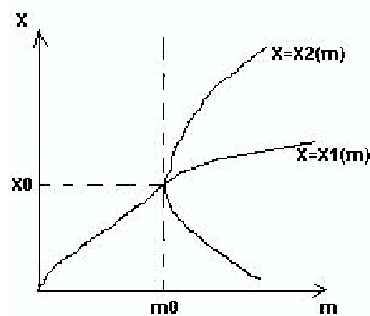


Рис. 9: Экстремальная точка бифуркации

## 16.3 Точка возврата.

Точка возврата – это точка двойной бифуркации, для которой ветви имеют общую касательную в точке  $(x_0, \mu_0)$  (см. рис. 10).

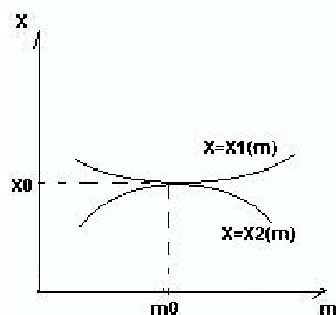


Рис. 10: Точка возврата

## 16.4 Тройная точка бифуркации.

Тройная точка бифуркации – это точка бифуркации, для которой имеем три ветви кривой равновесия. Выпняется равенство:

$$D = \left[ f''_{x\mu} \right]^2 (x_0, \mu_0) - f''_{xx}(x_0, \mu_0) \cdot f''_{\mu\mu}(x_0, \mu_0) = 0$$

## 17 Примеры динамических систем.

### 17.1 Модель "Хищник - Жертва".

Модель "Хищник - Жертва" (модель Лотка-Вольтера):

$P$  – "Хищник" (predator),  $V$  – "Жертва" (victim).

$$\frac{\delta P}{\delta t} = (b_p - d_p) \cdot P + \alpha PV, \quad \frac{\delta V}{\delta t} = (b_v - d_v) \cdot P - \alpha PV$$

$b$  – "born",  $d$  – "die",  $\alpha$  – вероятность поедания "хищником" "жертву".

Пример:

$b_p = 0, \quad b_v = 0, \quad d_p = 0.4, \quad d_v = 0, \quad \alpha = 0.005$  (см. рис. 11)

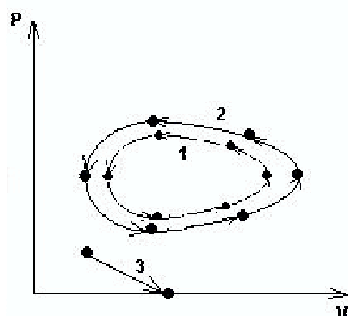


Рис. 11: "Хищник - Жертва"

### 17.2 Three traffic food web.

$R$  – resource (ресурсы),  $C$  – consumer (потребитель),  $P$  – predator (хищник) (см. рисю 12).

То есть  $C$  – consumer поедает только  $R$  – resource, а  $P$ -predator поедает и тех и других, с определенной вероятностью.

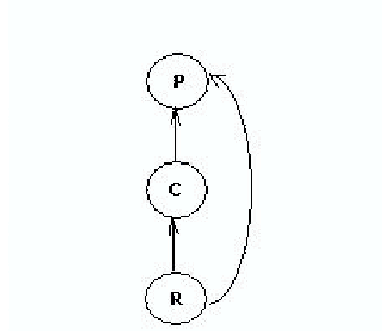


Рис. 12: Three traffic food web

Сама модель:

$$\begin{aligned}\frac{\delta R}{\delta t} &= R \left[ r \left( 1 - \frac{R}{K} \right) - \frac{\lambda_{RC}C}{1 + h_{RC}\lambda_{RC}R} - \frac{U_{RP}\lambda_{RP}P}{1 + U_{RP}\lambda_{RP}h_{RP}R + U_{CP}\lambda_{CP}h_{CP}C} \right], \\ \frac{\delta C}{\delta t} &= C \left[ \frac{e_{RC}\lambda_{RC}R}{1 + h_{RC}\lambda_{RC}R} - \frac{U_{CP}\lambda_{CP}P}{1 + U_{RP}\lambda_{RP}h_{RP}R + U_{CP}\lambda_{CP}h_{CP}C} - m_C \right], \\ \frac{\delta P}{\delta t} &= P \left[ \frac{U_{RP}\lambda_{RP}e_{RP}R + U_{CP}\lambda_{CP}e_{CP}C}{1 + U_{RP}\lambda_{RP}h_{RP}R + U_{CP}\lambda_{CP}h_{CP}C} - m_P \right],\end{aligned}$$

