ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «ОРЕНБУРГСКИЙ ГОСУДАРСТВЕННЫЙ АГРАРНЫЙ УНИВЕРСИТЕТ»

Методические рекомендации для самостоятельной работы обучающихся по дисциплине

Б1.В.ДВ.03.01 Теория вероятностей и математическая статистика

Направление подготовки 10.03.01 Информационная безопасность

Профиль подготовки Безопасность автоматизированных систем

Квалификация выпускника бакалавр

Форма обучения очная

СОДЕРЖАНИЕ

1. ОРГАНИЗАЦИЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ

1.1. Организационно-методические данные дисциплины

№ п.п.	Наименование темы	Общий объем часов по видам самостоятельной работы				
		подготовка курсового проекта (работы)	подготовка реферата/эссе	индивидуальные домашние задания (ИДЗ)	самостоятельное изучение вопросов (СИВ)	подготовка к занятиям (ПкЗ)
1	2	3	4	5	6	7
1	Классическое определение вероятности события. Геометрические вероятности. Относительная частота наступления события и статистическая вероятность. Формулы умножения и сложения вероятностей случайных событий	-	-	-	-	2
2	Зависимые события. Условная вероятность. Формула полной вероятности события. Вероятности гипотез. Формула Байеса. Повторение испытаний: формулы Бернулли, локальные и интегральные теоремы Лапласа, формула Пуассона, простейший поток событий.	-	-	-	-	2
3	Понятие случайной величины примеры. Виды случайных величин. Закон распределения вероятностей. Функция распределения случайных величин. Свойства. Плотность распределения вероятностей. Числовые характеристики:	-	-	-	-	2

	математическое ожидание, свойства; дисперсия, свойства; среднее квадратичное отклонение и его свойства.					
4	Законы распределения ДСВ: биноминальный и Пуассона. Законы распределения вероятностей НСВ: равномерное распределение, показательное распределение. Нормальное распределение вероятностей НСВ. Правило трех сигм.	-	-	-	-	6
5	Многомерные случайные величины, их числовые характеристики	-	-	-	2	4
6	Задачи математической статистики. Статистический материал. Статистические параметры распределения. Статистические оценки параметров распределения	-	-	-	-	-
7	Интервальные оценки параметров статистического распределения. Необходимость их введения. Доверительные интервалы. Доверительные вероятности. Доверительные интервалы для оценки математического ожидания нормального распределения. Доверительные интервалы для оценки среднего квадратического отклонения нормального распределения.	-	-	-	-	2
8	Понятие статистической гипотезы. Виды гипотез.	-	-	-	-	4

	Статистический критерий. Критическая область. Мощность критерия. Критерии согласия: критерий Пирсона. Выравнивание рядов.					
9	Понятие функциональной, стохастической и корреляционной зависимости. Функция регрессии. Корреляционное отношение. Его свойства, значимость. Линейная функция регрессии. Коэффициент корреляции его.	-	-	-	8	6
10	Основные понятия теории марковских процессов. Простейший поток. Классификация марковских процессов	-	-	-	-	2
11	СМО, их свойства, классификация	-	-	-	-	2
Итого в соответствии с РПД		-	-	-	**10**	**32**

2. МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ ПО САМОСТОЯТЕЛЬНОМУ ИЗУЧЕНИЮ ВОПРОСОВ

5.2.1Многомерные случайные величины, их числовые характеристики

Нормальный закон распределения двумерной случайной величины

Непрерывная случайная величина (X,Y) имеет двумерное нормальное распределение, если ее плотностью распределения равна

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r_{xy}^2}}\exp\left\{-\frac{1}{2(1-r_{xy}^2)}\left[\frac{(x-m_x)^2}{\sigma_x^2}-2r_{xy}\frac{(x-m_x)(y-m_y)}{\sigma_x\sigma_y}+\frac{(y-m_y)^2}{\sigma_y^2}\right]\right\},$$

где $m_x=M[X]$, $m_y=M[Y]$, $s_x=\sqrt{D[X]}$,$s_y=\sqrt{D[Y]}$, $r_{xy}=\frac{k_{xy}}{\sigma_x\sigma_y}$ -коэффициент корреляции, $k_{xy}=M[(x-m_x)(y-m_y)]$- ковариация.
Отсюда следует, что

$$f_1(x)=\int_{-\infty}^{+\infty} f(x,y)dy=\frac{1}{\sigma_x\sqrt{2\pi}}\exp\left\{-\frac{(x-m_x)^2}{2\sigma_x^2}\right\},$$

$$f_2(y)=\int_{-\infty}^{+\infty} f(x,y)dx=\frac{1}{\sigma_y\sqrt{2\pi}}\exp\left\{-\frac{(y-m_y)^2}{2\sigma_y^2}\right\},$$

т.е. если (X,Y) распределена нормально, то и каждая ее составляющая случайная величина распределена нормально.

