МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ КЫРГЫЗСКОЙ РЕСПУБЛИКИ

КЫРГЫЗСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМ.И.РАЗЗАКОВА

ФАКУЛЬТЕТ ТРАНСПОРТА И МАШИНОСТРОЕНИЯ

КАФЕДРА «АВТОТРАНСПОРТА»

ПРИМЕНЕНИЕ ОДНОФАКТОРНЫХ РЕГРЕССИОННО-КОРРЕЛЯЦИОННЫХ УРАВНЕНИЙ ДЛЯ РЕШЕНИЯ ЗАДАЧ СТАТИСТИЧЕСКОГО ИССЛЕДОВАНИЯ

Методическое пособие для практических занятий по курсу ОСНОВЫ НАУЧНЫХ ИССЛЕДОВАНИЙ предназначены для студентов специальностей 552101. 01 Автомобили и автомобильное хозяйство и 552101.02. Эксплуатация и обслуживание транспортных и технологических машин и оборудования

БИШКЕК – 2010

«Рассмотрено»
На заседании кафедры
«Автотранспорт»
Пртокол №6. от 25.01.2010 г.

«Одобрено»
Методическим советом факультета транспорта и машиностроения
Протокол №6 от 28.01.2010 г.

УДК 629.113.254.

Составитель БЕКЕТАЕВ О.Б.

Применение однофакторных регрессионно-корреляционных уравнений для решения задач статического исследования. / КГТУ им. И. Раззакова; сост**.** О.Б. Бекетаев. – Б.: ИЦ «Текник», 2010. – 60 с.

Предназначены для студентов специальностей 552101. 01 Автомобили автомобильное хозяйство и 552101.02. Эксплуатация и обслуживание транспортных и технологических машин и оборудования.

Рецензент к. т. н., доцент Абакиров С.А.

Понятие о регрессионно – корреляционном анализе

Занятие 1

Математическая статистика – это сложная и многогранная наука. В настоящее время каждый инженер должен владеть методами статистического исследования. Регрессионно-корреляционный анализ (один из разделов математической статистики) основывается на ретроспективном взгляде на прошлое, т.е. на зарегистрированных результатах статистических наблюдений за прошедшее время.

Это позволяет получать, так называемые уравнения регрессии, предоставляющие собой зависимость условного среднего $M\left[y \middle/ x_1, x_2, \ldots x_n \right]$ (обозначаемое также $\hat{Y}_{расч}$ от связанных с ним факториальных признаков:

$$\hat{Y}_{расч} = M\left[y \middle/ x_1, x_2, \ldots, x_n \right] = \varphi(x_1, x_2, \ldots, x_n),$$

где $\hat{Y}_{расч}$ - условное расчетное среднее, (параметр оптимизации, функция отклика, результативный признак) и т.п.;

φ - аналитический вид функции;

x_1 - факторы, параметры или функциональные признаки, оказывающие влияние на функцию отклика.

Уравнения регрессии определяют математические зависимости между переменными физического процесса, поэтому их называют **математическими моделями**. Регрессионный анализ тесно связан с корреляционным анализом. При этом:

- основной задачей регриссионого анализа является установление вида функции φ, связывающей параметр оптимизации с факторами, оказывающими влияние на функцию отклика;
- основной задачей корреляционного анализа является определение тесноты связи, т.е. тесноты зависимости функции отклика от связанных с ней факторов.

Указанная характеристика определяется с помощью корреляционного момента связи κ_{yx} и коэффициента корреляции r_{yx}, с помощью матрицы моментов связи и матрицы коэффициентов корреляции.

В результате проведения эксперимента при однофакторном анализе получают поле опытных точек (рис.1)

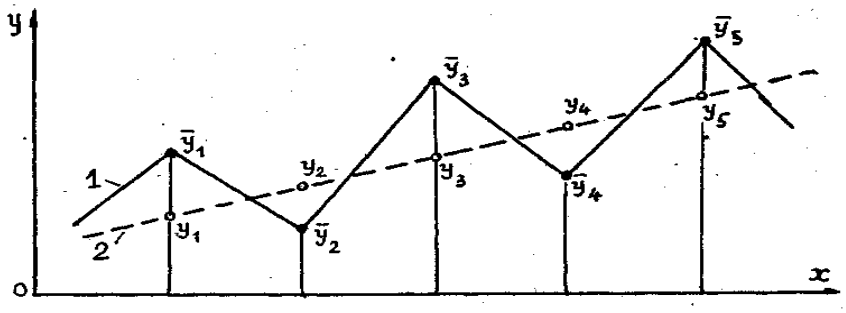

Рис.1. Поле опытных точек однофакторной регрессионной зависимости: 1- опытная линия регрессионной зависимости средних значений функции отклика $У_{ср}$ от фактора x; 2- расчетная (теоретическая) линия регрессии, представленная в виде прямой

При проведении и статистической обработке экспериментов устанавливают следующий порядок:

1. Если будет проводиться активный эксперимент, то производится выбор математической модели, с помощью которой предполагается описывать изучаемый процесс. При отсутствии сведений о виде модели, то в начале избирается линейная модель в соответствии с которой намечается центр плана и уровни факторов;
2. Проверяется воспроизводимость эксперимента;
3. Вычисляют коэффициенты модели;
4. Производится дискриминация (сравнение) различных моделей и выбирается лучшая;
5. Отобранная модель проверяется на адекватность;
6. Строится доверительный коридор и раструб разброса среднего результата по каждому из факторов;
7. Вычисляют простые (парные) коэффициенты корреляции;
8. Проверяют значимость вычисленных коэффициентов корреляции;
9. Вычисляют частные (парциальные) коэффициенты корреляции и на основе этого строят диаграмму влияния каждого из факторов на исследуемую функцию отклика.

Линейная модель

(Вывод формул для определения коэффициентов однофакторной модели).

Если рассматривается однофакторная регрессионная зависимость, то, как уже отмечалось выше, результаты наблюдений могут быть представлены в виде поля опытных точек (рис.1.)

Уравнение однофакторной регрессионной зависимости записывается так:

$$\hat{Y}_{расч} = B_{\circ} x_{\circ} + B_1 x_1 \qquad (1)$$

где $\hat{Y}_{расч}$ - среднее расчетное значение функции отклика;

x_1 - фактор, оказывающий влияние на функцию отклика;

B_1 - коэффициент при факторе x_1;

$B_\circ$ - начальная ордината;

$x_\circ$ - фиктивное переменное, равное единице.

Очевидно, что, при выравнивании поля опытных точек прямой линией 2(рис.1), она должна занимать такое положение на плоскости χoy, чтобы сумма квадратов отклонений расчетной прямой относительно опытных точек была бы минимальной, т.е чтобы:

$$U = \sum_{i=1}^{n} \left(Y_{i расч} - \overline{Y}_{i опытн}\right)^2 = \sum_{i=1}^{n} \left(B_\circ x_{\circ i} - B_1 x_{1i} - \overline{Y}_i\right)^2 \to \min, \quad (2)$$

где $\overline{Y}_{i опытн}$ - среднее опытное значение функции отклика, полученное по сечениям графика для опытной линии регрессии;

$Y_{i расч}$ - среднее расчетное значение функции отклика, рассчитанное по сечениям для прямой линии;

U - функция, называемая невязкой:

i – номер сечения графика $(i = 1,2,3,...,k)$

Требование (2) называют методом наименьших квадратов.

Дифференцируя указанную функцию U (невязку) по переменным $B_\circ$ и B_1, получаем:

$$\frac{\partial U}{\partial B_\circ} = 2\left(B_\circ x_{\circ 1} - B_1 x_{11} - \overline{Y}_1\right)x_{\circ 1} + ... + 2\left(B_\circ x_{\circ n} - B_1 x_{1n} - \overline{Y}_n\right)x_{\circ n} = 0;$$

$$\frac{\partial U}{\partial B_\circ} = 2\left(B_\circ x_{\circ 1} - B_1 x_{11} - \overline{Y}_1\right)x_{11} + ... + 2\left(B_\circ x_{\circ n} - B_1 x_{1n} - \overline{Y}_n\right)x_{1n} = 0.$$

Раскрывая скобки и группируя переменные получаем систему двух уравнений с двумя неизвестными, называемые системой нормальных уравнений.

$$\left.\begin{aligned} B_\circ \sum_{i=1}^{n} x_{\circ i} x_{\circ i} + B_1 \sum_{i=1}^{n} x_{\circ i} x_{1i} &= \sum_{i=1}^{n} x_{\circ i} \overline{Y}_i; \\ B_\circ \sum_{i=1}^{n} x_{1i} x_{\circ i} + B_1 \sum_{i=1}^{n} x_{1i} x_{1i} &= \sum_{i=1}^{n} x_{1i} \overline{Y}_i. \end{aligned}\right\} \quad (3)$$

Разделив левую и правую части нормальных уравнений на число испытаний ***n,*** получаем:

$$B_{\circ}\cdot 1+B_1\bar{x}_1=\bar{Y};\qquad B_{\circ}\cdot\bar{x}_1+B_1\alpha_2(x)=\alpha_{11}(x,\ y)\quad (4)$$

где $\bar{x}_1=\sum_{i=1}^{n}x_{1i}/n$ - опытное общее среднее арифметическое фактора;

$\bar{y}=\sum_{i=1}^{n}y_i/n$ - опытное общее среднее арифметическое функции отклика;

$\alpha_2(x)=\sum_{i=1}^{n}x_{1i}^{\,2}/n$ - второй начальный момент фактора;

$\alpha_{11}(x,y)=\sum_{i=1}^{n}x_iy_i/n$ - второй смешанный начальный момент.

Решая полученную систему нормальных уравнений, например, методом Крамера, получаем выражения для определения искомых коэффициентов математической модели

$$B_1=\frac{\Delta B_1}{\Delta\Gamma\Lambda}=\begin{Vmatrix}1 & \bar{y}\\ \bar{x}_1 & \alpha_{11(x,y)}\end{Vmatrix}\Big/\begin{Vmatrix}1 & \bar{x}_1\\ \bar{x}_1 & \alpha_2(x)\end{Vmatrix}=\frac{\alpha_{11}(x,y)-\bar{x}_1\bar{y}}{\alpha_2(x)-(\bar{x}_1)^2}=\frac{K_{yx}}{D[x]}\quad (5)$$

Непосредственно из равенства (4) получаем:

$$B_{\circ}=\bar{Y}-B_1\bar{x}_1\qquad (6)$$

Занятие 2

Проиллюстрируем на примере порядок получения расчетного уравнения однофакторной регрессионной зависимости при условии, что обработке подлежат пассивные эксперименты.

Пример 1. Исследуется тяжесть травм имеющих место на одном из автомобильно-производственном предприятии по годам в течение пяти лет. Статистическими наблюдениями были зафиксированы следующие значения тяжести травм (табл.1)

Таблица 1

№ строки	*Уопытн*		Факториальный признак χ				m_{y_i}
			$\chi=1$ (1976г.)	$\chi=2$ (1977г.)	$\chi=3$ (1978г.)	$\chi=4$ (1979г.)	
1	Тяжесть травм (число дней – бюллетеней, выданных в течение года)	12 11 10 9 8 7 6 5 4 3	 2 1 2	 2 1 2	 2 1 2	 2 2 2	2 1 2 2 3 3 2 2 2 2
2	Число выданных бюллетеней по годам n_{x_j}		5	5	5	6	
3	Среднее значение функции отклика по годам $\bar{y}_j$		1	7	8	4	
4	Дисперсии функции отклика по годам (по сечениям) $S^2\{y\}_j$		1	1	1	0,8	

Примечание. Дисперсии, вычисленные по сечениям графика или по точкам факторного пространства, называют построчечными дисперсиями.

Из таблицы видно, что по мере течения времени тяжесть травм быстро уменьшается.

Требуется:

1. Построить график опытной регрессионной зависимости тяжести травм от времени.
2. Проверить воспроизводимость эксперимента;
3. Аппроксимировать опытную линию регрессии прямолинейной зависимостью;
4. Аппроксимировать опытную линию регрессии графиком, отвечающим показательной функции;
5. Произвести дискриминацию математических моделей, отвечающих линейной зависимости и регрессии, отвечающей показательной функции;
6. Проверить отобранную математическую модель на адекватность;
7. Спрогнозировать тяжесть травм на последующий год.

Решение. Строим график опытной линии регрессии.

1. Вычисляем опытные средние значения функции отклика по годам пятилетки (по сечениям) по формуле:

$$\overline{Y}j = \sum_{i=1}^{n} Y_{ij} \cdot m_{y_i} \Big/ n_{x_j}$$

где n_{x_j} - число наблюдений в сечении;

m_{y_i} - число повторений наблюдений.

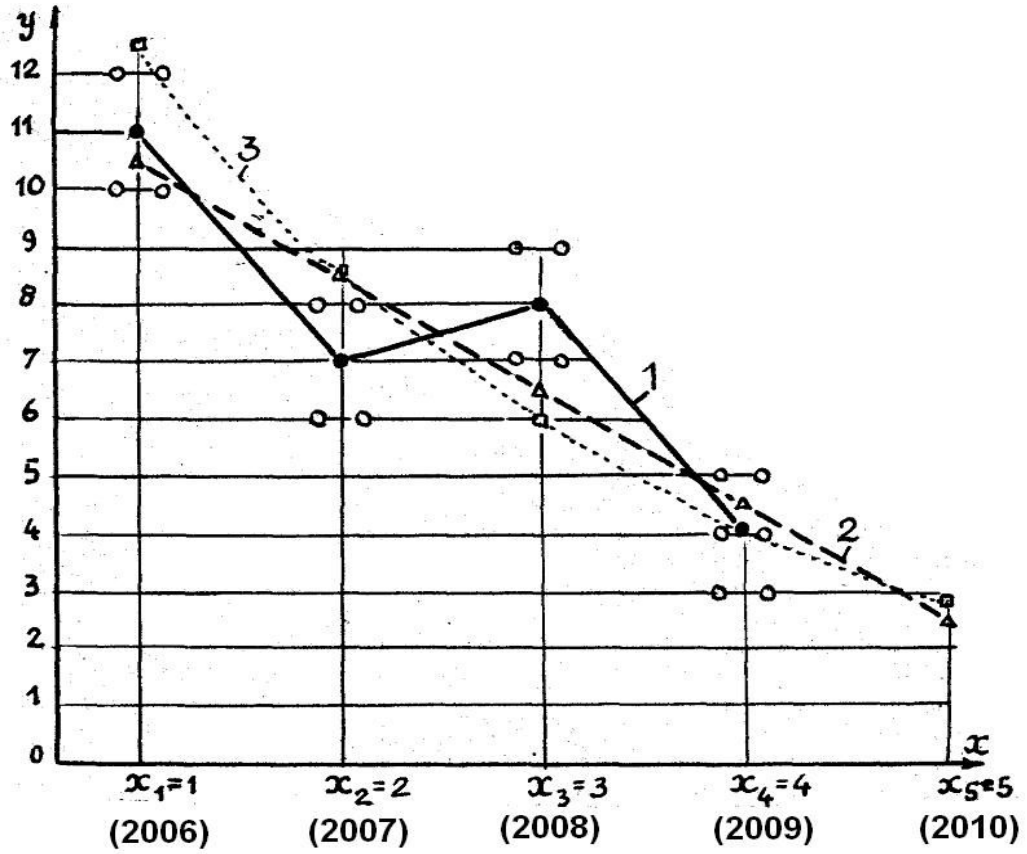

Рис. 2. Графики опытной линии регрессии (1); линии регрессии отвечающей прямолинейной зависимости (2); линии регрессии отвечающей показательной зависимости (3); 0 - опытные точки

Для первого сечения получаем

$$\overline{Y}_1 = \sum_{j=1}^{3} Y_{ij} \cdot m_{y_i} \Big/ n_{xj} = \frac{10 \cdot 2 + 11 \cdot 1 + 12 \cdot 2}{5} = 11$$

Аналогичное для последующих сечений (табл. 1).