где  $r$  – максимальная скорость роста ресурсов  $[0.3h^{-1}]$ ,

$K$  – "вместимость" ресурсов в окружающую среду  $\left[0..20\frac{mgC}{L}\right]$ ,

$\lambda_R$  – скорость поиска потребителем ресурсов  $\left[0.037\frac{L}{mg\cdot h}\right]$ ,

$\lambda_{RP}$  – скорость поиска хищником ресурсов  $\left[0.025\frac{L}{mg\cdot h}\right]$ ,

$\lambda_P$  – скорость поиска хищником потребителей  $\left[0.025\frac{L}{mg\cdot h}\right]$ ,

$U_{RP}$  – вероятность нападения хищника на ресурс  $[0 \dots 1]$ ,

$U_{CP}$  – вероятность нападения хищника на потребителя  $[0 \dots 1]$ ,

$h_{RC}$  – время, необходимое потребителю для переваривания ресурсов  $[3h]$ ,

$h_{RP}$  – время, необходимое хищнику для переваривания ресурсов  $[4h]$ ,

$h_{CP}$  – время, необходимое хищнику для переваривания потребителя  $[4h]$ ,

$e_{RC}$  – эффективность усвоения ресурса потребителем  $[0.6]$ ,

$e_{RP}$  – эффективность усвоения ресурса хищником  $[0.36]$ ,

$e_{CP}$  – эффективность усвоения потребителя хищником  $[0.6]$ ,



$m_C$  – скорость отмирания потребителя  $[0.03h^{-1}]$ ,

$m_P$  – скорость отмирания хищника  $[0.03h^{-1}]$ .

$f(R) = R \left[ r \left( 1 - \frac{R}{K} \right) \right]$  – функция прироста ресурсов, ограничивает систему (см. рис. 13)

Пример:

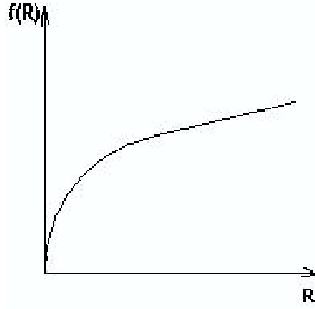


Рис. 13: функция прироста ресурсов  $f(R)$

Сделаем упрощения  $h_{RC} = 0$ ,  $U_{RP} = U_{CP} = 1$ .

Промоделировав эту систему, нашли две точки равновесия:

1) когда вымерли хищники:

$$E_{RC} = \left( \frac{m_C}{\lambda_{RC} e_{RC}}; \quad \frac{r(K\lambda_{RC} e_{RC} - m_C)}{K\lambda_{RC}^2 e_{RC}}; \quad 0 \right).$$

2) когда вымерли потребители:

$$E_{RP} = \left( \frac{m_P}{\lambda_{RP}(e_{RP} - m_P h_{RP})}; \quad 0; \quad \frac{r e_{RP}(R\lambda_{RP}(e_{RP} - m_P h_{RP}) - m_P)}{K\lambda_{RP}^2(e_{RP} - h_{RP} m_P)^2} \right).$$

Получаем, что при определенных сочетаниях  $e_{RP}$  и  $K$  (при  $e_{CP} = 0.3$ ), могут получаться различные ситуации, когда выжившими остаются  $R, P, C$  (см. рис. 14).

И как следствие при определенном сочетании величин  $R, P, C$  могут быть различные пищевые цепочки (см. рис. 15).