Если коэффициент корреляции $r_{xy}=0$, т.е. величины X и Y некоррелированны, то легко получить, что $f(x,y)=f_1(x)\times f_2(y)$, т.е. для двумерного нормального распределения понятия некоррелированности и независимости равносильны.

Если X и Y - независимые случайные величины, имеющие нормальные распределения, то

$$P(a<X<b;g<Y<d)=P(a<X<b)\times P(g<Y<d)=$$

$$=\{\Phi(\frac{\beta-m_x}{\sigma_x})-\Phi(\frac{\alpha-m_x}{\sigma_x})\}\{\Phi(\frac{\delta-m_y}{\sigma_y})-\Phi(\frac{\gamma-m_y}{\sigma_y})\}$$

С геометрической точки зрения график плотности представляет собой “гору” с достаточно крутыми склонами, вершина которой находится в точке (m_x,m_y). Линиями уровня служат эллипсы

$$\left\{\left(\frac{x-m_x}{\sigma_x}\right)^2-2r_{xy}\left(\frac{x-m_x}{\sigma_x}\right)\left(\frac{y-m_y}{\sigma_y}\right)+\left(\frac{y-m_y}{\sigma_y}\right)^2\right\}=C=const$$

с центром в т. (m_x,m_y), большая полуось которых имеет при $r_{xy}>0$ положительный наклон к оси абсцисс 0x. По мере удаления от центра плотность нормального распределения очень быстро убывает и стремится к нулю.

На рис.1 представлены линии уровня- эллипсы, ограничивающие область, в которые случайный вектор попадает с вероятностями 0,5; 0,9; 0,99. Говорят, что X и Y связаны линейной корреляционной зависимостью, если обе функции регрессии Y на X и X на Y линейны.

Имеет место следующая важная теорема.

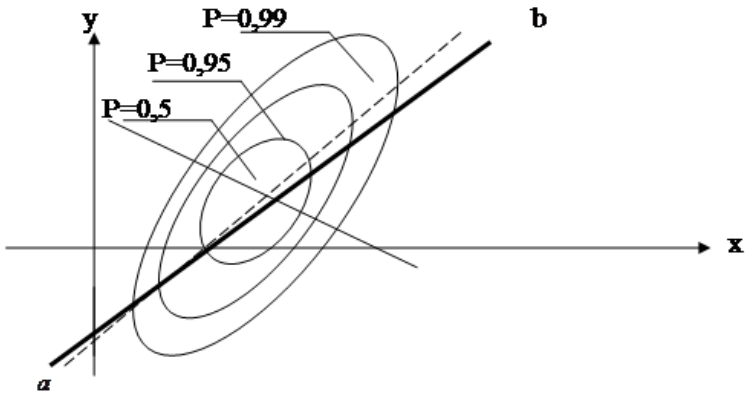

Рис.1

Теорема *(без доказательства)***. Если двумерная случайная величина (X,Y) распределена нормально, то X и Y связаны линейной корреляционной зависимостью**.

Например, если (X,Y) распределена нормально, причем $m_x=3$, $m_y=1$, $s_x=\sqrt{5}$, $s_y=1$, $k_{xy}=2$, то $r_{xy}=\frac{2}{\sqrt{5}}$ и

$$f(x,y)=\frac{1}{2\pi\sqrt{5}\cdot 1\sqrt{1-\frac{4}{5}}}\exp\left\{-\frac{1}{2(1-\frac{4}{5})}\left(\frac{(x-3)^2}{5}-2\frac{2}{\sqrt{5}}\frac{(x-3)(y-1)}{\sqrt{5}\cdot 1}+\frac{(y-1)^2}{1}\right)\right\}=$$

$$=\frac{1}{2\pi}\exp\left\{-\frac{1}{2}[(x-3)^2-4(x-3)(y-1)+5(y-1)^2]\right\},$$

Прямая среднеквадратической регрессии Y на X

y=kx+b, где k= $\frac{\sigma_y}{\sigma_x}r_{xy}$, b= m_y - k×m_x имеет вид y= $\frac{2}{5}x-\frac{1}{5}$.