2. На основе вычисленных опытных средних значений функции отклика по сечениям (по годам) строим график опытной линии регрессии (рис. 2).

Теперь необходимо проверить воспроизводимость эксперимента. В связи с этим напомним кратко сущность статистической оценки гипотез.

Краткие сведения об оценке статистических гипотез

При обработке опытных данных часто возникают задачи проверки ряда гипотез, например:
- проверяются гипотезы о принадлежности опытных данных (гистограмм) к нормальному закону, показательному закону и другим вероятностным законам;
- проверяются гипотезы об однородности двух дисперсии (например, для дискриминации двух математических моделей);
- проверяются гипотезы об однородности нескольких дисперсий (указанная процедура производится при проверке воспроизводимости опытов);
- проверяются гипотезы о статистической значимости коэффициентов полученной математической модели;
- проверяются гипотезы об адекватности полученной математической модели;
- проверяются гипотезы о согласованности показаний группы экспертов и т.п.

При выдвижении и принятии указанных гипотез могут иметь место следующие четыре случая:

1. Гипотеза H_o верна и принимается.
2. Альтернативная ей гипотеза H_1 состоящая в том, что гипотеза H_o верна, но ошибочно отвергается. Этот случай является противоположным по отношению к первому. Возникающую при этом ошибку называют ошибкой первого рода, а вероятность ее появления называют уровнем значимости и обозначают α.
3. Гипотеза Но неверна и отвергается.
4. Гипотеза Но неверна, но ошибочно принимается. Возникающую при этом ошибку называют ошибкой второго рода, а вероятность ее появления обозначают β.

Для решения отмеченных задач различными исследователями (например: Пирсоном, Колмогоровым, Кохреном, Бартлетом, Стьюдентом, Фишером, и др.) были предложены соответствующие критерии и заранее, при заданном уровне значимости: $\alpha = 0{,}1; \alpha = 0{,}05; \alpha = 0{,}02; \alpha = 0{,}01; \alpha = 0{,}0027 и \alpha = 0{,}001$ и т.д. были подсчитаны и составлены таблицы, в которых помещены критические (табличные) значения указанных критериев.

При этом область возможных значений каждого из критериев делят на две части:
- область принятия гипотезы;
- область непринятия гипотезы (так называемая критическая область), которая для различных критериев может быть левосторонней или правосторонней (см. рис. 3).

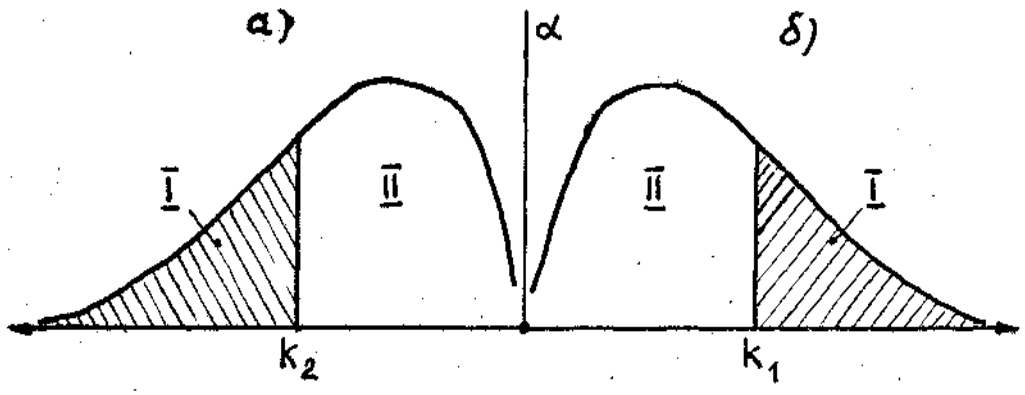

Рис.3. Левосторонняя (а) и правосторонняя (б) критические области, где К1 и К2 – критические точки или квантили: I – критическая область непринятия гипотезы и II- область принятия гипотезы

Для некоторых критериев применяются правосторонние критические области (см. табл. приложения 1), а для других левосторонние см. табл приложения 1).

При практической проверке рассматриваемых гипотез происходит сопоставление опытных значений критерия $K_{опытн}$ с табличным значением критерия $K_{табл}$ и далее в зависимости от соотношения.

$$K_{îïûòí} \quad \triangleright\triangleleft \hat{E}_{êðèò}$$

Принимают или отвергают выдвинутую гипотезу.

Порядок проверки статистических гипотез можно сформулировать так: Если опытное значение критерия $K_{опытн}$, вычисляемое при заданном уровне значимости α попадает в область принятия гипотезы, то гипотезу принимают. Если же опытное значение критерия $K_{опытн}$ попадает в критическую область, то гипотезу отвергают.

В общем случае квантиль- это обращенное значение функции распределения. Так например, (рис. 5), доверительной вероятности $P_д = 90\%,$ отвечает 95%-ный квантиль. При этом уровень значимости составляет α =0,05.

<u>Квантиль</u> - это абсцисса, отсекающая от площади ограниченной, например, кривой плотности распределения случайной величины, заданный процент площади. Например, медиана – это 50%-ный квантиль. Квантиль, отсекающий 25% площади называется квартилем. При этом различают нижний 25%-ный квантиль и верхний 75%-ный квантиль. Имеются таблицы квантилей, составленные для различных законов распределения, например, для нормального закона, показательного закона и т.п. Критерии Пирсона, Кохрена, Стьюдента и Фишера также являются квантилями (см. приложения).

Уровень значимости α - это вероятность отклонения истинной гипотезы, представляет собой нечто иное как риск изготовителя, т.е вероятность отклонения годной партии изделий.

На первый взгляд кажется, что чем меньше уровень значимости α, тем строже условия проверки гипотезы. Например, при α = 0,05 разрешается совершить ошибку первого рода в пяти случаях из ста. При α =0,01 разрешается совершить ошибку первого рода в одном случае из ста. Однако с ошибкой первого рода (вероятностью отвергнуть верную гипотезу $H_{\circ}$), связана ошибка

второго рода β (вероятность того, что гипотеза будет принята, в то время когда она на самом деле неверна, например, принять негодную партию изделий).

Вероятность ошибки второго рода β зависит от характера проверяемой гипотезы, от способов проверки и от многих других причин, что сильно усложняет ее оценку.

Известно, что вероятность ошибки второго рода β тем меньше, чем выше уровень значимости α, так как при этом увеличивается число отвергаемых гипотез. Качественное соотношение между ошибками первого и второго родов показано на рис. 4.

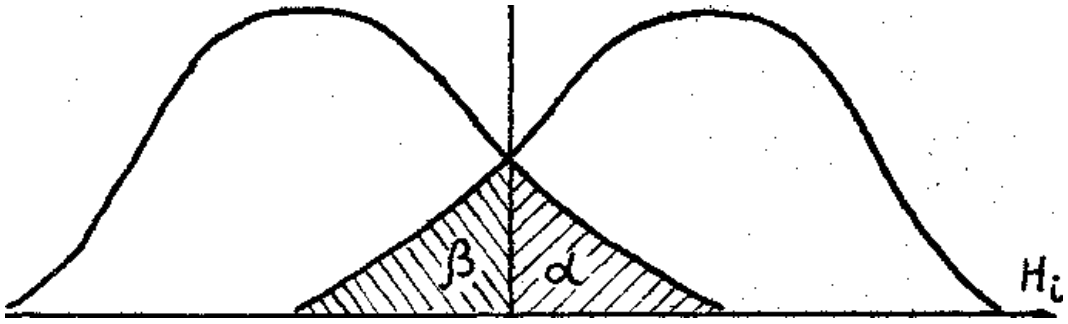

Рис 4. Качественное соотношение между ошибками первого и второго родов, H_i - различные гипотезы

Из рис. 4 видно, что «ужесточение» условий, т.е. повышение строгости скорей будет при повышении уровня значимости α, т.е. более строгими условия будут при α = 0,10, чем при α = 0,01.

Занятие 3

Проверка воспроизводимости эксперимента

После проведения опытов, как уже отмечалось выше, производится проверка воспроизводимости эксперимента. Поясним сущность сказанного.

При фиксации (во время экспериментов) наблюдаемых значений функции отклика в различных точках факторного пространства) если проводится однофакторный эксперимент, то в различных точках сечения графика (рис.1) могут возникать ошибки, например, ошибки замера, ошибки отсчета по шкалам приборов, ошибки от влияния неучтенных и неуправляемых факторов и т.п. Значения указанных ошибок (называемых уровнем «шума» установки) характеризуются отвечающими им в различных точках построчечными дисперсиями:

$$S^2\{Y_1\}, S^2\{Y_2\}, \ldots, S^2\{Y_n\}.$$

Если построчные дисперсии в различных точках факторного пространства мало отличаются друг от друга, то говорят, что дисперсии однородны и имеет место хорошая воспроизводимость эксперимента. В противном случае говорят, что дисперсии неоднородны.

Однородные дисперсии можно осреднить и получить дисперсию воспроизводимости всего эксперимента, которая используется при статистическом анализе значимости рассчитанных коэффициентов

$$B_\circ, B_1, \ldots, B_n$$

и при проверке полученной математической модели на адекватность. Если же построчечные дисперсии в различных точках неоднородны, тогда должна быть найдена и устранена причина возникающих отклонений.

В связи с этим, приступая к статистической обработке полученной информации, вначале производят проверку воспроизводимости эксперимента, т.е. проверку однородности построчечных дисперсий функции отклика в различных точках факторного пространства.

Таким образом, возникает одна из задач статистики – задача сравнения нескольких (построчечных) дисперсий.

При этом выдвигаются две гипотезы:

- все построчечные дисперсии однородны, и, следовательно, имеется хорошая воспроизводимость эксперимента;
- дисперсии неоднородны и воспроизводимость эксперимента плохая.

Указанная задача может решаться:

- с помощью критерия Кохрена, если число опытов в каждой точке сечения графика(в каждой точке факторного пространства) одинаково;
- с помощью критерия Бартлета, если число опытных точек в каждом из сечений графика различно.

Простым (более удобным) является критерий Кохрена.

Проверка правдоподобия гипотезы об однородности построчечных дисперсий с помощью критерия Кохрена записывается в виде следующего альтернативного условия, отвечающего правосторонней критической области (см. рис. 3 и табл. приложения 1)

$$G_{\substack{êîõð\\ îïûòí}} = \frac{S^2\{Ó_1\}_{\max}}{\sum_{i=1}^{n} S^2\{Ó_i\}} = \frac{\left[\sum_{i=1}^{n} \frac{(Ó_{ij} - \bar{Ó}_i)^2 m_{yi}}{n_{xj} - 1}\right]_{\max}}{\sum_{i=1}^{n}\sum_{j=1}^{nxj} \frac{(Ó_{ij} - \bar{Ó}_i)\cdot m_{yi}}{n_{xj} - 1}} =$$

$$= \begin{cases} < G_{\substack{êîõð\\ òàáë}}\begin{pmatrix}\alpha\\ k=r-1\\ n\end{pmatrix} & \text{- гипотеза об однородности дисперсий не отвергается;} \\ > G_{\substack{êîõð\\ òàáë}}\begin{pmatrix}\alpha\\ k=ê-1\\ n\end{pmatrix} & \text{- гипотеза отвергается,} \end{cases}$$

где $G_{\substack{êîõð\\ îïûòí}}$ - опытное значение критерия Кохрена;

$G_{\substack{êîõð\\ òàáë}}$ - табличное (критическое) значение критерия Кохрена;

$S^2\{Ó_i\}_{\max}$ - максимальное значение построчечной дисперсий;

$\sum_{i=1}^{n} S^2\{Y_i\}$ - сумма построчечных дисперсии;

α - уровень значимости (ошибка первого рода)

$k = r - 1$ - число степеней свободы. В рассматриваемом примере $K = n_{xj} - 1 = 5$. В статистике числом степеней свободы называется разность между числом опытов и числом констант, которые уже вычислены по результатам этих опытов;

$r(n_{xj})$ - число опытов в сечении графика (число параллельных опытов); в рассматриваемой задаче $r = 5$;

n - число сечений (число складываемых дисперсий).

Вычисляем для рассматриваемого примера построчечные дисперсии, по формуле:

$$S^2\{Y_i\} = \sum_{j=1}^{n_{xj}} \frac{(Y_{ij} - \bar{Y}_j)\cdot m_{yi}}{n_{xj} - 1},$$

где 1-ставится для того, чтобы дисперсия не была смещенной.

Для первого сечения получаем

$$S^2\{Y_1\}=\sum_{j=1}^{n_{xj}}\frac{\left(Y_{ij}-\overline{Y}_j\right)^2\cdot m_{yi}}{n_{x_1}-1}=\frac{(10-11)^2\cdot 2+(11-11)^2\cdot 1+(12-11)^2\cdot 2}{5-1}=1$$

Аналогично для последующих сечений (табл.1, строка 4), следовательно, сумма всех построчечных дисперсии равна

$$\sum_{i=1}^{n}S^2\{Y_i\}=\sum_{i=1}^{n}\sum_{j=1}^{n_{xj}}\frac{\left(Y_{ij}-\overline{Y}_j\right)\cdot m_{y_1}}{n_{x_j}-1}=1+1+1+0{,}8=3{,}8.$$

Следовательно, для рассматриваемого примера, опытное значение критерия Кохрена составляет:

$$G\underset{опытн}{кохр}=\frac{1}{3{,}8}=0{,}263.$$

Табличное (критическое) значение критерия Кохрена при уровне значимости α =0,05, числе степеней свободы числителя $K=r-1=5-1=4$, и числе степеней свободы знаменателя (числе сечений графика) n=**4** (см. приложения 1), составляет:

$$G\,кохр\underset{табл}{\begin{pmatrix}\alpha=0{,}05\\ K=r-1=5-1=4\\ n=4\end{pmatrix}}=0{,}62,$$

Следовательно

$$G\underset{опытн}{кохр}=0{,}263<G\,кохр\underset{табл}{}=0{,}62.$$

Это значит, что при α =0,05 опытное значение критерия Кохрена попадает в область принятия гипотезы и, следовательно, гипотеза об однородности дисперсий не отвергается (см. рис. 5)

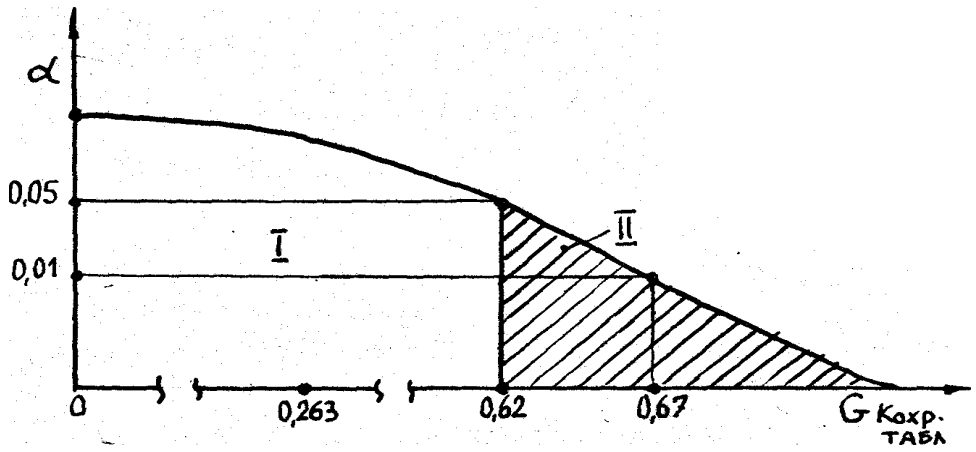

Рис 5. Область принятия гипотезы при уровне значимости α =0,05: I- область принятия гипотезы; II область непринятия гипотезы (критическая область)

Если изменить условия проверки гипотезы об однородности дисперсий и принять α =0,01, тогда

$$G\,кохр_{табл}\begin{pmatrix}\alpha=0{,}01\\ K=4\\ n=4\end{pmatrix}=0{,}67.$$

Как видим и для других условий опытное значение критерия Кохрена также попадает в область принятия гипотезы. Следовательно, гипотеза об однородности дисперсий не отвергается.