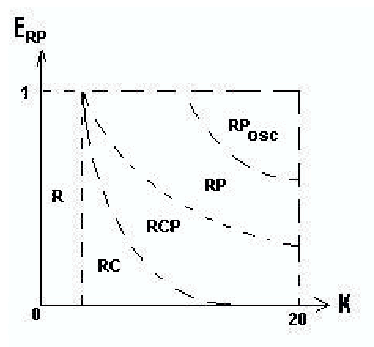


Рис. 14:  $e_{RP}$  и  $K$

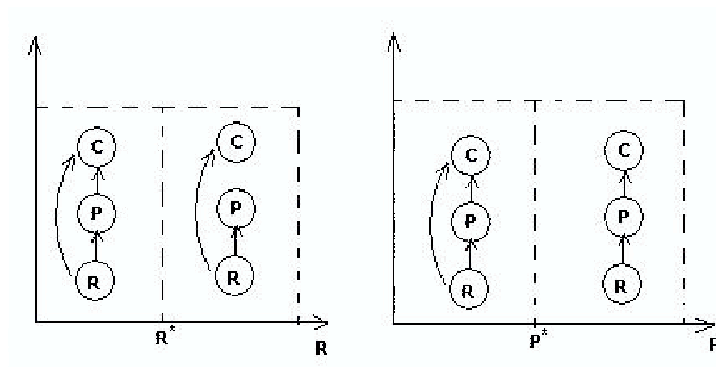


Рис. 15: различные пищевые цепочки

## Случайные числа.

### 18 Генерация случайных чисел.

#### 18.1 Схема получения случайных чисел.

Для получения случайных чисел, требуемого закона распределения необходимо:

1. Формирование случайных чисел  $U_i$ , равномерно распределенных на  $[0 \dots 1]$ ,  $i = \overline{1 \dots n}$
2. Программный переход от  $U_i$  к случайному числу  $X_j$ , имеющему требуемое распределение  $F(x)$ .

## 18.2 Генерация равномерно распределенных случайных чисел.

$N$  идентичных триггеров со счетным входом, каждый из которых фиксирует независимый поток импульсов от лампы или радиоаппаратуры. Такой поток – простейший, при интенсивности  $\lambda$  и интервалами отсчета  $\Delta t$ , и в интервале  $[0 \dots 1]$  различаются между собой на  $e^{-2\lambda\Delta t}$ . При достаточно больших  $\lambda\Delta t$  дисбаланс между этими значениями незаметен.

## 18.3 Метод обратных функций.

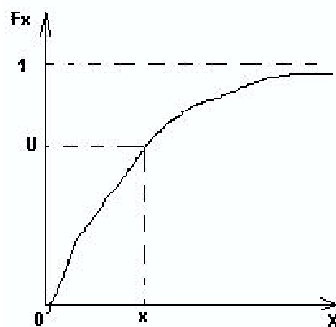


Рис. 16:  $F(x)$  и  $x$

$U$  – случайная величина, равномерно распределенная на  $[0 \dots 1]$ ,  $F[X] = U \Rightarrow X = F^{-1}(U)$  (см. рис. 16).

1. Экспоненциальное распределение:

$$F(x) = 1 - e^{-\lambda x}, \quad 1 - e^{-\lambda X} = U \Rightarrow X = -\frac{\ln(1 - U)}{\lambda}.$$

Замена:  $(1-U)$  так же равномерно распределена на  $[0 \dots 1]$  как и  $U$ . Представим  $U$  в виде  $U = m * 2^p$ . Следовательно  $\ln U$  мы можем вычислить как  $\ln U = -\ln 2(p - 2.6797 + 4.0391m - 1.3594m^2)$  с погрешностью не превышающей 0.001.

Правда нужно учитывать что все равномерно сгенерированные числа на  $[0,1)$  имеют ненулевую вероятность, что  $U = 0$ .

2. Распределение Релея:

$$F(x) = 1 - e^{-\frac{x^2}{2\sigma^2}} \Rightarrow X = \sigma\sqrt{-2\ln U}$$

3. Распределение Вейбула:

$$F(x) = 1 - e^{-\frac{x^k}{T}} \Rightarrow X = \sigma\sqrt[k]{-T\ln(1-U)}$$

4. Распределение Коши, с медианой  $\mu$  и масштабом  $\sigma$ :

$$f(x) = \frac{\sigma}{\pi(\sigma^2 + (x - \mu)^2)}, \quad F(X) = \frac{1}{2} - \frac{1}{\pi} \cdot \arctg\left(\frac{X - \mu}{\sigma}\right) \Rightarrow X = \sigma \cdot \operatorname{tg}\left(\pi(U - \frac{1}{2})\right)$$

5. Логистическое распределение:

$$F(x) = \frac{1}{1 + e^{-\frac{x-a}{b}}} \Rightarrow X = a - b \cdot \ln\left(\frac{1}{U} - 1\right).$$

Пример:

$$a = 0, \quad b = 1 : F(X) = \frac{1}{1 + e^{-x}} \Rightarrow X = \ln\left(\frac{U}{1-U}\right)$$