Пример. Случайные величины X и Y независимы и нормально распределены с $m_X=m_Y=0; D(X)=D(Y)=1.$ Найти вероятность того, что случайная точка (X,Y) попадет в кольцо $k=\{(x,y):4\le x^2+y^2\le 9\}$

Решение: Так как случайные величины X и Y независимы, то они не коррелированы и, следовательно, r = 0. Подставляя $m_x=m_y=0, \sigma_x=\sigma_y=1, r=0$ в (C), получаем $x^2+y^2=C^2$,

то есть эллипс равной вероятности выродился в круг равной вероятности. Тогда

$$P\{(X,Y)\in k\}=P(3)-P(2)=\left(1-\exp\left(-\frac{9}{2}\right)\right)-\left(1-\exp\left(-\frac{4}{2}\right)\right)=e^{-2}-e^{-4,5}\approx 0.1242.$$

Ответ:0,1242.

5.2.3 Понятие функциональной, сто-хастической и корреляционной зависимости. Функция регрес-сии. Корреляционное отноше-ние. Его свойства, значимость. Линейная функция регрессии. Коэффициент корреляции его.

Нелинейные регрессионные модели. Автокорреляция

Многие важные связи в экономике являются **нелинейными**, например, ПФ (зависимости между объемом производства, трудом и капиталом и т.д.), функция спроса (зависимости между спросом на какой – либо товар или услуги, доходом населения и ценами на этот товар). Если в результате анализа пришли к выводу, что в регрессионной модели функция $f(\overline{X},\overline{A})$ *нелинейная*, то обычно поступают так:

• подбирают такие *преобразования* анализируемых переменных $y, x_1, x_2, \ldots, x_n$, которые позволили бы представить искомую зависимость в виде **линейного** соотношения между ***новыми переменными***:

$\tilde{y}=\varphi_0(y)$, $\tilde{x}_1=\varphi_1(x_1)$, …, $\tilde{x}_n=\varphi_n(x_n)$ – преобразования, $\tilde{y}=a_0+a_1\tilde{x}_1+a_2\tilde{x}_2+\cdots+a_n\tilde{x}_n+e_i$, где $i=1,\ldots,k$

– это процедура **линеаризации модели**.

• при невозможности линеаризации модели исследуют регрессионную зависимость (нелинейную). Это значительно сложнее.

Различают **два класса** нелинейных **регрессий**:

• регрессии, нелинейные *относительно* включенных в анализ *объясняющих переменных*(факторов), но ***линейные по оцениваемым параметрам***;

• регрессии, ***нелинейные по оцениваемым параметрам***.

К **первому классу** относятся следующие функции:

• полиномы разных степеней – $y = a + bx + cx^2 + \varepsilon$,
$y = a + bx + cx^2 + dx^3 + \varepsilon$, и т.д.

• гипербола – $y = a + \frac{b}{x} + \varepsilon$.

Ко второму классу относятся функции:

• степенная – $y = a \cdot x^b \cdot \varepsilon$;

• показательная – $y = a \cdot b^x \cdot \varepsilon$;

• экспоненциальная – $y = e^{a+bx} \cdot \varepsilon$.

Нелинейная регрессия по включенным переменным представляет сложности в оценке параметров. Она определяется методом наименьших квадратов, ибо эти функции линейны по параметрам.

Пример 1

В уравнении полинома второй степени $y = a_0 + a_1 x + a_2 x^2 + \varepsilon$, сделаем замену $x = x_1$, $x^2 = x_2$, тогда имеем $y = a_0 + a_1 x_1 + a_2 x_2 + \varepsilon$ – уравнение линейной регрессии.

В уравнении полинома $y = a_0 + a_1 x + a_2 x^2 + \cdots + a_p x^p + \varepsilon$,

сделаем замену ,…, $x^p = x_p$,

получим уравнение вида $y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_p x_p + \varepsilon$.