Это значит, что теперь можно приступить к определению коэффициентов математической модели.

Заметим также, что поскольку построчечные дисперсии однородны, то это позволяет вычислить дисперсию воспроизводимости всего эксперимента, величина которой определяется путем осреднения построчечных дисперсий по формуле

$$S^2\{Y\}_{воспр}=\sum_{i=1}^{n}\sum\frac{\left(Y_{ij}-\overline{Y}_j\right)^2}{n(r-1)}=\sum_{i=1}^{n}\frac{S^2\{Y_i\}}{n};\qquad(7)$$

где n – число сечений (число складываемых дисперсий).

Для рассматриваемого примера получаем:

$$S^2\{Y\}_{воспр}=\sum_{i=1}^{n}\frac{S^2\{Y_i\}}{n}=\frac{1+1+1+0{,}8}{4}=\frac{3{,}8}{4}=0{,}95$$

Вычисление коэффициентов однофакторной линейной модели

Рассматривая график опытной линии регрессии (ломаная линия, рис 1) Принимаем решение выравнивать указанную опытную зависимость прямой линией, коэффициенты, которой определяются с помощью вышеприведенных формул (5) и (6).

Расчет значений коэффициентов $B_{\circ}uB_1$ производим в следующем порядке:

1. Находим общие средние арифметические значения факториального и результативного признаков:

$$M^*[x]=\overline{x}=\sum_{i=1}^{5}x_i\frac{m_{xi}}{n}=1\cdot\frac{5}{21}+2\frac{5}{21}+3\frac{5}{21}+4\frac{6}{21}=2{,}57;$$

$$M^*[y]=\overline{y}=\sum_{i=1}^{10}\acute{O}_i\frac{m_{yi}}{n}=10\cdot\frac{2}{21}+11\cdot\frac{1}{21}+\ldots+5\cdot\frac{2}{21}=7{,}33.$$

2. Находим общие несмещенные дисперсии и средние квадратические отклонения факториального и результативного признаков:

$$\widetilde{D}[x]=\frac{n}{n-1}\left[\sum_{i=1}^{5}\frac{x_i^2 m_{xi}}{n}-(\bar{x})^2\right]=\frac{21}{20}\left[1^2\cdot\frac{5}{21}+2^2\cdot\frac{5}{21}+3^2\frac{5}{21}+4^2\cdot\frac{6}{21}-(2{,}57)^2\right]=1{,}364;$$

$$\widetilde{D}[x]=\frac{n}{n-1}\left[\sum_{i=1}^{10}\frac{Y_i^2 m_{xi}}{n}-(\bar{y})^2\right]=\frac{21}{20}\left[10^2\cdot\frac{5}{21}+11^2\cdot\frac{5}{21}+\ldots+3^2\frac{2}{21}-(7{,}33)^2\right]=7{,}33;$$

$$\widetilde{\delta}(x)=S\{x\}=\sqrt{1{,}364}=1{,}168;$$

$$\widetilde{\delta}(y)=S\{y\}=\sqrt{7{,}633}=2{,}762.$$

3. Находим несмещенный момент связи и коэффициент корреляции*

$$\widetilde{K}_{yx}=\frac{n}{n-1}\left[\sum_{i=1}^{n}x_i y_i\frac{m_{yi}}{n}-\overline{xy}\right]=\frac{21}{20}\left[1\cdot 10\cdot\frac{2}{21}+1\cdot 11\cdot\frac{1}{21}+\ldots+-(2{,}57)(7{,}33)\right]=-2{,}73$$

$$\widetilde{r}_{yx}=\widetilde{K}_{yx}/\delta(y)\cdot\delta(x)=-2{,}73/2{,}762\cdot 1{,}168=-0{,}846.$$

Следовательно, матрица несмещенных моментов связи и матрица коэффициентов корреляции запишутся так

$$\widetilde{K}_{yx}=\begin{Vmatrix}D[Y] & \widetilde{K}_{yx}\\ \widetilde{K}_{xy} & D[x]\end{Vmatrix}=\begin{Vmatrix}7{,}633 & -2{,}73\\ -2{,}73 & 1{,}364\end{Vmatrix};$$

$$\widetilde{R}_{yx}=\begin{Vmatrix}1 & r_{yx}\\ r_{xy} & 1\end{Vmatrix}=\begin{Vmatrix}1 & -0{,}846\\ -0{,}846 & 1\end{Vmatrix}$$

5. Вычисляем значения коэффициентов линейной парной регрессионной зависимости

$$B_1=\frac{\widetilde{K}_{yx}}{D[x]}=\frac{-2{,}73}{1{,}364}=-2;$$

$$B_\circ=\bar{y}-B_1\bar{x}=7{,}33-(-2)\cdot 2{,}57=12{,}47.$$

и, следовательно, теоретическое (расчетное) уравнение линейной регрессионной зависимости запишется в виде:

$$\hat{Y}_{расч}=12{,}47-2x.$$

*** Парный коэффициент корреляции r_{yx} впервые в статистике был предложен Френсисом Гальтоном и затем развит в работах К. Пирсона. Коэффициент корреляции r_{yx} представляет собой скалярное произведение двух векторов. Кроме коэффициента корреляции в статистике применяются также другие коэффициенты, с помощью которых определяется зависимость взаимосвязанных величин. Так, например, имеется коэффициент ранговой корреляции Спирмена, коэффициент конкордации Кендела и** др.

Указанное расчетное уравнение представляет собой математическую модель, с помощью которой (после статистической оценки значимости ее коэффициентов и проверки на адекватность) может описываться изучаемый процесс, т.е. процесс снижения травм на автопредприятии.

Наносим полученную расчетную линию регрессии на график (см. рис 2.)

если x =1, тогда y = 12,47 – 2 =10,47;
если x =2, тогда y =12,47 – 4 =8,47;
если x =3, тогда y =12,47 -6 =6,47;
если x =4, тогда y =12,47 -8 =4,47;
если x =5, тогда y =12,47 -10 =2,47;
если x =6, тогда y =12,47 -12 = 0,47.

6.Вычисляем сумму квадратов отклонений расчетных значений результативного признака от опытных (см. табл. 2).
Для полученной прямой линии указанная сумма квадратов составляет

$$\sum_{i=1}^{n}\left(Y_{iопытн} - Y_{iрасч}\right)^2 = 5.$$

Занятие 4

Однофакторная показательная регрессия

Вывод формул для определения коэффициентов модели

При решении многих инженерных и экономических задач выравнивание опытной линии регрессии, как уже отмечалось выше , может производиться с помощью показательной функции

$$\hat{Y}_{расч} = B_{\circ} \cdot \left(B_1\right)^x. \qquad (8)$$

Логарифмируя уравнение (8) приводим его к виду прямой линии

$$\lg\left(\hat{Y}_{расч}\right)=\lg B_{\circ}+\lg B_1\cdot x$$

или

$$y'=B_{\circ}+B_1'\cdot x.$$

Выражения для вычисления коэффициентов $B_{\circ}$ и B_1 запишутся так

$$\begin{cases} B_1'=\dfrac{\widetilde{K}_{xy'}}{D[x]}=\dfrac{\alpha_{11(xy')}-\bar{x}\cdot\bar{y}'}{D[x]}=\dfrac{\sum\limits_{i=j=1}\dfrac{x_i\lg(\bar{y}_i)}{n}-\sum\limits_{i=1}^{n}\dfrac{x_i}{n}\cdot\sum\limits_{i=1}^{n}\dfrac{\lg(\bar{y}_i)}{n}}{D[x]} \\ B_{\circ}'=\left(\lg\bar{y}_i\right)-B_1'\bar{x}=\sum\limits_{i=1}^{n}\dfrac{\lg(\bar{y}_i)}{n}-\dfrac{\widetilde{K}_{x\bar{y}'}}{D[x]}\cdot\bar{x}. \end{cases} \tag{9}$$

Покажем порядок вычисления коэффициентов $B_{\circ}$ и B_1 отвечающих однофакторной показательной регрессии применительно к рассматриваемому выше примеру (табл. 2).

Подставляя полученные значения сумм в формулы (9) для определения $B_{\circ}'$ и B_1' получаем

$$B_1'=\frac{\sum\limits_{i=1}^{4}\dfrac{x_i\lg(\bar{y}_i)}{n}-\sum\limits_{i=1}^{4}\dfrac{x_i}{n}\cdot\sum\limits_{i=1}^{4}\dfrac{\lg(\bar{y}_i)}{n}}{D[x]}=\frac{1{,}962-(2{,}57)(0{,}848)}{1{,}364}=-0{,}159$$

Потенцируя, получаем $B_1=0{,}693$;

$$B_{\circ}'=\sum_{i=1}^{n}\frac{\lg(\bar{y}_i)}{n}-B_1'\cdot\bar{x}=0{,}848+(0{,}159)(2{,}57)=1{,}256,$$

Потенцируя, получаем $B_{\circ}=18.$

Таким образом, уравнение регрессии для рассматриваемого примера, отвечающее показательной зависимости, имеет вид

$$\hat{Y}_{расч}=B_{\circ}\left(B_1\right)^{x_1}=18\cdot(0{,}693)^{x_1}. \tag{10}$$

Наносим кривую, отвечающую показательной зависимости на график (рис.2)

если x=1, то y=12,474;
если x=2, то y=8,64;
если x=3, то y=5,98;
если x=4, то y=4,15;
если x=5, то y=2,87;
если x=6, то y=2,00.

Вычисляем сумму квадратов отклонений расчетных значений результативного признака от их опытных значений для показательной функции(см. табл. 2).

$$\sum_{i=1}^{n}\left(\overline{y}_{iопытн} - y_{iрасч}\right)^2 = 8{,}9^{*}.$$

*** Аппроксимация опытных линий регрессии для однофакторной зависимости может также производиться и другими аналитическими зависимости, например:**
- степенной функцией;
- логарифмической функцией;
- гиперболической функцией и др.

Расчет коэффициентов B_0 и B_1 для однофакторной показательной регрессионной зависимости

Таблица 2

x_i	$\overline{y_i}$	$\lg(\overline{y_i})$	$x\lg(\overline{y_i})$	Квадраты отклонений опытной линии регрессии	
				от прямой линии	от показательной функции
1	11	1,04139	1,04139	(11- 10,47)²= 0,28	(11 - 12,47)² = 2,16
2	7	0,84510	1,6902	(7 - 8,47)² = 2,16	(7 - 8,64)² = 2,68
3	8	0,90309	2,7092	(8 - 6,47)² = 2,34	(8 – 5,98)² = 4,04
4	4	0,60206	2,40824	(4 - 4,47)² = 0,22	(4 - 4,15)² = 0,022
$\overline{x}=2{,}57$	$\overline{y}=7{,}33$	$\sum_{i=1}^{n}\frac{\lg(\overline{y}_i)}{n}=0{,}848$	$\sum_{i=1}^{n}\frac{x_i\lg(\overline{y}_i)}{n}=$ =1,962	$\sum_{i=1}^{n}\frac{(\overline{y}_i опытн - y_i расч)^2}{n}$ =5	$\sum_{i=1}^{n}\frac{(\overline{y}_i опытн - y_i расч)^2}{n}$ = 8,9

Примечание: $\overline{y}_i опытн$ и $y_i расч$ - соответственно опытное и расчетное значения функции отклика**.**

Дискриминация математических моделей

Для выбора наивыгоднейшей из числа сравниваемых нескольких математических моделей и, следовательно, для отсева или дискриминации худших, вначале вычисляются дисперсии аппроксимации(называемые также ошибками неадекватности) для каждой из сравниваемых моделей по формуле:

$$S^2\{y\}_{\substack{\text{àïïðîêñ}\\(\text{íåàäåêâ})}} = \sum_{i=1}^{n} \frac{\left(\bar{y}_{\text{íîìîòí}} - y_{\text{iðàñ÷}}\right)^2}{n-d-1}, .(11)$$

где $\bar{y}_{iопытн}$ и $y_{iрасч}$ - соответственно опытное и расчетное значения функции отклика;

n – полный объем выборки (число испытаний);

d – число значащих коэффициентов расчетного уравнения;

1- ставится для того, чтобы оценка не была смещенной.

Для рассматриваемого примера (табл. 2) дисперсии неадекватности для прямой линии

$$S^2\{y\}н.а = \sum_{i=1}^{n} \frac{\left(\bar{y}_{iопытн} - y_{iрасч}\right)^2}{n-d-1} = \frac{5}{21-2-1} = 0{,}277.$$

для показательной функции

$$S^2\{y\}н.а = \sum_{i=1}^{n} \frac{\left(\bar{y}_{iопытн} - y_{iрасч}\right)^2}{n-d-1} = \frac{8{,}9}{21-2-1} = 0{,}494.$$

Дискриминация математических моделей производится попарным сравнением дисперсий двух сравниваемых математических моделей.

При этом выдвигаются две гипотезы:

- сравниваемые дисперсии однородны, т.е разница между ними незначима (нулевая гипотеза):
- сравниваемые дисперсии неоднородны, т.е разница между ними значима.

Дискриминация сравниваемых моделей производится с помощью критерия Фишера в виде следующего альтернативного условия, отвечающего правосторонней критической области,

$$F_{опытн} = \frac{S^2\{y\},больш}{S^2\{y\}меньш} = \frac{\sum_{i=1}^{n} \frac{\left(\bar{y}_{ijопытн} - y_{iрасч}\right)^2}{n-d-1}}{\sum_{i=1}^{n} \frac{\left(\bar{y}_{iопытн} - y_{iрасч}\right)^2}{n-d-1}} =$$

$$= \begin{cases} > F_{табл}\begin{pmatrix} \alpha = 0{,}05 \\ K_1 = n_1 - 1 \\ K_2 = n_1 - 1 \end{pmatrix} \text{- разница между дисперсиями значима;} \\ \leq F_{табл}\begin{pmatrix} \alpha = 0{,}05 \\ K_1 = n_2 - 1 \\ K_2 = n_2 - 1 \end{pmatrix} \text{- разница между дисперсиями незначима,} \end{cases}$$

где α - уровень значимости (вероятность ошибки первого рода);

K_1 - число степеней свободы большей дисперсии;

K_2 - число степеней свободы меньшей дисперсии;

n_1 - полный объем выборки, по которой вычислена большая несмещенная дисперсия;

n_2 - полный объем выборки, по которой вычислена меньшая дисперсия.

Для сравниваемых в рассматриваемом примере двух математических моделей дисперсионное отношение равно:

$$F_{опытн} = \frac{S^2\{y\}больш}{S^2\{y\}меньш} = \frac{0{,}494}{0{,}277} = 1{,}78,$$

где $S^2\{y\}больш$ и $S^2\{y\}меньш$ - дисперсии неадекватности соответственно экспоненты и прямой линии.

Расчетное значение критерия Фишера при α=0,05 равно:

$$F_{расч}\begin{pmatrix} \alpha = 0{,}05 \\ K_1 = 21 - 1 = 20 \\ K = 21 - 1 = 20 \end{pmatrix} = 2{,}2.$$

Как видим, при уровне значимости α=0,05

$$F_{опытн} = 1{,}78 < F_{табл} = 2{,}2.$$

Это значит, что при α= 0,05 опытное значение критерия Фишера не попадает в критическую область.

Следовательно, обе математические модели можно считать равноценными.

Учитывая, что прямая линия является более простой формой, поэтому принимаем решение выравнивать опытную линию регрессии прямой линией.

Статистическая оценка значимости коэффициентов линейной модели

После дискриминации, т.е. после отбора лучшей математической модели производиться статистическая оценка значимости коэффициентов указанной модели. В рассматриваемой задаче – это коэффициенты прямой линии $B_{\circ}$=12,47 и $B_1 = -2$.

Оценка коэффициентов производится по очереди. При этом выдвигаются две гипотезы:

- коэффициент значим;
- коэффициент незначим.