Нахождение максимума

$$F(X) = U = F^n(X) \Rightarrow X = F^{-1}(U^{\frac{1}{n}}),$$

и минимума

$$T_n(X) = 1 - (1 - F(X))^n = U \Rightarrow X = F^{-1}(1 - U^{\frac{1}{n}}).$$

6. Треугольное распределение:

$$F(X) = \frac{2}{a} \left(X - \frac{X^2}{2a}\right) \Rightarrow X = a(1 - \sqrt{1-U})$$

7. Распределение Паретто:

$$F(X) = 1 - \left(\frac{b}{x}\right)^2 \Rightarrow X = \frac{b}{\sqrt{1-U}}$$

## 18.4 Аппроксимация обратной функции распределения.

Для нахождения случайных чисел с заданным распределением необходимо решить уравнение, которое зачастую решается итерационно, в этом случае лучше использовать таблично-приближенный метод – Кусоно-линейная аппроксимация обратной функции:  
строится  $g(U) = F^{-1}(U)$  – кусоно-линейная функция.

1. (Авторы: Оде и Ивэнс) аппроксимация  $g(U)$ :

$$g(U) = t + \frac{A(t)}{B(t)}, \quad \text{где } t = \sqrt{-2 \ln(U)}$$

, где коэффициенты многочленов  $A(t)$  и  $B(t)$  при соответствующих степенях  $t$  даны в таблице. Применяется для  $10^{-20} \leq U < \frac{1}{2}$

2. Аппроксимация  $g(U)$ :

$$X = \frac{2.30753 + 0.27061 \cdot t}{0 + 0.99229 \cdot t + 0.04481 \cdot t^2}, \quad t = \sqrt{-2 \ln U'}, \quad U' = \max\{U; 1-U\}$$

с погрешностью  $3 \cdot 10^{-3}$ .

## 18.5 Дискретные распределения.

Для формирования дискретной случайной величины непрерывная функция распределения заменяется ступенчатой (см. таблицу 6).

Таблица 6: Дискретные распределения.

Распределение	Параметры	$P(x = i)$	Диапазон
Пуассона	$\lambda > 0$	$e^{-\lambda} \cdot \frac{\lambda^i}{i!}$	$i \geq 0$
Биномиальное	$n, p$	$C_n^i p^i (1-p)^{n-i}$	$i = 0..n$
Отрицательное биномиальное	$n \geq 1, p$	$C_{n+1+i}^i \cdot (1-p)^{n+i}$	$i \geq 0$
Логарифмическое	$0 < \theta < 1$	$-\frac{\theta^i}{i \cdot \ln(1-\theta)}$	$i \geq 1$
Геометрическое	$0 < p < 1$	$p(1-p)^{i-1}$	$i \geq 1$

## 18.6 Метод последовательных сравнений.

Метод последовательных сравнений является аналогом метода обратных функций, заключается в переборе  $X$ , пока не окажется:

$$F(X-1) = \sum_{i < X} p_i < U < \sum_{i \leq X} p_i = F(X)$$

, при этом:  $P(x = i) = F(i) - F(i - 1) = p_i$

Пример:

1. Распределение Пуассона:

ш.1  $X = 0, \quad b = e^{-\lambda}, \quad S = b.$

ш.2 Сформировать равномерно распределенную величину  $U$ .

ш.3 Пока  $U > S$

$$X = X + 1, \quad b = b \cdot \frac{\lambda}{X}, \quad S = S + b.$$

ш.4 Вернуть значение  $X$ .

2. Биномиальное распределение:

$$b_0 = (1 - p)^n, \quad b = b \cdot \frac{n - x}{x + 1} \cdot \frac{p}{1 - p}$$

3. Отрицательное биномиальное распределение:

$$b = b \cdot \frac{n + x}{x + 1} \cdot \frac{p}{1 - p}$$

4. Геометрическое распределение:  $b = b \cdot (1 - p)$

## 18.7 Моделирование случайной величины. Воспроизведение вероятностной схемы.

- Распределение Эрланга с параметрами  $\lambda$  и  $r$ :

$$X = \sum_{k=1}^r \left( -\frac{1}{\lambda} \ln U_k \right) = -\frac{1}{\lambda} \sum_{k=1}^r \ln U_k = -\frac{1}{\lambda} \ln \left( \prod_{k=1}^r U_k \right)$$

- Гамма-распределение:

ш.1 цикл

Сформировать равномерно распределенные случайные числа  $U_1$  и  $U_2$ .

$$\text{Вычислить } S_1 = U_1^{\frac{1}{\alpha}}, \quad S_2 = U_2^{\frac{1}{1-\alpha}}.$$

пока  $S_1 + S_2 \leq 1$

ш.2 Сформировать равномерно распределенно случайное число  $U_3$ .