Значит, *полин ом*любого порядка сводится к *линейной регрессии* с ее методами оценивания параметров и *проверки гипотез*. На практике чаще используется *парабола*, то есть полином второй степени. В этом случае определяется ***x***, при котором достигается максимальное (или минимальное) значение результативного признака. Для этого применяют необходимое условие экстремума функции $\hat{y}_x = a + bx + cx^2$, то есть находят производную и приравнивают ее к нулю $\hat{y}'_x = b + 2cx = 0$, откуда $x_0 = -\frac{b}{2c}$. Эту зависимость целесообразно применять, если в определенном интервале значений фактора меняется характер связи рассматриваемых признаков: прямая связь меняется на обратную или наоборот.

Применение МНК для оценки параметров параболы второй степени приводит к следующей системе нормальных уравнений:

$$\begin{cases} \sum_{i=1}^{k} y_i = k \cdot a + b \cdot \sum_{i=1}^{k} x_i + c \cdot \sum_{i=1}^{k} x_i^2, \\ \sum_{i=1}^{k} y_i \cdot x_i = a \cdot \sum_{i=1}^{k} x_i + b \cdot \sum_{i=1}^{k} x_i^2 + c \cdot \sum_{i=1}^{k} x_i^3, \\ \sum_{i=1}^{k} y_i \cdot x_i^2 = a \cdot \sum_{i=1}^{k} x_i^2 + b \cdot \sum_{i=1}^{k} x_i^3 + c \cdot \sum_{i=1}^{k} x_i^4. \end{cases}$$

Решение ее возможно по правилу Крамера или Гаусса.

Такого рода функцию можно наблюдать в *экономике труда* при изучении зависимости *заработной платы от возраста*. С определенного возраста заработная плата может убывать ввиду старения организма, снижения производительности труда.

Такой же характер носит кривая Лаффера (зависимость поступления в бюджет от налоговой ставки).

Такая зависимость может быть использована для характеристики зависимости урожайности от количества внесенных удобрений. С увеличением количества удобрений урожайность растет лишь до достижения оптимальной дозы вносимых удобрений. Дальнейший рост удобрений оказывается вредным для растения и урожайность снижается.

Чаще исследователь имеет дело лишь с отдельными сегментами параболы.

Следует отметить также, что параметры параболы не всегда могут быть логически истолкованы.

Пример 2

Рассмотрим пример гиперболической зависимости. Она может быть использована для характеристики связи между удельными расходами сырья, материалов, топлива и объемов выпускаемой продукции, временем обращения товаров и товарооборотом (на микроуровне), а также макроуровне. Например, зависимость себестоимости (результирующий признак ***y***) от объемов выпускаемой продукции (фактор ***x***).

Классическим примером гиперболической зависимости является кривая Филлипса, характеризующая нелинейное соотношение между нормой безработицы (фактор ***x***) и процентом прироста заработной платы (результирующий признак***y***).

Заменив $z = \frac{1}{x}$, получим линейное уравнение регрессии $y = a + bz + \varepsilon$, оценки параметров которого могут быть найдены МНК. Система нормальных уравнений будет иметь следующий вид:

$$\begin{cases} \sum_{i=1}^{k} y_i = k \cdot a + b \cdot \sum_{i=1}^{k} \frac{1}{x_i}, \\ \sum_{i=1}^{k} \frac{y_i}{x_i} = a \cdot \sum_{i=1}^{k} \frac{1}{x_i} + b \cdot \sum_{i=1}^{k} \frac{1}{x_i^2}. \end{cases}$$

или после замены $\begin{cases} \sum_{i=1}^{k} y_i = k \cdot a + b \cdot \sum_{i=1}^{k} z_i, \\ \sum_{i=1}^{k} y_i \cdot z_i = a \cdot \sum_{i=1}^{k} z_i + b \cdot \sum_{i=1}^{k} z_i^2. \end{cases}$

Если получить здесь матрицу X, то она будет иметь следующий вид:

$$X = \begin{pmatrix} 1 & \frac{1}{x_1} \\ \vdots & \vdots \\ 1 & \frac{1}{x_k} \end{pmatrix}_{k \times 2}$$ (матрица имеет *k*строк и 2 столбца).

Так, для кривой Филипса $\hat{y} = 0{,}00679 + 0{,}1842 \frac{1}{x}$ (здесь ***a***=0,00679 и ***b***=0,1842) величина ***a*** означает, что с ростом уровня безработицы *темп прироста заработной платы* стремится к нулю. Соответственно можно определить тот уровень безработицы, при котором заработная плата оказывается стабильной и темп ее прироста равен нулю.