Оценка значимости коэффициентов производиться с помощью критерия Стьюдента, записанного в виде следующего альтернативного условия, отвечающего левосторонней критической области.

$$t_{опытн} = \frac{\left|B_j\right|}{S\{B_j\}} = \frac{\left|B_j\right|}{\sqrt{c_{ii}}S\{y\}_{воспр}} =$$

$$= \begin{cases} \geq t_{табл}\begin{pmatrix}\alpha \\ K = n-1\end{pmatrix} \\ < t_{табл}\begin{pmatrix}\alpha \\ K = n-1\end{pmatrix} \end{cases}$$

где $t_{опытн}$ - опытное значение критерия Стьюдента;

$t_{табл}$ - табличное (критическое) значение оцениваемого коэффициента;

$\left|B_j\right|$ - абсолютное значение оцениваемого коэффициента;

$S^2\{B_j\}$ - дисперсия коэффициента модели;

$S^2\{y\}$ - дисперсия воспроизводимости всего эксперимента;

c_{ii} - диагональный элемент обратной матрицы;

α - уровень значимости;

n - полный объем выборки;

$K = n-1$ - число степеней свободы.

Покажем порядок определения элементов c_{ii}.

Для рассматриваемого примера исходная и транспонированная матрицы записывается так

$$X_{исх} = \begin{Vmatrix} x_{01} & x_{11} \\ x_{02} & x_{12} \\ x_{03} & x_{13} \\ x_{04} & x_{14} \end{Vmatrix} = \begin{Vmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{Vmatrix}; \; X^T = \begin{Vmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{Vmatrix}$$

Матрица моментов равна

$$X^T X=\begin{Vmatrix}1 & 1 & 1 & 1\\ 1 & 2 & 3 & 4\end{Vmatrix}\begin{Vmatrix}1 & 1\\ 1 & 2\\ 1 & 3\\ 1 & 4\end{Vmatrix}=\begin{Vmatrix}4 & 10\\ 10 & 30\end{Vmatrix};\det\left(X^T X\right)=20.$$

Присоединенная матрица от транспонированной равна

$$(X^T X)=\begin{Vmatrix}30 & -10\\ -10 & 4\end{Vmatrix}.$$

И, следовательно, обратная матрица запишется так

$$(X^T X)^{-1}=(X^T X)_{\det(X^T X)}=\frac{1}{20}\begin{Vmatrix}30 & -10\\ -10 & 4\end{Vmatrix}=\begin{Vmatrix}3/2 & -1/2\\ -1/2 & 1/5\end{Vmatrix}=\begin{Vmatrix}C_{00} & C_{01}\\ C_{10} & C_{11}\end{Vmatrix}$$

Таким образом, $C_{00}=3/2$ и $C_{11}=1/5$

Оцениваем значимость коэффициента $B_0=12{,}47$

$$t_{опытн}=\frac{|B_\circ|}{\sqrt{C_{11}}S\{y\}_{воспр}}=\frac{|12{,}47|}{\sqrt{3/2}\sqrt{0{,}95}}=10\ ;$$

$$t_{табл}\begin{pmatrix}\alpha=0{,}05\\ K=21-1=20\end{pmatrix}=2{,}09.$$

Как видим,

$$t_{опытн}=10>t_{табл}=2{,}09.$$

Значит, коэффициент $B_0=12{,}47$ значим.

Оцениваем значимость коэффициента $B_1=-2$

$$t_{опытн\ (для B_1)} = \frac{|B_1|}{\sqrt{C_{22}}S\{y\}_{воспр}} = \frac{|-2|}{\sqrt{1/5}\sqrt{0{,}95}} = 4{,}59$$

$$t_{табл}\begin{pmatrix}\alpha = 0{,}05\\ K = 20\end{pmatrix} = 2{,}09.$$

Следовательно

$$t_{опытн} = 4{,}59 > t_{табл} = 2{,}09.$$

Как видим, оба коэффициента полученной математической модели (прямая линия) статистически значимы.

Это позволяет теперь приступить к проверке полученной математической модели на адекватность.

Проверка математической модели на адекватность

После дискриминации математических моделей и статистической оценки значимости коэффициентов проверяется отобранная математическая модель на адекватность.

При этом выдвигаются две гипотезы:

- модель адекватная;
- модель неадекватная.

Проверка правдоподобности гипотезы об адекватности производиться с помощью критерия Фишера, записываемого в виде следующего альтернативного условия, отвечающего правосторонней критической области

$$F_{опытн\ .} = \frac{S^2\{y\}_{н.а}}{S^2\{y\}_{воспр\ .}} = \begin{cases} \le F_{табл\ .}\begin{pmatrix}\alpha = 0{,}05\\ K_1 = n - d - 1\\ K_2 = n - 1\end{pmatrix} \\ \rangle F_{табл\ .}\begin{pmatrix}\alpha = 0{,}05\\ K_1 = n - d - 1\\ K_2 = n - 1\end{pmatrix} \end{cases},$$

где α - уровень значимости;

n - число всех испытаний для отобранной модели;

d - число значащих коэффициентов для отобранной модели;

r - число параллельных опытов.

Для рассматриваемого примера, (для прямой линии) дисперсионное отношение, т.е. опытное значение критерия Фишера составляет:

$$F_{опытн}=\frac{S^2\{y\}_{н.а}}{S^2\{y\}_{воспр}}=\frac{r\cdot\sum\limits_{i=1}^{n}\frac{\left(\bar{y}_{iопытн}-y_{iрасч}\right)^2}{n-d-1}}{\frac{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{r}\left(y_{ij}-\bar{y}_j\right)^2}{n(r-1)}}=\frac{5\cdot0{,}277}{0{,}95}=1{,}45.$$

Табличное значение критерия Фишера, как уже отмечалось выше, составляет:

$$F_{табл}=\begin{pmatrix}\alpha=0{,}05\\ K_1=n-d-1=21-2-1=18\\ K_2=n-1=21-1=20\end{pmatrix}=2{,}2.$$

Следовательно, при уровне значимости α=0,05, опытное значение критерия Фишера для прямой линии меньше табличного, т.е.

$$F_{опытн}=1{,}45<F_{табл}=2{,}2.$$

Это значит, что избранная математическая модель (прямая линия) адекватно описывает изучаемое явление.

Занятие 5

Построение доверительного интервала среднего результата функции линейной однофакторной регрессии

Полученное выше уравнение математической модели в виде прямой линии y=12,47-2x основывается на небольшом числе наблюдений и значит в реальной действительности имеет место разброс указанной расчетной линии относительно ее неизвестного истинного значения. Указанный разброс в каждом сечении графика характеризуется доверительным интервалом, отвечающим заданной доверительной вероятности Рд.

Половина доверительного интервала, как известно из основных положений математической статики, равна среднему квадратическому отклонению среднего результата, умноженному на обращенное значение функций Стьюдента, т.е.

$$\boldsymbol{\delta[\![M^*(x)]\!]=\mathfrak{S}[\![M^*(x)]\!]\cdot S^{-1}(n,P_Д)=\frac{\mathfrak{S}\{y\}_{воспр.}}{\sqrt{n}}\cdot S^{-1}(n,P_Д),}\qquad(12)$$

где

$\delta[\![\mathbf{M}^*(x)]\!]$ - половина доверительного интервала разброса среднего результата;

$\mathfrak{S}[\![\mathbf{M}^*(x)]\!] = \frac{\mathfrak{S}\{y\}\text{воспр.}}{\sqrt{n}}$ - среднее квадратическое отклонение среднего результата;

$\mathbf{S}^{-1}(n, \mathbf{P}_Д)$ - обращенное значение функций Стьюдента

n – полный объем выборки:

$\mathbf{P}_Д$ – доверительная вероятность.

Покажем, для рассматриваемого примера порядок вычисления половины доверительного интервала разброса среднего результата при $\mathbf{P}_Д$ =0,98;

$$\mathfrak{S}^2\{Y\}\text{воспр.} = 0,95;\ \mathbf{S}^{-1}(n = 21;\ \mathbf{P}_Д = 0,90) = 1,81.$$

и значит, для заданных условий для прошедшего периода времени половина доверительного интервала разброса среднего результата составляет:

$$\delta[\![\mathbf{M}^*(x)]\!] = \frac{\mathfrak{S}\{Y\}\text{воспр.}}{\sqrt{n}} \cdot \mathbf{S}^{-1}(n = 21;\ \mathbf{P}_Д = 0,90) = \frac{0,95}{\sqrt{21}} \cdot 1,81 = 0,362$$

Поясним сказанное.

Понятие доверительного интервала разброса среднего результата.

Напомним, что вероятность попадания случайной величины в интервал **(a‹x‹b)** для случайной величины, распределенной нормально (см. рис. 6), равна

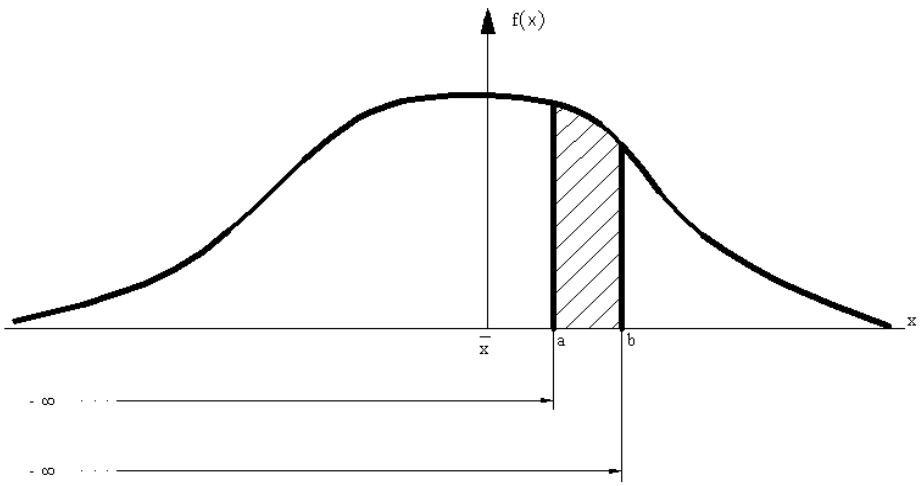

Рис.6. Вероятность попадания случайной величины в интервал ***(a‹x‹b)***

$$\mathbf{P}(a < x < b) = \int_{-\infty}^{b} \frac{1}{\mathfrak{S}_x\sqrt{2\pi}} e^{-(x-\bar{x})^2/2\mathfrak{S}_x^2} dx - \int_{-\infty}^{a} \frac{1}{\mathfrak{S}_x\sqrt{2\pi}} e^{-(x-\bar{x})^2/2\mathfrak{S}_x^2} dx \ ,$$

где **x**- частные значения случайной величины.

Переходя к центрированному и нормальному значению случайной величины

$$t = \frac{x - \bar{x}}{\sigma_x},$$

получаем:

$$\mathbf{P(a < x < b) = \Phi^*\left(\frac{b - \bar{x}}{\sigma_x}\right) - \Phi^*\left(\frac{a - \bar{x}}{\sigma_x}\right)}, \qquad (13)$$

где $\Phi^*(t)$ – табличная функция распределения для нормального закона (табл. 2 приложения 1).

Наряду с табличной функцией распределения в теории вероятностей применяется также функция Лапласа

$$\mathbf{F_{лапаласа}} \int_0^x \frac{1}{\sigma_x\sqrt{2\pi}} e^{-(x-\bar{x})^2/2\sigma_x^2} dx$$

которая после нормирования и центрирования записывается так

$$\Phi_0(t) = \Phi_0\left(\frac{x - \bar{x}}{\sigma_x}\right) \qquad \text{(табл.3, приложение 1)}$$

Вероятность попадания в заданный интервал **(a‹x‹b)** с помощью табличной функции Лапласа записывается так

$$\mathbf{P(a < x < b) = \frac{1}{2}\left[\Phi_0\left(\frac{b - \bar{x}}{\sigma_x}\right) - \Phi_0\left(\frac{a - \bar{x}}{\sigma_x}\right)\right]}. \qquad (14)$$

Перед квадратной скобкой ставится 1/2 , потому, что подынтегральная функция четная.

Если требуется вычислить вероятность попадания в симметричный интервал

$[-\delta, +\delta]$, тогда $\mathbf{b = \bar{x} + \delta; a = \bar{x} - \delta}$ (рис. 7)

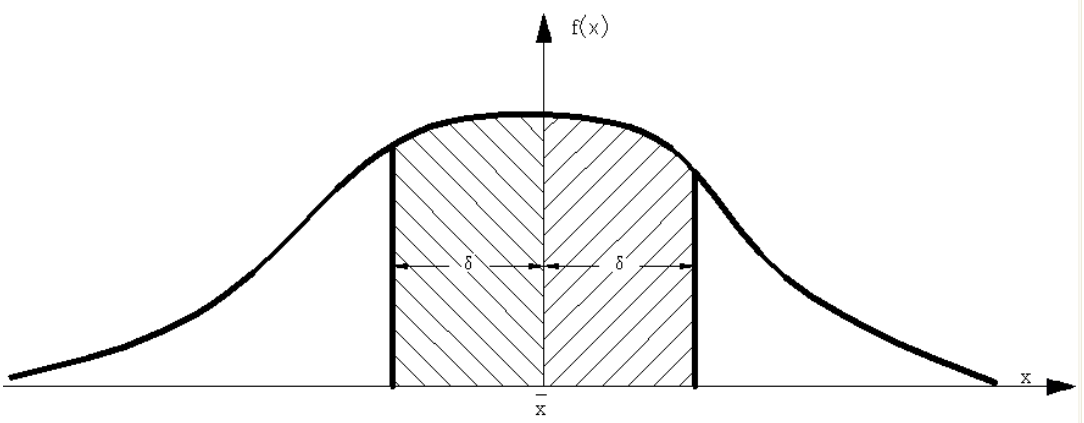

Рис. 7. Вычисление вероятности попадания в симметричный интервал

и значит:

$$\mathbf{P(-\delta < x < +\delta) = \frac{1}{2}\left[\Phi_0\left(\frac{\bar{x} + \delta - \bar{x}}{\sigma_x}\right) - \Phi_0\left(\frac{\bar{x} - \delta - \bar{x}}{\sigma_x}\right)\right] = \Phi_0\left(\frac{\delta}{\sigma_x}\right)} \qquad (15)$$

Разрешая равенство (15) относительно δ, получаем

$$\boldsymbol{\delta = \mathfrak{S}_x \cdot \phi^{-1}\left[P_Д(-\delta < x < +\delta)\right]}, \qquad (16)$$

где $\mathbf{P_Д(-\delta < x < +\delta)}$ - называется доверительной вероятностью (надежностью) и обозначается $P_Д$;

$\boldsymbol{\phi^{-1}(P_Д)}$ - обращенное значение функции Лапласа.

Таким образом, половина доверительного интервала разброса самой случайной величины x равна среднему квадратическому отклонению, умноженному на обращенное значение функции Лапласа.

Поскольку в рассматриваемом вопросе речь идет о разбросе не самой случайной величины x, а о разбросе её среднего результата, поэтому выведем формулу для определения среднего квадратического отклонения среднего результата. Как известно

$$\mathbf{M^*[x] = \sum_{i=1}^{n} x_i / n}.$$

Осуществляя операцию нахождения дисперсии, получаем

$$\mathbf{D\{M^*[x]\} = \frac{1}{n^2} \cdot D \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \frac{D[x]}{n^2} = \frac{n}{n^2} \cdot D[x] = D[x]/n}$$

и значит

$$\mathfrak{S}[\![M^*(x)]\!] = \mathfrak{S}_x/\sqrt{n}. \qquad (17)$$

Таким образом, половина доверительного интервала разброса среднего результата (называемая точностью), определяется по формуле

$$\delta[\![M^*(x)]\!] = \frac{\mathfrak{S}_x}{\sqrt{n}} \cdot \phi^{-1}(P_Д), \qquad (18)$$

где $\boldsymbol{\phi^{-1}(P_Д)}$ - обращенное значение табличной функции Лапласса.

n- объем выборки.