ш.3

$$X = -\frac{S \ln U_3}{\lambda(S_1 + S_2)}$$

ш.4 Вернуть  $X$ .

- Нормальное распределение:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Известно, что сумма  $n$ -величин с математическим ожиданием  $Z^{(1)}$  и дисперсией  $d^{(1)}$  при  $n \rightarrow \infty$  стремится к нормальному распределению с параметрами  $Z = nZ^{(1)}, d = d^{(1)}$ .

Пример:

Пусть  $Z^{(1)} = \frac{1}{2}, d^{(1)} = \frac{1}{12}$

1.

$$X(n) = \sqrt{\frac{12}{n}} \left( \sum_{i=1}^n U_i - \frac{n}{2} \right), \quad X = \mu + \sigma X(n).$$

Но этот метод не является точным и используется другой:

2.

$$Y = X(5) - \frac{3(X(5)) - X^3(5)}{100}$$

Кроме того можно использовать другие методы. Например распределение Релея. Распределение Релея описывает радиальное отклонение от центра прицеливания при равной дисперсии по осям  $X$  и  $Y$ .

3. Метод Бокса и Мюллера:

$$Z_1 = W \cdot \cos(2\pi U_2), \quad Z_2 = W \cdot \sin(2\pi U_2), \quad W = \sqrt{-2 \ln U_1}.$$

4. Метод Марсилья:

ш.1 формировать равномерно распределенные случайные числа  $U_1$  и  $U_2$ .

ш.2  $S = U_1^2 + U_2^2$ .

ш.3 Если  $S > 1$ , то ш.1.

ш.4 Вычислить

$$W = \sqrt{-\frac{2 \ln S}{S}}.$$

ш.5 Выдать  $X = W \cdot U_1$ ,  $Y = W \cdot U_2$ .

## 18.8 Генерация пары коллерированных величин.

Если нам нужна пара случайных величин  $X$  и  $Y$  таких, что: коэффициент корреляции  $X$  и  $Y$  был равен  $\rho$ , и соответствующие математические ожидания  $x, y$  и дисперсии  $\sigma_x^2, \sigma_y^2$  то:

$$Y = aX + b + z, \text{ где } a = \frac{\rho\sigma_y}{\sigma_x}, \text{ } b = y - ax, \text{ и } z = \sqrt{1 - \rho^2}\sigma_y\eta,$$

где  $\eta$  – случайное число, распределенное по нормальному закону с параметрами  $\bar{\eta} = 0, \sigma_{\eta} = 1$ .

sectionМоделирование этнических систем.  
Л.Н. Гумилев: "Этногенез и биосфера земли."

Этнос – некая совокупность людей, которые ощущают себя как единое целое, не отделяющая себя от других элементов(см. рис. 17). Консорция



Рис. 17: Этнос

– временная организация людей, объединенных одной целью.

Гумилев вводит понятие пассионарность – избыток пассионарной энергии, подавляющей в человеке инстинкт самосохранения и способствующей совершению сверхнапряжения. Самосохранение – стремление человека сохранить свою жизнь и жизнь своих близких.



Гумилев подразделяет людей в этносе на три категории (см. рис. 18).

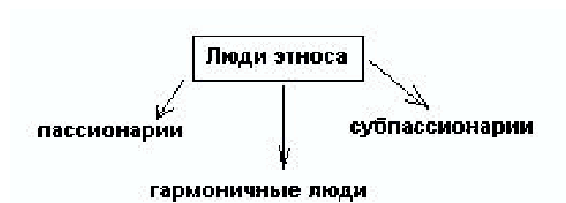


Рис. 18: Люди этноса

$$\text{Пассионарное напряжение} = \frac{\text{пассионарная энергия}}{\text{количество членов этноса}}$$

Гумилев разделяет процесс этногенеза на несколько стадий (см. рис. 19):