При $b < 0$, имеем медленно *повышающуюся* функцию с верхней асимптотой $y = a$ при $\chi \to \infty$.

Примером может служить взаимосвязь доли расходов ***y*** на товары длительного пользования и общих сумм расходов (или доходов)***x***. Немецкий статистик Энгель на основании исследования семейных расходов сформулировал закономерность – с ростом дохода доля расходов, расходуемых на продовольствие, уменьшается. Соответственно с увеличением доходов доля дохода, расходуемая на непродовольственные товары, будет возрастать. Имеем гиперболическую зависимость вида $y = a - \frac{b}{x} + \varepsilon$. При $y = 0$, получаем $a = \frac{b}{x}$, откуда $x = \frac{b}{a}$.

Вместе с тем равносторонняя гипербола не является единственно возможной функцией для описания кривой Энгеля. **Уоркинг**и **С.Лизер** для этих целей использовал и **полулогарифмическую** зависимость

$$y = a + b \cdot \ln x + \varepsilon.$$

Заменив $z = \ln x$, опять получим линейное уравнение $\hat{y} = a + b \cdot z$. Оценка параметров ***a*** и ***b*** может быть найдена МНК. Система нормальных уравнений имеет вид:

$$\begin{cases} \sum_{i=1}^{k} y_i = k \cdot a + b \cdot \sum_{i=1}^{k} \ln x_i, \\ \sum_{i=1}^{k} y_i \cdot \ln x_i = a \cdot \sum_{i=1}^{k} \ln x_i + b \cdot \sum_{i=1}^{k} (\ln x_i)^2. \end{cases}$$

Возможны и иные нелинейные модели. Однако если нет каких-либо теоретических обоснований в использовании какого-либо вида кривых, то нужно использовать такие, чтобы для преобразованных переменных получить более простую модель регрессии.

До сих пор преобразования затрагивали только факторы. Рассмотрим случай, когда преобразовывается не только фактор, но и результирующий признак.

Если имеется зависимость вида $y = \frac{1}{a + bx + \varepsilon}$, то есть $\frac{1}{y} = a + bx + \varepsilon$, то она приводится к линейной регрессии преобразованием $\tilde{y} = \frac{1}{y}$, получаем $\tilde{y} = a + bx + \varepsilon$. При вычислении МНК оценок в качестве вектора наблюдаемых значений надо использовать вектор $\vec{\tilde{y}} = \left(\frac{1}{y_1}, \ldots, \frac{1}{y_k} \right)$. Эта ***зависимость*** полезна при изучении ***спроса на товар y*** в зависимости от его ***цены x.***

При зависимости $y = \frac{x}{a + bx + x\varepsilon}$ линеаризация достигается преобразованием: $\tilde{x} = \frac{1}{x}$, тогда $\tilde{y} = a\tilde{x} + b + \varepsilon$.

Экспоненциальная (показательная) зависимость.

Широкий класс экономических показателей характеризуется приблизительно ***постоянным темпом*** относительно прироста во времени. Этому соответствует зависимость вида: $y = a \cdot e^{bx + \varepsilon}$.

Относительный прирост ***y***за единицу времени ***x***:

$$\frac{1}{y}\frac{dy}{dx} = \frac{1}{y} \cdot a \cdot e^{bx + \varepsilon} = b.$$

Линеаризация достигается здесь переходом к переменным $\tilde{y} = \ln y$, тогда $\tilde{y} = \tilde{a} + bx + \varepsilon$, где $\tilde{a} = \ln a$.

Имея наблюдения $(x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k)$ и формируя столбец $\vec{\tilde{y}} = (\ln y_1, \ln y_2, \ldots, \ln y_k)^T$, с помощью МНК получают оценки $\tilde{a}$, b, а затем $a = e^{\tilde{a}}$.

Приводима к линейному виду и **логистическая функция**

$$y = \frac{a}{1 + b \cdot e^{-cx + \varepsilon}}.$$

Ее можно записать в виде: $b \cdot e^{-cx + \varepsilon} = \frac{a}{y} - 1$. Тогда логарифмируя данное выражение $\ln b - cx + \varepsilon = \ln\left(\frac{a}{y} - 1 \right)$ и полагая $z = \ln\left(\frac{a}{y} - 1 \right)$ и $B = \ln b$, получим линейную зависимость $z = B - cx + \varepsilon$.

Логистическая кривая используется для описания поведения показателей, имеющих определенные «***уровни насыщения***».