Если число наблюдений не велико $(n < 30)$, то в этом случае более точные результаты будут получены, если вместо обращенной функции Лапласса поставить обращенное значение функции Стьюдента (приложения 1)

$$\boldsymbol{\delta[\![M^*(x)]\!] = \frac{\mathfrak{S}_x}{\sqrt{n}} \cdot S^{-1}(n, P_Д)}. \qquad (19)$$

Что и требовалось доказать.

Для рассматриваемого примера случайной величиной служит **y**. при этом дисперсия воспроизводимости составляет $S^2\{y\}воспр. = 0,95$, а число всех опытов равно 21, поэтому среднее квадратическое отклонение среднего результата составляет:

$$\mathfrak{S}[\![M^*(y)]\!] = \frac{S\{y\}воспр}{\sqrt{n}} = \frac{0,95}{\sqrt{21}} = 0,2.$$

Обращенное значение функции Стьюдента равно

$$S^{-1}(n = 21;\ P_Д = 0,90) = 1,81.$$

и значит, половина доверительного интервала разброса среднего результата составляет

$$\delta[\![M^*(y)]\!] = \mathfrak{S}[\![M^*(y)]\!] \cdot S^{-1}(n = 21;\ P_Д = 0,90) = 0,2 \cdot 1,81 = 0,362,$$

На основании этого строим (для прошедшего периода времени) доверительный коридор разброса среднего результата, отвечающий доверительной вероятности $P_Д = 0,90$ (рис. 8).

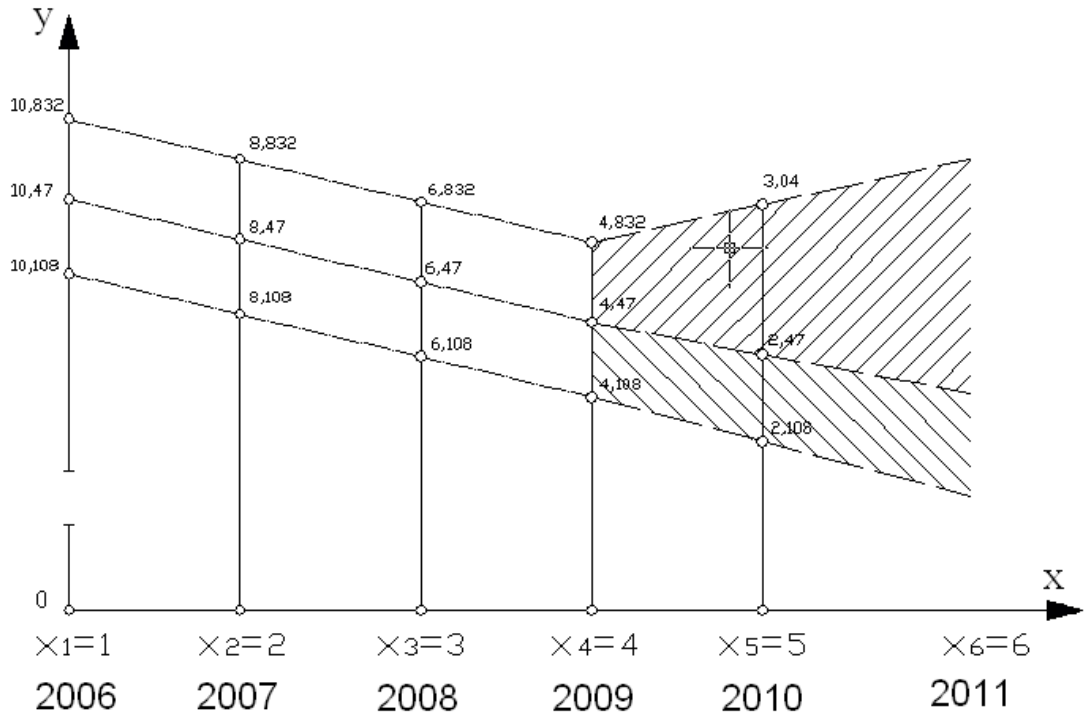

Рис. 8. доверительный коридор разброса среднего результата за прошедшее время, отвечающий доверительной вероятности $P_Д = 0,90$. *На рисунке также нанесен раструб разброса для прогнозируемого периода времени на 2006 – 2011 гг.*

Границы доверительного коридора вычисляются по формуле

$$y_{граничн}\ y_{ср} \pm \delta[\![M^*(x)]\!], \qquad (20)$$

т. е. верхняя граница коридора равна сумме расчетного значения среднего результата и половины доверительного интервала. Нижняя граница равна разности среднего результата и половины доверительного интервала.
В математической системе y_{CP} называют регулярной составляющей (тренд), под которым понимается описание процесса, очищенного от случайных помех.

Построение раструба прогноза линейного тренда

Очевидно, что для периода предсказания на положение верхний и нижний границ доверительного коридора разброса среднего результата оказывают влияние многие не контролируемые переменные. В связи с этим, коридор для периода прогноза значительно увеличивается и представляет собой раструб (рис. 8).

Границы раструба прогноза зависят:

- от продолжительности периода предшествующего наблюдения. Обозначим эту величину через **n** (равную числу сечений графика). Чем больше период предшествующего наблюдения, тем меньше раструб;

- от величины периода упреждения (или предсказания). Обозначим эту величину **L**. Чем больше **L,** тем больше раструб.

Половина доверительного интервала разброса среднего результата для периода прогноза вычисляется по формуле:

$$\delta[\![\mathrm{M}^*(y)]\!]_{\text{прогн}} = \delta[\![\mathrm{M}^*(y)]\!] \cdot \sqrt{\frac{n+1}{n} + \frac{3(n+2\cdot L-1)^2}{n(n^2-1)}}. \qquad (21)$$

Для рассматриваемого примера по формуле (21) для 1980 года имеем

$$\delta[\![\mathrm{M}^*(y)]\!]_{\text{прогн}} = 0.362\sqrt{\frac{4+1}{4} + \frac{3(4+2\cdot 1-1)^2}{4(4^2-1)}} = 0,572.$$

Аналогично получаем для 2011 г. (рис,8)

$$\delta[\![\mathrm{M}^*(y)]\!]_{\text{прогн}} = 0.362\sqrt{\frac{4+1}{4} + \frac{3(4+2\cdot 2-1)^2}{4(4^2-1)}} = 0,695.$$

Как видим, по мере увеличения периода предсказания величина раструба разброса среднего результата увеличивается.
Аналогично решаются другие подобные примеры

ПРИМЕНЕНИЕ ФАКТОРНОГО АНАЛИЗА ДЛЯ РЕШЕНИЯ ЗАДАЧ АВТОМОБИЛЬНОГО ТРАНСПОРТА

Занятие 6

Системный подход к исследованию и прогнозированию развития явлений и процессов, имеющих место при решении задач автомобильного транспорта, требует учета по возможности всей совокупности факторов, оказывающих влияние на параметр оптимизации и учета взаимосвязей между ними.

В связи с этим, в процессе исследования прогнозист вынужден выбирать компромиссионый вариант между числом переменных и сложностью и трудоемкостью анализа и прогноза.

Для снижения размерности описаний сложных объектов применяются различные способы, и в том числе факторный анализ.

Факторный анализ в настоящее время широко применяется при анализе экспериментальных данных как в естественных (технических и экономических), так и в гуманитарных науках. Факторный анализ проник в социологию, экономику, геологию, метеорологию, технику и, в том числе, в исследование функционирования автопроизводственных предприятий. К настоящему времени разработано не мало различных методов факторного анализа и их модификаций.

Столь широкий интерес к приложению методов факторного анализа обусловливается тем обстоятельством, что эти методы позволяют с некоторым приближением решать одну из наиболее распространенных задач научного исследования, а именно, задачу построения той или иной схемы классификации, т.е компактного содержательного описания исследуемого явления на основе обработки больших информационных массивов.

Рассмотренный выше регрессионно-корреляционный анализ позволяет получать закономерности, объективно существующие в экономических и технических системах, так, например, определять:

- надежность автомобилей и дорожно-строительных машин и их агрегатов, зависящую от целого ряда показателей (индикаторов);
- экономический показатель, зависящий от цены приобретения машины, массы и металлоемкости машины, от эксплуатационных издержек, потребляемой энергии и т.п.;
- технологичность, зависящую от трудоемкости изготовления машины, показателя себестоимости изготовления, тягово-эксплуатационных свойств и т.п.;
- топливную экономичность автомобилей, зависящую от технического состояния машин, качества дорог, квалификации водителей, условий эксплуатации и т.п.

Однако все перечисленные показатели не могут быть непосредственно измерены. Они определяются с помощью показателей – индикаторов более низкой ступени иерархии, значения которых фиксируются при проведении испытаний.

Многие из указанных частных признаков- показателей (обозначаемых $x_1, x_2, ..., x_n$) взаимосвязаны и часто дублируют друг друга.

В такого рода ситуациях возникает естественное желание и необходимость: сжать, рафинировать, сконцентрировать, агрегировать имеющуюся информацию, т.е. выразить большое число исходных (частных) признаков x_j через меньшее число более общих теоретических понятий или внутренних характеристик процесса или явления, обозначаемых $P_1, P_2, ..., P_n$ и называемых базовыми параметрами или факторами, которые конденсируют исходный набор признаков, и этим решить задачу снижения размерности описания.

Иначе говоря, явления и процессы в определенной области исследований, несмотря на свою разнородность и изменчивость, могут быть описаны относительно небольшим числом функциональных единиц, параметров, или базовых факторов.

Факторный анализ применяется так же, как отмечалось выше, для оценки непосредственно неизмеряемой величины.

Сущность факторного анализа

Как уже отмечалось выше, на основе статистических наблюдений, т.е. на основе информационного массива получают исходную матрицу $X_{исх}$, устанавливающую зависимость параметра оптимизации Y от связанных с ним факторов $x_1, x_2, \ldots, x_n$.

Так, например, при исследовании эффективности функционирования различных видов автомобилей одним из параметров оптимизации является качество машин. Факториальными показателями (индикаторами) могут служить:

- наработка на отказ (x_1);
- средний межремонтный ресурс (x_2);
- средний срок службы до списания (x_3);
- средний ресурс до капитального ремонта (x_4);
- коэффициент полезного действия (КПД) (x_5);
- трудоемкость изготовления (x_6);
- техническая производительность (x_7);
- трудоемкость обслуживания (x_8);
- энергоемкость (x_9);
- коэффициент взаимозаменяемости частей и агрегатов (x_{10});
- эксплуатационная производительность (x_{11})

и др. показатели.

При анализе топливной экономичности автомобиля одним из параметров оптимизации служит средний расход ГСМ на 100 км пробега, или расход ГСМ на один рабочий день. Факториальными признаками могут служить:

- степень исправности системы зажигания (x_1) ;
- степень исправности системы питания (x_2);
- среднее давление в шинах (x_3);
- квалификация водителей (x_4);
- средняя скорость движения (x_5);
- состояние и качество дорог (x_6);
- соответствие потребляемого ГСМ данному виду машины и т.п. (табл.11)

Отмеченные выше зависимости могут быть представлены в виде табл.3.

Таблица исходных данных, т.е информационный массив в общем виде может быть представлен в виде матрицы

$$X_{исх} = \left\| \begin{array}{cccccccc} y_1 & x_{11} & x_{12} & x_{13} & \dots & x_{1j} & \dots & x_{1n} \\ y_2 & x_{21} & x_{22} & x_{23} & \dots & x_{2j} & \dots & x_{2n} \\ \vdots & & & & & & & \\ y_i & x_{i1} & x_{i2} & x_{i3} & \dots & X_{ij} & \dots & x_{in} \\ \vdots & \vdots & \vdots & \vdots & & \vdots & & \vdots \\ y_{m1} & x_{m1} & x_{m2} & x_{m3} & \dots & x_{mj} & \dots & x_{mn} \end{array} \right\|$$

Таблица 3

Исходная таблица зависимости параметра оптимизации от факториальных признаков

Объекты наблюдения (индивидумы), т.е. номера автомобилей или ДСМ	Функция отклика $У_{опытн}$	Показатели или признаки (индикаторы), описывающие данное явление					
		x_1	x_2	x_3	x_4	...	x_n
1	y_1	x_{11}	x_{12}	x_{13}	x_{14}	...	x_{1n}
2	y_1	x_{21}	x_{22}	x_{23}	x_{24}	...	x_{2n}
⋮	⋮	⋮	⋮	⋮	⋮		⋮
m	y_1	x_{m1}	x_{m2}	x_{m3}	x_{m4}	...	x_{mn}
Среднее арифметическое по каждому из признаков	$\tilde{У}_{ср}$	$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_3$	$\tilde{x}_4$	...	x_n

Построчечные дисперсии	$S^2\{y_i\}$	$S^2\{y_1\}$	$S^2\{y_2\}$	$S^2\{y_3\}$	$S^2\{y_4\}$	...	$S^2\{y_n\}$

Матрица $X_{исх}$ имеет m - строк по числу объектов наблюдения и $n-$столбцов по числу частных показателей или признаков (индикаторов), описывающих данное явление.

Так, в экономических исследованиях в качестве объектов наблюдений выступают предприятия, транспортные объединения, виды продукции, различные виды автомобилей и т.п.Каждый из наблюдаемых объектов может быть охарактеризован целым набором показателей или признаков (индикаторов). Например, при оценке качества и технического устройства автомобилей такими признаками, как уже отмечалось выше, могут служить показатели надежности, производительности, экономический показатель, технологичность, эргономичность и т.п.

Однако, такая таблица (информационный массив, табл. 6) недостаточна для понимания явления. Только регрессионно - корреляционный или факторный анализ позволят выявить закономерности, скрытые за набором указанных чисел.

Применяя правила регрессионно – корреляционного анализа на основе исходной матрицы по формуле второго центрального смешанного момента, может быть вычислена матрица корреляционных моментов связи K_{ij} и матрицы простых (парных) коэффициентов корреляции R_{ij}.

Применяя факторный анализ все частные признаки x_1, x_2, x_3 и параметр оптимизации предварительно подвергают операции центрирования и нормирования, т.е. переходят к стандартизованным переменным z_{ij} по формуле

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\delta_i}$$

Это значит, что для стандартизованных переменных z средние арифметические значения факторов равны нулю, а средние квадратические отклонения равны единице.

При таком условии простые (парные) коэффициенты корреляции выражаются с помощью более простых зависимостей

$$r_{ik} = \frac{Kx_i x_k}{\delta_{x_i}\delta_{x_k}} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_{ij} - \bar{x}_i)(x_{kj} - \bar{x}_k)}{\delta_{x_i}\delta_{x_k}} = \frac{1}{n-1}\sum_{i=1}^{n} z_{ij} z_{kj} \quad (22)$$

Корреляция как скалярное произведение

В факторном анализе коэффициент корреляции r_{ij}, рассматривается как скалярное произведение двух векторов

$$r_{x_1x_2} = |x_1| \cdot |x_2| \cdot \cos\left(\tilde{o}_1 \hat{\tilde{o}}_2 \right).$$

Действительно, если коэффициент корреляции равен нулю, то это значит, что Cos между векторами равен 90° ($\cos 90^\circ = 0$) т,е, вектора перпендикулярны друг другу и ^{0} являются независимыми.

Если же угол между векторами равен О, то $\cos 0^\circ = 1$, и между векторами существует жесткая зависимость.

Если угол между векторами лежит в пределах от нуля до 90^0, то в этом случае скалярное произведение двух векторов (коэффициент корреляции) равно проекции одного вектора на другой.

Заметим также, что коэффициенты корреляции r_{ij} вычисляются с ошибками, т.е. имеет место разброс вычисленных коэффициентов корреляции относительно их истинных значений. Указанный разброс характеризуется дисперсиями разброса.