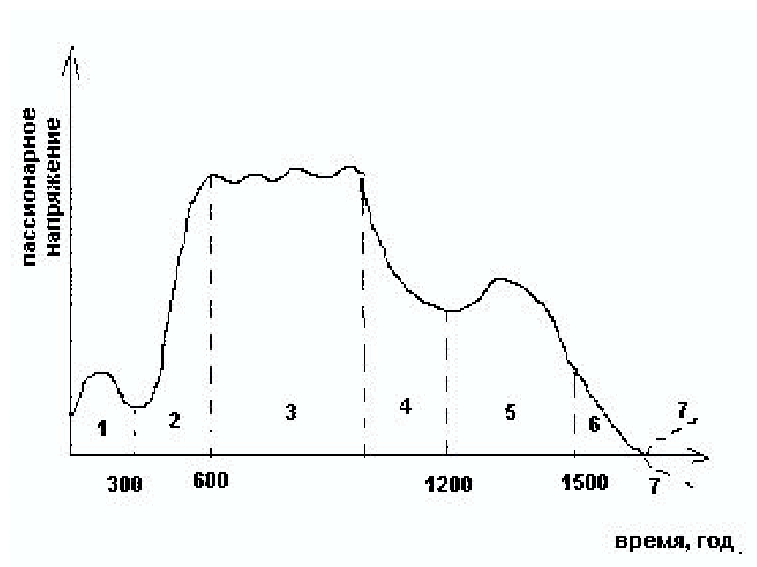


Рис. 19: Процесс этногенеза

1. Пассионарный толчок.  
Происходит некий толчок и появляется много пассионариев.
2. Фаза подъема.  
Происходит рост и подъем этноса. Людям этноса присуще стремление к новому.
3. Акматическая фаза.  
Во время этой фазы происходят войны, революции ...
4. Фаза надлома.  
Многих пассионариев убивают, высылают, в основном потому что этнос "устал" от напряжения.
5. Инерционная фаза.  
Размеренная жизнь, без войн. осподство гармоничных людей.
6. Фаза обскурации.  
Уничтожение этноса как системы.
7. Фаза регенерации(этнос возрождается в качестве другого) или реликта(существует, доживает, растворившись в других этносах).

Гумилев выделяет структуру этноса(см рис. 20):

1. Организация.  
Это ситема управления этносом(гос. строй, религиозные системы ...).
2. Наука и техника.  
(Достижения науки и техники ...).
3. Культура и искусство.
4. Ландшафт.  
Тот ареал, в котором находится этнос.

## 18.9 Модель этнической системы. Автор: А.К.Гуц

$$\overline{X} = X(P, H, S, O, T, C, L))$$

$$\frac{\delta \overline{X}}{\delta t} = F(x, t) \quad - \text{энергетическая формула.}$$

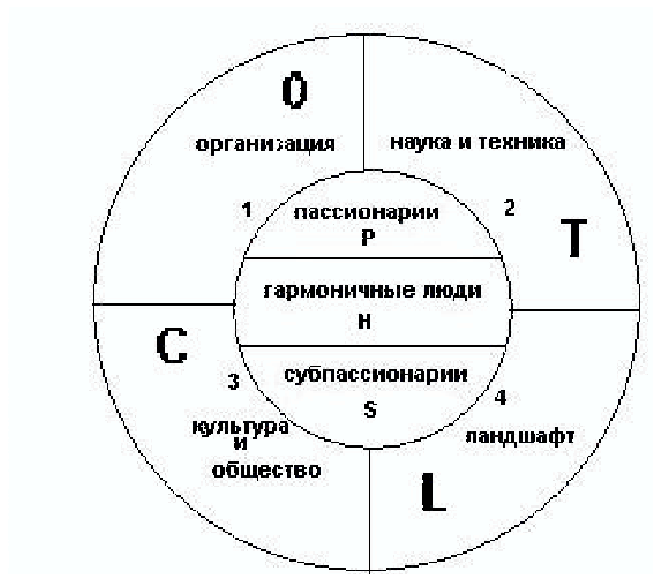


Рис. 20: Структура этноса

Пример:

$$\frac{\delta P}{\delta t} = \pi(t) - C_{PO}P - C_{PT}T, \quad \frac{\delta O}{\delta t} = C_{PO}P$$

Для моделирования этой системы нужно задать все параметры этой системы:  $C_{PO}, C_{PT} \dots$

$$\Pi = \frac{P + H + S}{N(t)} \quad \text{кривая пассинарного напряжения(см. рис. 21).}$$

### 18.10 Модель этнического поля.

Взаимодействие между членами этноса осуществляется согласно этническому полю(см. рис. 22).  $G$  – Область распространения этнического поля.