Степенной вид.

В экономических исследованиях распространены зависимости степенного вида: $y = a_0 \cdot x_1^{a_1} \cdot x_2^{a_2} \cdot \dots \cdot x_k^{a_k} \cdot \varepsilon$.

При преобразовании , $\tilde{x}_i = \ln x_i$, где $i = 1,\dots,k$.

Получаем: $\tilde{y} = \tilde{a}_0 + a_1\tilde{x}_1 + a_2\tilde{x}_2 + \dots + a_k\tilde{x}_k + \ln\varepsilon$, где $\tilde{a}_0 = \ln a_0$. Зависимости степенного вида играют важную роль при построении и анализе ПФ. В этом случае коэффициенты $a_1, a_2, \dots, a_k$ являются эластичностями признака у по объясняющим переменным $x_1, x_2, \dots, x_k$.

Действительно, коэффициент эластичности определяется по формуле $E_i = \frac{\partial y}{\partial x_i} \cdot \frac{x_i}{y}$. В нашем случае

$$\frac{\partial y}{\partial x_i} = \frac{\partial}{\partial x_i}\left(a_0 \cdot x_1^{a_1} \cdot x_2^{a_2} \cdot \dots \cdot x_i^{a_i} \cdot \dots \cdot x_k^{a_k}\right) = a_0 \cdot a_i \cdot x_1^{a_1} \cdot x_2^{a_2} \cdot \dots \cdot x_i^{a_i-1} \cdot \dots \cdot x_k^{a_k}$$

$$= \frac{a_0 \cdot a_i \cdot x_1^{a_1} \cdot x_2^{a_2} \cdot \dots \cdot x_i^{a_i} \cdot \dots \cdot x_k^{a_k}}{x_i} = \frac{a_i}{x_i} \cdot y$$, тогда

$$E_i = \frac{\partial y}{\partial x_i} \cdot \frac{x_i}{y} = \frac{a_i}{x_i} \cdot y \cdot \frac{x_i}{y} = a_i$$.

Эластичность показывает, на сколько процентов возрастает у, если затраты *i*-го ресурса x_i увеличить на 1%.

Вообще, нелинейные модели регрессии, нелинейные по оцениваемым параметрам подразделяются на два типа: внутренне линейные и внутренне нелинейные.

В ***первом*** случае модель с помощью соответствующих преобразований может быть приведена к линейному виду. Например, степенная функция . Данная модель нелинейная по оцениваемым параметрам, но может быть сведена к линейной путем логарифмирования, то есть $\ln y = \ln a + b \cdot \ln x + \ln\varepsilon$, введем замену $\tilde{y} = \ln y$, $\tilde{a} = \ln a$, $\tilde{x} = \ln x$, $\tilde{\varepsilon} = \ln\varepsilon$, получаем уравнение линейной регрессии $\tilde{y} = \tilde{a} + b\tilde{x} + \tilde{\varepsilon}$.

Во ***втором*** случае модель не может быть сведена к линейной функции. Например, модель вида $y = a + b \cdot x^c + \varepsilon$ или $y = a \cdot \left(1 - \frac{1}{1 - x^b}\right) + \varepsilon$.

В этом случае для *оценки параметров* используются итеративные процедуры, успешность которых зависит от вида уравнений и особенностей применяемых итеративных подходов.

Замечание. В моделях нелинейных по оцениваемым параметрам, но приводимых к линейному виду, МНК применяется к преобразованным уравнениям. Если в линейной модели и в моделях, нелинейных по переменным, при оценке параметров исходят из критерия $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \to \min$, то в моделях, нелинейных по оцениваемым параметрам, требование МНК применяется не к исходным значениям результативного признака, а к их преобразованным величинам, то есть $\ln y$, $\frac{1}{y}$ и другим. Так в степенной функции оценка пара-

метров основывается фактически на минимизации суммы квадратов отклонений в логарифмах:

$$\sum_{i=1}^{n}(\ln y_i - \ln \hat{y}_i)^2 \to \min$$

.

Вследствие этого оценка параметров для линеаризуемых функций МНК оказывается несколько смещенной (заниженной).

При использовании линеаризуемых функций, затрагивающих преобразования зависимой переменной y, следует проверять наличие предпосылок МНК, чтобы они не нарушались при преобразовании.

Преобразование случайного члена.