Полная дисперсия разброса коэффициента корреляции состоит из следующих компонент:

$$\delta_r^2 = \left\lfloor \delta_{r_1}^2 + \delta_{r_2}^2 + ... + \delta_{r_n}^2 + \delta_{r_{cïåö}}^2 + \delta_{r_{rîø}}^2 \right\rfloor \qquad (22)$$

где $\left(\delta_{r_1}^2 + \delta_{r_2}^2 + ... + \delta_{r_n}^2 \right)$ - общая дисперсия;

$\delta_{r_{спец}}^2$ - специфическая дисперсия;

$\delta_{r_{ош}}^2$ - дисперсия, обуславливаемая ошибками;

Общая дисперсия – это та часть полной дисперсии, которая коррелирует с другими переменными (не менее чем с двумя другими переменными).

Специфическая дисперсия – это та часть полной дисперсии, которая присуща лишь одной переменной и не коррелирует со всеми остальными.

Дисперсия, обусловленная ошибками замера, ошибками выборки, неточностями инструментов, применяемых для наблюдений и регистрации показателей, неучетом изменившихся условий эксперимента и т.п.

Если равенство пронормировать, т.е. разделить на полную вероятность, то в этом случае, получаем:

$$1 = \underbrace{\left[\delta_{r_1}^2 + \delta_{r_2}^2 + \ldots + \delta_{r_n}^2\right]}_{h_r^2 - общность} + \underbrace{\left[\delta_{r_{спец}}^2 + \delta_{rош}^2\right]}_{U_r^2 - характерность} \qquad (23)$$

или

$$1 = h_r^2 + U_r^2. \qquad (24)$$

Корреляционная матрица K_{ij} является квадратной и симметричной.

Если диагональные элементы корреляционной матрицы равны единице, то такая корреляционная матрица называется полной.

Если исходные показатели выражаются в стандартизованных переменных Z, то в этом случае моменты связи численно равны коэффициентам корреляции. В матричной форме матрица коэффициентов корреляции записывается так

$$R_{ij} = K_{ij} = \frac{1}{n-1} ZZ^T, \qquad (25)$$

где R_{ij} - матрица простых (парных) коэффициентов корреляции, размерностью $n \times n$;

K_{ij} - матрица моментов связи;

$1/(n-1)$ - поправка на число наблюдений(скаляр);

Z - матрица стандартизованных исходных переменных порядка $m \times n$;

Z^T - транспонированная матрица стандартизованных исходных данных.

Как будет показано ниже, сущность факторного анализа состоит в том, что исходная (полная) матрица(размерностью $n \times n$) коэффициентов корреляции R_{ij}, вычисленная на основе исходных переменных $x_1, x_2, \ldots x_n$ аппроксимируется матрицей коэффициентов корреляции базовых показателей меньшей размерности (порядка $n \times r$, где $r < n$), называемой факторной матрицей.

Факторная матрица имеет столько столбцов, сколько было выделено базовых факторов,и столько строк, сколько было переменных (индикаторов) в исходной матрице $X_{исх}$.

Занятие 7

Математические основы факторного анализа

Целью любого метода, применяемого при факторном анализе является представление величины Z_{ij}, элемента матрицы Z, в виде линейной комбинации нескольких базовых факторов. Обычно считают, что значение Z_{ij} может быть выражено в виде линейной зависимости от базовых факторов

$$Z_{ij} = a_{i1}P_{1j} + a_{i2}P_{2j} + ... + a_{ir}P_{rj}, \qquad (26)$$

где a_{ij} - постоянные коэффициента при базовых факторах, называемые факторными нагрузками, которые надо определить. Факторная нагрузка, как будет показано ниже, это корень квадратный из той части дисперсии переменной, которая обуславливает данным фактором, т.е. это корреляция между переменной и фактором;

$P_{1j} - P_{rj}$ - значения базовых факторов Y_j - того индивидуума;

r - число базовых факторов, количество которых значительно меньше числа исходных признаков- индикаторов, где r значительно меньше $n(r < n)$. Равенство (26) представляет собой основную модель факторного анализа.

Используя матричную форму записи можно записать:

$$Z = A \cdot P, \qquad (27)$$

где $A = (a_{ij})$- матрица факторного отображения (порядка $m \times r$), она же матрица коэффициентов регрессии по базовым факторам. Она отражает связь исходных признаков с базовыми факторами P_j;

$P = (P_{ij})$ - матрица базовых факторов; ее элементы отвечают каждому из индивидиумов,

В уравнении (27) известна лишь матрица Z. Матрицы $A_и P$ неизвестны. Это значит, что если не будет наложено ограничений на матрицу, A то может быть получено множество решений в смысле количества базовых факторов и отвечающих им значений коэффициентов a_{ij}. Задачей факторного анализа является определение матрицы A, т.е. матрицы, элементы которой называются факторными нагрузками.

Очевидно, что если будет найдена матрица A^* и известно Z, тогда не трудно будет найти матрицу базовых факторов P и задача по снижению размерности исследуемого явления будет решена.

Если подставить выражение (26) в выражение (27), тогда получим

$$R = \frac{1}{n-1} ZZ^T = \frac{1}{n-1} (AP)(AP)^T = \frac{1}{n-1} APP^T A^T,$$

или

$$R = A\left[\frac{1}{n-1} PP^T\right] A^T = ACA^T$$

где $C = \frac{1}{n-1} PP^T = c_{ij}$ - матрица корреляции между базовыми факторами.

Если наложить на это равенство условие некоррелированности базовых факторов, т.е. положить $C = E$, тогда

$$R = AA^T, \qquad (28)$$

Полученное выражение называется фундаментальной теоремой факторного анализа.

Фундаментальная теорема утверждает, что полная корреляционная матрица R_{ij}

(размерности $n \times n$) может быть воспроизведена с помощью факторного отображения A и матрицы корреляций C между базисными факторами.

Из уравнения (28), что процедура факторного анализа в известной степени обратима:

Если задана матрица A, то может быть найдена матрица R и наоборот, если задана матрица R , то по ней может быть найдена матрица A. Проиллюстрируем сказанное на примере.

Пример 3. Пусть в результате эксперимента при обработке исходных стандартизованных переменных, характеризующих, например, надежность автомобиля или ДСМ заданного вида, была получена следующая матрица простых парных коэффициентов корреляции

$$R_{ij} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{Vmatrix} r_{11} & r_{12} & r_{13} & r_{14} \\ r_{21} & r_{22} & r_{23} & r_{24} \\ r_{31} & r_{32} & r_{33} & r_{34} \\ r_{41} & r_{42} & r_{43} & r_{44} \end{Vmatrix} \end{array} = \begin{Vmatrix} 1{,}0 & 0{,}72 & 0{,}45 & 0{,}045 \\ 0{,}72 & 1{,}0 & 0{,}40 & 0{,}040 \\ 0{,}45 & 0{,}40 & 1{,}0 & 0{,}025 \\ 0{,}045 & 0{,}04 & 0{,}025 & 1{,}0 \end{Vmatrix} \qquad (29)$$

Рассматривая матрицу (29) видим, что первая и вторая переменные (индикаторы) сильно коррелированны. Третья переменная с первыми двумя связана слабее, а четвертая не зависит от всех предыдущих. Следуя обычной регрессионно-корреляционной процедуре можно было бы проверить значимость каждого коэффициента корреляции и отсеять незначащие коэффициенты. Однако при факторном анализе поступают иначе: указанную матрицу R приближенно

заменяют матричным произведением базового фактора P_1 на его транспонированное значение P_1^T.

Элементы указанного базового фактора P_1 называются факторными нагрузками. Для рассматриваемой матрицы они составляют:

$$P_1 = \begin{Vmatrix} a_1 = 0,9 \\ a_2 = 0,8 \\ a_3 = 0,5 \\ a_4 = 0,05 \end{Vmatrix}.$$

Убедимся, что базовый фактор P_1 эквивалентен матрице R :

$$\widetilde{R} = P_1 P_1^T = \begin{Vmatrix} 0,9 \\ 0,8 \\ 0,5 \\ 0,05 \end{Vmatrix} \begin{Vmatrix} 0,9 & 0,8 & 0,5 & 0,05 \end{Vmatrix} =$$

$$= \begin{Vmatrix} 0,81 & 0,72 & 0,45 & 0,045 \\ 0,72 & 0,64 & 0,4 & 0,04 \\ 0,45 & 0,4 & 0,25 & 0,025 \\ 0,045 & 0,04 & 0,025 & 0,003 \end{Vmatrix}$$

И, значит, полная корреляционная матрица R размерностью 4×4 может быть приближенно заменена матрицей, образуемой одним базовым фактором P_1 размерностью 4×1.

Для рассматриваемого примера вместо вычисления 10 элементов матрицы R достаточно найти четыре факторные нагрузки, образующие базовый фактор

$$P_1 = \begin{Vmatrix} a_1 = 0,9 \\ a_2 = 0,8 \\ a_3 = 0,5 \\ a_4 = 0,05 \end{Vmatrix}.$$

Этим самым будет учтена вся информация, содержащаяся в матрице R .

Погрешность будет иметь место только в диагональных элементах.

Отметим, что диагональные элементы (вычисляемые с погрешностями) называются общностями и обозначаются, $h_1, h_2, h_3, \ldots, h_n$, а, составленная матрица и обозначается $\widetilde{R}$.

Значение общности, записываемое на главной диагонали корреляционной матрицы, выражает ту часть корреляции переменной с самой собой, которую можно приписать влиянию общих факторов [см. ниже].

Если бы на главной диагонали записывались корреляции каждой переменной с самой собой и с учетом специфической дисперсии и дисперсии обусловленной ошибкой замера, то эти корреляции были бы равны ±1.

Порядок определения общностей и базовых факторов рассматривается ниже.

Рассмотрим теперь более сложный случай, когда исходная матрица коэффициентов корреляции заменяется двумя базовыми факторами P_1 и P_2.

Пример 4. Предположим, что задана редуцированная матрица $\widetilde{R}.$, которая имеет вид:

$$\widetilde{R} = \begin{Vmatrix} 0{,}8125 & 0{,}7225 & 0{,}0850 & 0{,}08 \\ 0{,}7225 & 0{,}6425 & 0{,}08 & 0{,}075 \\ 0{,}085 & 0{,}08 & 0{,}6425 & 0{,}5625 \\ 0{,}08 & 0{,}07 & 0{,}5625 & 0{,}4925 \end{Vmatrix}.$$

Требуется воспроизвести ее с помощью базовых факторов. Если мы поступим так, как это было показано в вышерассмотренном примере, т.е. заменим матрицу одним базовым фактором

$$P_1 = (0{,}90;\quad 0{,}80;\quad 0{,}0;\quad 0{,}05),$$

Тогда в первом приближении можно записать

$$R' \cong P_1 P_1^T \begin{Vmatrix} 0{,}9 \\ 0{,}8 \\ 0{,}05 \\ 0{,}05 \end{Vmatrix} \cdot (0{,}9 \quad 0{,}8 \quad 0{,}05 \quad 0{,}05) = \begin{Vmatrix} 0{,}81 & 0{,}72 & 0{,}045 & 0{,}045 \\ 0{,}72 & 0{,}64 & 0{,}04 & 0{,}04 \\ 0{,}045 & 0{,}04 & 0{,}025 & 0{,}025 \\ 0{,}045 & 0{,}04 & 0{,}025 & 0{,}025 \end{Vmatrix}.$$

Как видим, матрица R' воспроизводит исходную матрицу со значительно большими погрешностями, чем в первом случае.

Погрешность равна разности двух матриц и составляет:

$$\Delta R = \widetilde{R} - R' = \widetilde{R} - P_1 P_1^T = \begin{Vmatrix} 0{,}0025 & 0{,}0025 & 0{,}0400 & 0{,}0350 \\ 0{,}0025 & 0{,}0025 & 0{,}0400 & 0{,}0350 \\ 0{,}0400 & 0{,}0400 & 0{,}6400 & 0{,}5600 \\ 0{,}0350 & 0{,}0350 & 0{,}5600 & 0{,}4900 \end{Vmatrix}$$

Если остаточную матрицу ΔR заменить вторым базовым фактором $P_2 = (0{,}05;0{,}05;0{,}80;0{,}70)^T$, то в этом случае исходная матрица R_{ij} может быть уже с меньшей погрешностью заменена двумя базовыми факторами

$$\widetilde{R} = (P_1/P_2)(P_1/P_2)^T = \begin{Vmatrix} 0{,}90 & 0{,}05 \\ 0{,}80 & 0{,}05 \\ 0{,}05 & 0{,}80 \\ 0{,}05 & 0{,}70 \end{Vmatrix} \begin{Vmatrix} 0{,}90 & 0{,}80 & 0{,}05 & 0{,}05 \\ 0{,}05 & 0{,}05 & 0{,}80 & 0{,}70 \end{Vmatrix} =$$

$$= \begin{Vmatrix} 0{,}8125 & 0{,}7225 & 0{,}085 & 0{,}080 \\ 0{,}7225 & 0{,}6425 & 0{,}080 & 0{,}075 \\ 0{,}085 & 0{,}080 & 0{,}425 & 0{,}5625 \\ 0{,}080 & 0{,}0750 & 0{,}56 & 0{,}4925 \end{Vmatrix}$$

Как видим, в рассматриваемом примере исходная редуцированная корреляционная матрица $\widetilde{R}.$ состоящая из 16 элементов может быть заменена двумя базовыми факторами P_1 и P_2.

Матрица факторных нагрузок, отвечающая введенным двум базовым факторам, имеет вид

$$A = \begin{array}{c|c} \mathbf{P_1} & \mathbf{P_1} \\ \hline \begin{Vmatrix} a_{11} = 0{,}90 \\ a_{21} = 0{,}8 \\ a_{31} = 0{,}05 \\ a_{41} = 0{,}05 \end{Vmatrix} & \begin{Vmatrix} a_{12} = 0{,}05 \\ a_{22} = 0{,}05 \\ a_{32} = 0{,}80 \\ a_{42} = 0{,}70 \end{Vmatrix} \end{array}$$

и, значит, классическая модель факторного анализа имеет вид:

$$R \approx AA^T, \text{ или } AA^T = R + D^2, \qquad (30)$$

где D^2 - погрешность, возникающая вследствие замены диагональных элементов (равных единице) общностями.

Указанное равенство может быть сформулировано так: редуцированная матрица коэффициентов корреляции $\widetilde{R}.$ может быть заменена произведением редуцированной матрицы A базовых факторов на отвечающую ей транспонированную матрицу A^T.

В общем виде полная факторная матрица базовых факторов может быть представлена в виде табл. 4.

Таблица 4

Классификация факторов

Переменные (индикаторы)	Общие факторы (базовые)	Характерные факторы	Общности
	P_1 P_2 ... P_r	1 2 … r	P_1 P_2 ...P_r
	Факторные нагрузки		
1	a_{11} a_{12} ...a_{1r}	U_{11}	h_1^2
2	a_{21} a_{22} ...a_{2r}	U_{22}	h_2^2
⋮	⋮ ⋮ ⋮	⋱	
m	a_{m1} a_{m2} ...a_{mr}	U_{mr}	h_r^2
Вклады факторов	V_1 V_2 ... V_r		$\sum_{j=1}^{m} V_i = \sum_{j=1}^{m} h_i^2$

Термин общий, подчеркивает тот факт, что данный фактор имеет существенное значение для анализа всех переменных r_{ij} , точнее, если хотя бы две его нагрузки отличаются от нуля.

Характерный – подчеркивает другой факт, а именно, что он относится только к соответствующей переменной r .

Если все нагрузки значительно отличаются от нуля, то такой общий фактор называют генеральным (табл. 5)

Таблица 5

Факторы*

Переменные	A	B	C	È 1	È 2	È 3	È 4
r_1	+	+		+			
r_2	+	+	+		+		
r_3	+		+			+	
r_4	+		+				+

* B и C – общие факторы;

A – генеральный фактор;

$И_1, И_2, И_3 и И_4$ - характерные факторы.