$u_i(x, y, t)$  – плотность этнического поля.

$U_i(t) = \int_G u_i(x, y, t) dx dy$  – пассионарная энергия.

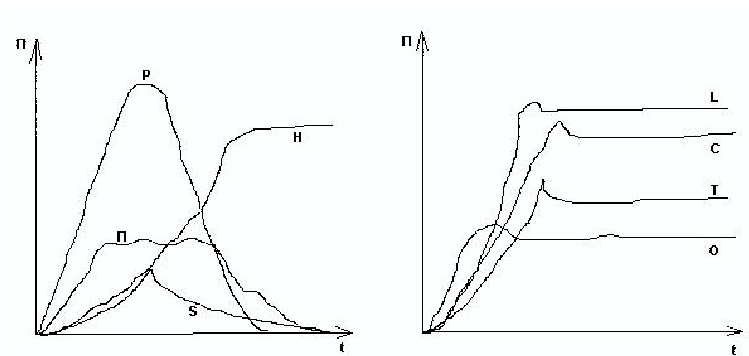


Рис. 21: Кривые пассионарного напряжения

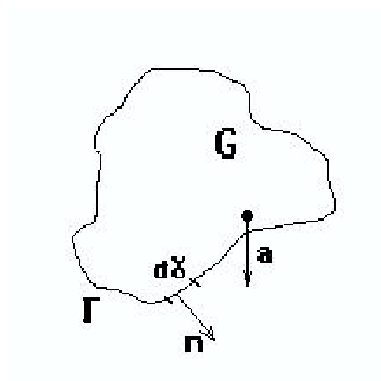


Рис. 22: распространение этнического поля

$$U_i(t_1) - U_i(t_1) = \int_{t_1}^{t_2} (R_i(t) + P_i(t) + T_i^+(t) + T_i^-(t) + K_i(t)) dt,$$

– изменение пассионарной энергии.

$R_i = \oint_{\Gamma} \xi_i(x, y, t) \frac{\delta u_i}{\delta n}(x, y, t) d\gamma$  – процесс равномерного распространения энергии.

$\xi_i(x, y, t)$  – свойство ландшафта.

$$P_i(t) = \oint_{\Gamma} -(\vec{a}_i, \vec{n}) u_i(x, y, t) d\gamma, \quad \text{– куда этнос хочет переместиться.}$$

$$\vec{a}_i = -\nabla \varphi_i(x, y, t)$$

$$T_i^+(t) = \int \int_G \beta_i^+(x, y, t) u_i(x, y, t) dx dy$$

– приток пассивной энергии за счет индукции.

$$T_i^-(t) = - \int \int_G \beta_i^-(x, y, t) u_i(x, y, t) dx dy$$

– убывание энергии.

$$K_i = - \int \int_G \sum_{j=1}^k \gamma_{ij}(x, y, t) \cdot u_i(x, y, t) \cdot u_j(x, y, t) dx dy$$

– поток энергии, "всасываемой" при столкновении с другими этносами.

Подставив все эти формулы в начальное уравнение и продифференцировав, мы получим:

$$\frac{\delta \varphi}{\delta t} = \xi_i \Delta u_i + \nabla \xi_i \nabla u_i + \nabla \varphi_i \nabla u_i + \nabla \varphi_i u_i + u_i \cdot \left( \beta_i^+ + \beta_i^- - \sum_{j=1}^k \gamma_{ij} u_j \right), \quad i = \overline{1..k}.$$

Где

$$\varphi(x, y, t) = \frac{\lambda_2(t)}{2\mu(t)} e^{-\mu(t)((x_0(t)-x)^2 + (y_0(t)-y)^2)} + C(t)$$

$$\xi(x, y) = I_\Omega(\zeta_\xi \cdot l_w)(x, y)$$

$\zeta_\xi$  – функция, описывающая зависимость коэффициента  $\xi$  от типа ландшафта.  $l_w$  – функция задающая тип ландшафта в узле сетки.

$$\beta_i^+(t) = \max\{0, \beta^0 - \beta^1(t - T^0)\},$$

$T^0$  – момент времени,  $\beta^0, \beta^1$  – константы.

$$\beta_i^-(x, y) = I_\Omega(\xi_\beta \cdot l_w)(x, y)$$

$I_\Omega$  – оператор интерполяции по области  $\Omega$ .  $\xi_\beta$  – определяют потерю энергии от типа ландшафта.