Методы определения общностей и базовых факторов. Центроидный метод

Преобразование корреляционной матрицы R вычисленной на основе исходной матрицы $X_{исх}$, в матрицу базовых факторов может производиться с помощью различных методов, например:

- с помощью метода главных факторов;
- с помощью метода главных компонент;
- с помощью центроидного метода;
- с помощью метода максимального правдоподобия;
- с помощью метода минимальных остатков и других.

По дидактическим соображениям, рассмотрим вначале центроидный метод. Центроидный метод является упрощенным аппроксимационным вариантом метода главных факторов. Его основное достоинство – доступность для ручного счета. В связи с указанным, рассмотрим порядок определения базовых факторов на конкретном числовом примере с помощью центроидного метода.

Пример 5. Предположим, что на основе информационного массива была получена следующая полная корреляционная матрица (табл. 6) для шести признаков – индикаторов).

Таблица 6

	$^{ч}11$	$^{ч}12$	$^{ч}13$	$^{ч}14$	$^{ч}15$	$^{ч}16$
$^{ч}11$	1	0,229	0,400	0,297	0,116	0,232
$^{ч}12$		1	0,568	0,537	0,432	0,154
$^{ч}13$			1	0,487	0,436	0,071
$^{ч}14$				1	0,545	0,092
$^{ч}15$					1	0,016
$^{ч}16$						1

Требуется преобразовать указанную корреляционную матрицу в матрицу базовых факторов.

Решение;

1. Для этого, прежде всего, составляем редуцированную матрицу, $\widetilde{R}$ у которой по главной диагонали записаны общности h^2. Определение общностей представляет известную трудность. Дело в том, что общности h^2 не могут быть определены экспериментальным путем. Существуют различные способы определения общностей, один из которых состоит в том, что по главной диагонали записываются наибольшие значения коэффициентов корреляции в столбце или в строке, что одно и то же, т.к. корреляционная матрица R симметрична. Применяя этот

метод для рассматриваемого примера ,получаем редуцированную матрицу коэффициентов корреляции (табл.7).

2. Определяем нагрузки первого базового фактора:

а) суммируем элементы каждой строки (столбца) включая h_i с учетом алгебраических знаков и получаем $\sum r_i$ (см. предпоследний столбец в табл.7.

Таблица 7

Редуцированная матрица коэффициентов корелляции

	ч11	ч12	ч13	ч14	ч15	ч16	чi	a_i
ч11	0,400	0,299	0,400	0,297	0,166	0,232	1,744	0,500
ч12	0,229	0,568	0,568	0,534	0,432	0,154	2,555	0,733
ч13	0,400	0,568	0,568	0,487	0,436	0,071	2,530	0,726
ч14	0,297	0,534	0,487	0,545	0,545	0,092	2,500	0,717
ч15	0,116	0,434	0,436	0,545	0,545	0,016	2,058	0,590
ч16	0,234	0,154	0,071	0,160	0,016	0,232	0,765	0,219
$\sum r_i$	1,744	2,555	2,530	2,500	2,058	0,765	12,152	
a_i	0,500	0,733	0,726	0,717	0,590	0,219		

а) Складываем все суммы предпоследнего столбца и находим

$$T = \sum\sum r_i .$$

Для рассматриваемого примера эта сумма равна Т=12,152 и, следовательно,

$$\sqrt{T} = \sqrt{12{,}152} = 3{,}48859.$$

б) Находим нагрузки первого фактора по формуле

$$a_i = \Sigma r_i \big/ \Sigma\Sigma r_i .$$

Для рассматриваемого примера они составляют (см. последнюю строку): 0,50;0,733 … и т.д.

3. Определяем нагрузки для второго базового фактора. Как было показано выше (см. пример 4) для определения нагрузок второго базового фактора необходимо найти остаточную матрицу ΔR

$$\Delta R = \widetilde{R} - R' = \widetilde{R} - P_1 P_1^T .$$

Покажем порядок подсчета первых остатков в остаточной матрице:

$$\Delta r_{z_1 z_1} = 0{,}400 - (0{,}500 \times 0{,}500) = 0{,}150;$$

$$\Delta r_{z_1 z_2} = 0{,}299 - (0{,}500 \times 0{,}733) = -0{,}067;$$

$$\Delta r_{z_1 z_3} = 0{,}400 - (0{,}500 \times 0{,}726) = 0{,}037;$$

$$\Delta r_{z_1 z4} = 0{,}297 - (0{,}500 \times 0{,}717) = -0{,}061;$$

$$\Delta r_{z1 z5} = 0{,}116 - (0{,}500 \times 0{,}500) = -1{,}179$$

$$\Delta r_{z_1 z_6} = 0{,}232 - (0{,}500 \times 0{,}219) = 0{,}123.$$

Аналогично получаем:

$$\Delta r_{z_2 z_1} = 0{,}299 - (0{,}500 \times 0{,}733) = -0{,}067;$$

$$\Delta r_{z_2 z_2} = 0{,}568 - (0{,}733 \times 0{,}733) = 0{,}031;$$

$$\Delta r_{z_2 z_3} = 0{,}568 - (0{,}733 \times 0{,}726) = 0{,}036$$

На основе этого составляется матрица первых остатков (табл. 8).

Матрица первых остатков

	$\div_{11}$	$\div_{12}$	$\div_{13}$	$\div_{14}$	$\div_{15}$	$\div_{16}$
$\div_{11}$	(0,150)	-0,067	0,037	-0,061	-0,179	0,123
$\div_{12}$	-0,067	(0,031)	0,036	0,009	0	-0,06
$\div_{13}$	0,037	0,036	0,041	-0,033	0,008	-0,088
$\div_{14}$	-0,061	0,009	-0,033	(0,031)	0,122	-0,065
$\div_{15}$	-0,179	0	0,008	0,122	(0,197)	-0,145
$\div_{16}$	0,123	0,006	-0,088	-0,065	-0,145	(0,185)
$\sum_{i=1}^{n} \div_i$	0,003	0,003	0,001	0,003	0,003	0,004

Если сейчас поступить так, как это было показано при определении первого фактора (табл.8), то оказывается, что сумма элементов по столбцам будет близка к нулю; положительные и отрицательные значения r_{ij} уравновешиваются. В связи с этим, предварительно необходимо выполнить операцию обращения алгебраических знаков в матрице остатков. Для этого меняем знаки одного из признаков в одном столбце и в одноименной строке. Далее факторные нагрузки второго фактора на признаки рассчитываются по той же схеме, т.е. сумма эле-

ментов соответствующего столбца делится на сумму всех элементов матрицы. Окончательное значение факторных нагрузок получается путем восстановления первоначальных знаков у тех признаков, у которых они были изменены. Для рассматриваемого примера, применяя центроидный метод, была получена следующая матрица базовых факторов (табл. 9)

Таблица 9

Переменные (показатели)	Базовые факторы		
	P1	P2	P3
1	0,500	0,365	0,145
2	0,733	-0,118	0,075
3	0,726	-0,095	0,312
4	0,717	-0,220	-0,155
5	0,590	-0,404	-0,194
6	0,219	0,365	-0,099

Извлечение факторов продолжается до достижения некоторого заранее установленного порога.

Надо заметить, что операция нахождения базовых факторов будет надежной только тогда, когда известны общности. С другой стороны общности могут быть определены только тогда, когда известны базовые факторы. Таким образом, получается нечто похожее на заколдованный круг. Поэтому процедура факторного анализа состоит в том, что начинают с первого приближения факторных нагрузок, рассчитанных любым методом. Затем, путем последовательных приближений получают все более точные (более надежные) их значения, т.е. применятся метод итерации.

Занятие 8

Интерпретация результатов факторного анализа

Пусть при проведении испытаний над 33 объектами (под объектами подразумевается автомобили) по вышеотмеченным 11 показателям была получена следующая исходная информационная матрица данных.

Пример 6.

$$X_{исх} = \begin{Vmatrix} x_{11} & x_{12} & \cdots & x_{1.11} \\ x_{21} & x_{22} & \cdots & x_{2.11} \\ \vdots & \vdots & & \vdots \\ x_{33.1} & x_{33.2} & \cdots & x_{33.11} \end{Vmatrix}$$

Таблица 10

Матрица частных коэффициентов корелляции R

Показа-тели	1	2	3	4	5	6	7	8	9	10	11
1	1	0,63	0,75	0,17	-0,25	-0,06	-0,32	0,31	-0,16	0	-0,35
2		1	0,62	0,13	0,04	-0,02	-0,06	0,37	0,09	0,06	-0,04
3			1	0,13	0,07	0,24	-0,01	0,35	-0,19	0,16	0,03
4				1	0,03	0,33	0,45	0,49	0,17	0,39	0,32
5					1	-0,10	0,48	-0,07	0,29	-0,18	0,50
6						1	0,52	0,61	0,24	0,82	0,52
7							1	0,46	0,37	0,60	0,94
8								1	0,34	0,82	0,32
9									1	0,43	0,29
10										1	0,54
11											1

На основе исходной матрицы была рассчитана матрица частных коэффициентов корреляции R (табл. 10).

Применяя один из методов факторного анализа матрица R была преобразована в матрицу трех базовых факторов $P_1, P_2 и P_3$ (табл. 11)

Таблица11

Матрица базовых факторов и факторных нагрузок

Показатели или признаки (индикаторы)	Базовые факторы			Общности h_j^2
	P_1	P_2	P_3	
	Факторные нагрузки			
1	-0,23	0,06	0,88	0,839
2	0,05	-0,02	0,86	0,751
3	-0,07	0,14	0,86	0,762
4	0,26	0,58	0,22	0,415
5	0,93	-0,18	0	0,895
6	0,04	0,87	-0,03	0,758
7	0,74	0,59	-0,13	0,912
8	0,06	0,83	0,37	0,831
9	0,45	0,47	-0,18	0,456
10	0,06	0,97	0	0,943
11	0,74	0,50	-0,13	0,811

Признаки получившие наибольшую факторную нагрузку	5,7 и 11	6,8 и 10	1,2 и 3	
Вклад фактора $\sum_{j=1}^{11} a_{ip}^2$	2,28	3,6	2,5	8,38

Произведем интерпретацию полученных результатов.

Как было показано выше при проведении регрессионно-корреляционного анализа основной задачей является получение математической модели явления, определение коэффициентов корреляции и детерминации и построение, на основе этого, диаграммы влияния каждого из факторов на параметр оптимизации . Следовательно, сами уравнения регрессии являются конечной целью исследования.

При проведении факторного анализа основной целью исследования является задача иного характера, а именно получение достаточно надежной факторной матрицы (см., например, табл. 14). Если эта задача решена, тогда возникает задача другой направленности, а именно: распознать природу полученных факторов. Это трудная и тонкая задача, требующая, прежде всего изменения позиции исследователя, применяющего факторный анализ. Теперь он должен превратиться из статиста, заботящегося в первую очередь о правильности и точности вычислений, в эксперта по проблеме, закономерности которой исследовались с помощью факторного анализа.

При решении задачи до получения факторной матрицы преобразования производились безотносительно к конкретному содержанию переменных, т.е. признаков или индикаторов.

После получения факторной матрицы ситуация существенно изменяется. Теперь уже необходимо взять на вооружение все наши знания о показателях (переменных), т.е. об индикаторах подвергшихся факторному анализу в объектах исследования.

Таким образом, интерпретация факторов сводится к анализу величины и, главное, знаков нагрузок. Поиск названия факторов – совершенно неформализованная процедура. Этот этап интерпретации целиком и полностью зависит от интуиции и уровня осведомленности исследователя в изучаемой им области, ибо процесс формирования названия факторов есть процесс поиска среди известных исследователю понятий именно того понятия, которое в наибольшей степени соответствует конституируемому нагрузками взаимоотношению в изменении исследуемых эмпирических признаков.

В рассматриваемом примере 6 первый базовый фактор определяется тесно связанными показателями 5, 7 и 11, который может быть интерпретирован как производительность машин.

Второй фактор определяется тесно связанными показателями 6, 8 и 10 и может интерпретироваться как технологичность машин.

Третий фактор определяется тесно связанными показателями 1, 2 и 3 и может интерпретироваться как надежность машин.

При этом показатели 4 и 9 отсеиваются как малозначащие.

Рассмотренный пример приведен лишь для пояснения сущности факторного анализа и пояснения последовательности действий исследователя. В рассмотренном примере решена задача типизации или классификации показателей.

С помощью факторного анализа могут также решаться задачи другого характера, например:

- задачи повышения эффективности функционирования автопроизводственных предприятий, функционирования СТОА, закрытых парков технического обслуживания и т. д.;

- задачи по определению оптимальных видов и конструкции автомобилей , для решения которых применяются показатели, имеющие многоступенчатую иерархическую структуру ;

- задачи по повышению производительности труда и т.п.

В некоторых случаях для анализа исследуемых систем могут быть составлены уравнения регрессии, построенные на базовых факторах. Такие модели обладают определенными преимуществами перед моделями обычного регрессионно-корреляцинного анализа.

Естественно, что при решении всех выше перечисленных задач необходимо основываться на опытном информационном массиве достаточно большого объема и задачи решать с помощью электронно-вычислительных машин.

ЛИТЕРАТУРА

1. Вентцель Е.С. Теория вероятностей. М.: Наука, 1969.

2. Завадский Ю.В. Статистическая обработка эксперимента. М.: Высшая школа,1976.

3. Завадский Ю.В. Методика статстической обработки эксперименталых данных. М: МАДИ, 1973.

4. Вентцель Е.С.,Овчаров Л.А. Теория вероятностей и ееинженерное приложение

Приложения

Таблицы для вероятностных расчетов

Плотность распределения нормального закона для нормированной и центрированной случайной величины *t*

	Сотые доли *t*									
f	0	1	2	3	4	5	6	7	8	9
0,0	0,398	398	398	398	398	398	398	398	397	397
0 1	397	396	396	395	395	394	393	393	392	391
0 2	391	390	389	388	387	386	385	384	383	382
0 3	381	380	379	377	376	375	373	372	171	369
0 4	368	366	365	363	362	360	358	357	355	353
0 5	352	350	348	346	344	342	341	339	337	335
0 6	333	331	329	327	325	323	320	318	316	314
0.7	312	310	307	305	303	301	298	296	294	292
0 8	289	287	285	282	280	278	275	273	270	268
0 9	266	263	261	258	256	254	251	249	246	244
1.0	0.242	239	237	234	232	229	227	225	222	220
1 1	217	215	213	210	208	205	203	201	198	196
1.2	194	191	189	187	184	182	180	178	175	173
1.3	171	169	166	164	162	160	158	156	153	151
1 4	149	147	145	143	141	139	137	135	133	131
1 5	129	127	125	123	121	120	118	116	114	112
1.6	110	109	107	105	104	102	100	098	097	095
1.7	094	092	090	089	087	086	084	083	081	080
1.8	079	077	076	074	073	072	070	069	068	066
1.9	065	064	063	062	060	059	058	057	056	055
2.0	0.054	052	051	050	049	048	047	046	045	044
2 1	044	043	042	041	040	039	037	037	037	036
2.2	035	034	033	033	032	031	031	030	029	029
2 3	028	027	027	026	025	025	024	024	023	029
2 4	022	021	021	020	020	019	019	018	018	018
2 5	017	017	016	016	015	015	015	014	014	013
2.6	013	013	012	012	012	011	011	011	011	ОЮ
2 7	010	010	009	009	009	009	008	008	008	008
2 8	007	007	007	007	007	006	006	ОО6	006	006
2 9	006	005	005	005	005	005	005	004	004	004
3,0	0,00^	1 004	ОС	Ю04	ООЗ	ООЗ	ООЗ	003	003	003

Примечание. Все значения- вероятностей, помещенные в табл. 3,4,6,9, и 10 приложения 1 меньше единицы. Поэтому в таблицах приведены лишь десятичные знаки, следующие после запятой, перед которыми при пользовании таблицами нужно ставить нуль.

Таблица 2

Функции распределения нормального закона

t	$\Phi^*(t)$	t	$\Phi^*(t)$	t	$\Phi^*(t)$	t	$\Phi^*(t)$
-3,90	0,00	-1,70	0,044	-0,10	0,460	+1,40	0,919
-3.20	0.00	-1.60	0.054	-o.oo	0.500	+1.50	0.933
-ЗЛ0	0.001	-1.50	0.066	+0.00	0.500	+1.60	0.945
-3.00	0.001	-1.40	0.080	+0.10	0.539	+1.70	0.955
-2.90	0.001	-1.30	0.096	+ 0.20	0.579	+1.80	0.964
-2.80	0.002	-1.20	0.115	+0.30	0.617	+1.90	0.971
-2.70	0.003	-1.10	0.135	+0.40	0.655	+2.00	0.977
-2.60	0.004	-1.00	0.158	+0.50	0.691 ■	+2.10	0.982
-2.50	0.006	-0.90	0.184	+0.60	0.725 I	+2.20	0.986
-2.40	0.008	-0.80	0.211	+0.70	0.758	! +2.30	0.989
-2.30	0.010	-0.70	0.242	+0.80	0.788	+2.40	0.991
-2.20	0.013	-0.60	0.274	+0.90	0.815	+2.50	0.993
-2.10	0.01 7	-0.50	0.308	+1.00	0.841	+2.60	0.995
-2.00	0.022	-0.40	0.344	+1.10	0.864	+2.80	0.997
-1.90	0.028	-0.30	0.382	+1.20	0.884	+3.00	0.998
-1,80	0,035	•-0,20	0,420	+1,30	0,903	+3,90	1,000

Таблица 3

Значения функции распределения Лапласа нормального закона для нормированной и центрированной случайной величины *t*

t	Сотые доли *t*									
	0'	1	2	3	4	5	6	7	8	9
0,0	000	008	016	023	031	039	047	055	063	071
0.1	079	087	095	103	111	119	127	135	142	150
0.2	158	166	174	181	189	197	205	212	220	228
0.3	235	243	251	258	266	273	281	288	296	303
0.4	310	318	325	332	■340	347	354	361	368	375
0.5	382	389	396	403	410	417	424	431	438	444
0.6	451	458	464	471	477	484	490	497	503	509
0.7	516	522	528	534	540	546	552	558	564	570
0.8	576	582	587	593	599	604	610	615	621	626
0.9	633	637	642	647	652	657	662	667	672	677
1.0	682	687	69 2	697	701	706	710	715	719	724
1Л	728	733	737	741	745	749	754	758	762	766
1.2	769	773	777	781	785	788	79 2	795	799	802
1.3	806	809	813	816	819	823	826	829	832	835
1.4	838	84l!	844	847	850	852	855	858	861	863
1.5	866	869	871	874	878	878	881	883	885	888
1.6	890	892	894	896	899	901	903	905	907	909
1.7	910	912	914	916	918	919	9 21	9 23	924	9 26
1.8	928	929	931	932	934	935	937	938	939	941
1.9	942	943	945	946	947	948	950	951	952	953
2.0	954	955	956	967	958	959	960	961	962	963
2.1	964	965	966	966	966	968	969	970	970	971
2.2	972	972	973!	974	974	975	976	976	977	978
2.3	978	979	979	980	980	981	981	982	982	983
2.4	983	984:	984	984	985	985	986	986	986	987
2.5	987	987	988	988	988	989	989	989	990	990
2.6	990	991!	991	991	991	992	992	992	992	992
2.7	993	993	993	993	993	994	994	994	994	994
2.8	994	995	995	995	995	995	995	995	996	996
2.9	996	996	996	996	996	996	996	997	997	997
3,0	997	997	99 7	997	997	997	997	997	997	998

Таблица 4

Критические точки (квантили) распределения χ^2 Пирсона в зависимости от уровня значимости α и числа степеней свободы *К*

К	Уровень значимо-				К	Уровень значимости α			
	0,01	0,05	0,10	0,20		0,01	0,05	0,10	0,20
1	6,3	3,8	2,7	. 1.6	16	32,0	26,2	23,5	20,4
2	9.2	5.9	4.0	3.2	17	33.4	27.5	24.7	21.6
3	11.3	7.8	6.2	4.6	18	34.8	28.8	25.9	22.7
4	13.2	9.4	7.7	5.9	19	36.1	30.1	27.2	23.9
5	15.0	11.0	9.2	7.2	20	37.5	31.4	28.4	25.0
6	16.8	12.5	10.6	8.5	21	38.9	32.6	29.6	26.1
7	18·4	14.0	12.0	9.8	22	40.2	33.9	30.8	27.3
8	20.0	15.5	13.3	11.0	23	41.6	35.1	32.0	28.4
9	21.6	16.9	14.6	12.2	24	42.9	36.4	33.1	29.5
10	23.2	18.3	15.9	13.4	25	44.3	37.6	34.3	30.6
11	24.7	19.6	17.2	14.6	26	45.6	38.8	35.5	31.7
12	26.2	21.0	18.5	15.8	27	46.9	40.1	36.7	32.9
13	27.6	22.3	19.8	16.9	28	48.2	41.3	37.9	34.0
14	29.1	23.6	21.0	18.1	29	49.5	42.5	39.8	35.1
15	30,5	24,9	22,3	19,3	30	50,8	43,7	40,2	36,2

Таблица 5

Критические значения вероятностей критерия Колмогорова

λ	***P(λ)***	λ	***P (λ)***	λ	***P(λ)***
0,0	1,0	0,7	0,711	1,4	0,040
0,1	1,0	0,8	0,544	1,5	0,022
0,2	1,0	0,9	0,393	1,6	0,012
0,3	1,0/	1,0	0,270	1,7	О.ООб
0,4	0,977	1,1	0,178	1,8	0,003
0,5	0,964	1,2	0,112	1,9	0,002
0,6	0,864	1,3	0,068	2,0	0,001

Таблица 6

Значения функции распределения Стьюдента вычисленные в зависимости от аргумента t и числа испытаний *n*

t / n	2	3	4	c;	6	7	8	9	10	20	∞
o,1	064	070	074	074	076	076	076	078	078	078	080
0.2	126	140	146	148	150	152	152	154	154	156	159
0.3	188	208	216	220	224	226	228	228	228	232	236
0,4	242	272	284	290	294	296	300	300	302	306	311
0,5	296	334	348	356	362	366	368	370	372	378	383
0.6	344	390	410	420	'426	430	434	430	476	444	452
0.7	388	444	466	478	484	490	494	496	498	508	516
0.6	430	492	518	532	540	546	550	554	556	566	576
0,9	466	538	566	580	590	598	602	606	608	620	632
1,0	500	578	608	626	636	644	650	654	656	670	683
1.1	530	614	648	668	678	686	692	696	700	716	726
1.2	558	648	684	704	716	724	730	736	740	766	776
1.3	582	676	716	736	750	758	766	770	774	.790	806
1,4	606	704	744	766	780	788	796	800	804	822	838
1,5	626	728	770	792	806	816	822	828	832	8.50	866
1.6	644	750	792	816	830	840	846	852	856	874	890
1.7	662	768	812	836	850	860	868	872	876	894	910
1.8	678	786	830	854	874*	878	886	890	894	912	928
1,9	692	802	846	870	884	894	900	906	910	928	943
2,0	704	816	860	884	898	908	914	920	924	940	954
2.2	728	842	884	908	920	930	936	940	944	960	972
2.4	748	862	904	926	938	946	952	956	960	974	984
2,6	766	676	9 20	940	952	960	964	968	972	982	991
3,0	796	904	942	960	970	976	980	984	984	992	997
3.4	818	924	958	972	980	986	988	990	992	996	999
3.8	836	938	968	980	988	992	994	996	997	998	999

Таблица 7

Критические точки распределения критерия Стьюдента t крит. (α, к) зависимости от уровня значимости α и числа степеней свободы К = n-1

К		**Уровень значимости** α				**К**	Уровень значимости α				
	о,1	0,05	0,02	0,01	0,001		0,1	0,05	0,02	0,01	0,001
1	6,31	12,71	31,82	63,66	636,62	18	1,73	2,10	2,55	2,88	3,92
2	2,92	4,30	6,97	9,93	31,60	19	1,73	2,09	2,54	2,86	3,88
3	2,35	3,18	4,54	5,84	12,94	20	1,73	2,09	2,53	2,85	3,85
4	2,13	2,78	3,75	4,60	8,61	21	1,72	2,08	2,52	2,83	3,82
5	2,02	2,57	3,37	4,03	6,86	22	1,72	2,07	2,51	2,82	3,79
6	1,94	2,45	3,14	3,71	5,96	23	1,71	2,07	2,50	2,81	3,77
7	1,90	2,37	3,00	3,50	>5,41	24	1,71	2,06	2,49	2,80	3,75
8	1,86	2,31	2,90	3,36	5,04	25	1,71	2,06	2,48	2,79	3,73
9	1,83	2,26	2,82	3,25	4,78	26	1,71	2,06	2,48	2,78	3,71
10	1,81	2,23	2,76	3,11	4,59	27	1,70	2,05	2,47	2,77	3,69
11	1,80	2,20	2,72	3,11	4,44	28	1,70	2,05	2,47	2,76	3,67
12	1,78	2,18	2,68	3,06	4,32	29	1,70	2,04	2,46	2,76	3,66
13	1,77	2,16	2,65	3,01	4,22	30	1,70	2,02	2,46	2,75	3,65
14	1,76	2,15	2,62	2,98	**4,14**	40	1,68	2,02	2,42	2,70	3,55
15	1,75	2,13	2,60	2,95	4,07	60	1,67	2,00	2,39	2,66	3,46
16	1,75	2,12	2,58	2,92	4,02	120	1,66	1,98	2,36	2,62	3,37
17	1,74	2,11	2,57	2,90	3,97	∞	1,65	1,96	2,33	2,58	3,29

Таблица 8

Квантили распределения критерия Фишера F крит.

k_2 \ k_1	1	2	3	4	5	6	12	24	∞
2.	18	19	19	19,3	19,3	19,3	19,4	19,5	19,5
3.	10	9.6	9.3	9.1	9.0	8.9	8.7	8.6	8.5
4.	7.7	6.9	6.6	6.4	6.3	6.2	5.9	5.8	5.6
5.	6.6	5.8	5.4	5.2	5.1	5.0	4.7	4.5	4.4
6.	6.0	5.1	4.8	4.5	4.4	4.3	4.0	3.8	3.7
7.	5.6	4.7	4.4	4.1	4.0	3.9	3.6	3.4	3.2
8	5.3	4.5	4.1	3.8	3.7	3.6	3.3	3.1	2.9
9.	5.1	4.3	3.9	3.6	3.5	3.4	3.1	2.9	2.7
10.	**5.0**	4.1	3.7	3.5	3.3	3.2	2.9	2.7	2.5
11.	4.8	4.0	3.6	3.4	3.2	3.1	2.8	2.6	2.4
12.	4.8	3.9	3.5	3.3	**3.1**	3.0	2.7	2.5	2.3
13.	4.7	3.8	3.4	3.2	3.0	2.9	2.6	2.4	2.2
14.	4.6	3.7	3.3	3.1	3.0	2.9	2.5	2.3	2.1
15.	4.5	3.7	3.3	3.1	2.9	2.8	2.5	2.3	2.1
16.	4.5	3.6	3.2	3.0	2.9	2.7	2.4	2.2	2.0
17.	4.5	3.6	3.2	3.0	2.8	2.7	2.4	2.2	2.0
18.	4.4	3.6	3.2	2.9	2.8	2.7	2.3	2.1	1.9
19.	4.4	3.5	О і Д.	2.9	2.7	2.6	2.3	2.1	1.8
20.	4.4	3.5	3.1	2.9	2.7	2.6	2.3	2.1	1.8
22.	4.3	3.4	3.1	2.8	2.7	2.6	2.2	2.0	1.8
24.	4.3	3.4	3.0	2.8	2.6	2.5	2.2	2.0	1.7
26.	4.2	3.4	3.0	2.7	2.6	2.4	2.1	1.9	1.7
28.	4.2	3.3	2.9	2.7	2.6	2.4	2.1	1.9	1.6
30.	4.2	**3.3**	2.9	2.7	2.5	2.4	2.1	1.9	1.6
40.	4.1	3.2	2.9	2.6	2.5	2.3	2.0	1.8	1.5
60.	4.0	3.2	2.8	2.5	2.4	2.3	1.9	1.7	1.4
120.	3.9	3.1	2.7	2.5	2.3	2.2	1.8	1.6	1.3
∞	3,8	3,0	2,6	2,4	2,2	2,1	1.8	1.5	1,0

Таблица 9

Квантили распределения критерия Фишера F к р и т . при α = 0,01, где K_1 и K_2 - соответственно числа степеней свободы большей и меньшей дисперсий

k_1	1	2	3	4	5	6	12	24	0O
2	98,5	99,0	99,2	99,3"	99,3	99,4	99,4	99,5	99,5
3	34,1	30,8	29,5	28,7	28,2	27,9	27,1	26,6	26,1
4	21,2	18,0	16,7	16,0	15,5	15,2	14,4	13,9	13,5
5	16,3	13,3	12,1	11,4	11,0	10,7	9,9	9,5	9,0
6	13,7	10,9	9,8	9,2	8,8	8,5	7,7	7,3	6,9
7	12,3	9,6	8,5	7,9	7,5	7,2	6,5	6,1	5,7
8	11,3	8,7	7,6	7,0	6,6	6,4	5,7	5,3	4,9
9	10,6	8,0	7,0	6,4	6,1	5,8	5,1	4,7	4,3
10	10,0	7,6	6,6	6,0	5,6	5,4	4,7	4,3	3,9
11	9,7	7,2	6,2	5,7	5,3	5,1	4,4	4,0	3,6
12	9,3	6,9	6,0	5,4	5,1	4,8	4,2	3,8	3,4
13	9,1	6,7	5,7	5,2	4,9	4,6	4,0	3,6	3,2
14	8,9	6,5	5,6	5,0	4,7	4,5	3,8	3,4	3,0
15	8,7	6,4	5,4	4,9	4,6	4,3	3,7	3,3	2,9
16	8,5	6,2	5,3	4,8	4,4	4,2	3,6	3,2	2,8
17	8,4	6,1	5,2	4,7	4,3	4,1	3,5	OgX	2,7
18	8,3	6,0	5,1	4,6	4,3	4,0	3,4	3,0	2,6
19	8,2	5,9	5,0	4,5	4.2	3,9	3,3	2,9	2,4
20	8,1	5,9	4,9	4.4	4,1	3,9	3,2	2,9	2,4
22	7,9	5,7	4,8	4,3	4,0	3,8	3,1	2,8	2,3
24	7,8	5,6	4,7	4,2	3,9	3,7	3,0	2,7	2,2
26	7,7	5,5	4,6	4,1	3,8	3,6	3,0	2,6	2,1
28	7,6	5,5	4,6	4,1	3,8	3,5	2,9	2,5	2,1
30	7,7	5,4	4,5	4,0	3,7	3,5	2,8	2,5	2,0
40	7,3	7,5	4,3	3,8	3,5	3,3	2,7	2,3	1,8
60	7,1	5,0	4,1	3,7	3,3	3,1	2,5	2,1	1,6
20	6,9	4,8	4,0	3,5	3,2	5,0	2,3	2,0	1,4
∞	6.6	4.6	3,8	3,3	3,0	2,8	2,2	1,8	1,0

Применение однофакторных регрессионно-корреляционных уравнений для решения задач статистического исследования

Составитель ***Бекетаев О.Б.***

Редактор ***Дмитриенко К.М.***
Тех. редактор ***Бейшеналиева А.И.***

Подписано к печати 08.12.2010 г. Формат бумаги 60x84 $^{1}/_{16}$.
Бумага офс. Печать офс. Объем 3,75 п.л. Тираж 50 экз. Заказ 297 Цена 24 с.

Бишкек, ул. Сухомлинова, 20. ИЦ "Текник" КГТУ им. И.Раззакова, т.: 54-29-43
e-mail: beknur@mail.ru