Как было отмечено выше, для получения качественных оценок существенную роль играет выполнимость определенных предпосылок МНК для случайного отклонения (нормальное распределение в том числе).

В случаях, не требующих совокупного логарифмирования с аддитивным случайным членом, выполнимость предпосылок имеет место и проблем с оцениванием не возникает.

Для иллюстрации возможных проблем со случайным членом рассмотрим функцию $y = a \cdot x^b$, по-разному дополнив ее случайным членом.

1. $y = a \cdot x^b \cdot e^{\varepsilon}$, прологарифмировав данное выражение, получим: $\ln y = \ln a + b \cdot \ln x + \varepsilon$. Использование данного соотношения для оценки параметров не вызывает осложнений, связанных со случайным отклонением.

2. , прологарифмировав данное выражение, получим: $\ln y = \ln a + b \cdot \ln x + \ln \varepsilon$, то есть процедура линеаризации приводит к преобразованию случайных отклонений ε в $\ln \varepsilon$.

Использование МНК для нахождения оценок параметров требует, чтобы отклонения удовлетворяли предпосылкам МНК, то есть $\ln \varepsilon \sim N(0, \sigma^2)$. Другими словами, вектор возмущений должен иметь *логарифмически нормальное* распределение.

3. Возьмем в качестве модели $y = a \cdot x^b + \varepsilon$. Логарифмирование этого соотношения не приводит к линеаризации $\ln y = \ln(a \cdot x^b + \varepsilon)$.

Таким образом, при использовании преобразований линеаризации необходимо особое внимание уделять рассмотрению *свойств случайных* отклонений, чтобы полученные оценки имели высокую статистическую значимость.

Корреляция для нелинейной регрессии.

Уравнение нелинейной регрессии, так же как и в линейной зависимости, дополняется показателем корреляции: **индексом корреляции**

$$R = \left(1 - \frac{\sigma_{ост}^2}{\sigma_y^2}\right)^{\frac{1}{2}},$$

Где, $\sigma_y^2 = \frac{1}{k}\sum_{i=1}^{n}(y_i - \bar{y})^2$ – общая дисперсия результативного признака,

а $\sigma_{ост}^2 = \frac{1}{k}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ – остаточная дисперсия.

Величина данного показателя $0 \le R \le 1$, чем ближе к *единице*, тем *теснее связь* рассматриваемых признаков, тем *более надежно* найденное уравнение регрессии.

Индекс корреляции *R* можно найти по формуле:

$$R = \sqrt{1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Парабола второй степени, как и полином более высокого порядка, при линеаризации принимает вид уравнения множественной регрессии.

Если же нелинейное относительно объясняющей переменной уравнение регрессии при линеаризации принимает форму линейного уравнения парной регрессии, то для оценки тесноты связи может быть использован линейный коэффициент корреляции, величина которого в этом случае совпадает с индексом корреляции:

$$R = R_{yx} = r_{yz},$$

где z– преобразованная величина признака – фактора.

Иначе обстоит дело, когда преобразование уравнения в линейную форму связаны с зависимой переменной ***y***. В этом случае линейный коэффициент корреляции по преобразованным переменным дает лишь приближенную оценку тесноты связи и численно не совпадает с индексом корреляции.

Так для степенной функции после перехода к логарифмически $\tilde{x} = \ln x$ линейному уравнению $\ln y = \ln a + b \ln x$. Обозначим $z = \ln y$, $\tilde{a} = \ln a$, тогда $z = \tilde{a} + b\tilde{x}$ и $r_{z\tilde{x}} = \eta_{\ln y \ln x}$ может быть найден линейный коэффициент корреляции не для фактических значений переменных ***x*** и ***y***, а для их логарифмов, то есть $\eta_{\ln y \ln x}$.

Как показали расчеты, значения R_{yx} и довольно близки (или и), поэтому и для нелинейных функций используются для характеристики тесноты связи линейные коэффициенты корреляции. Только следует принимать во внимание, что для линейной зависимости $y = a + bx$ и $x = A + By$ и $r_{yx} = r_{xy}$.

Индекс детерминации $R^2 = R^2_{yx}$ можно сравнивать с коэффициентом детерминации r^2_{yx} для обоснования возможности применения *линейной функции*. Чем больше кривизна линии регрессии, тем величина коэффициента детерминации меньше индекса детерминации R^2_{yx}. Близость этих показателей означает, что нет необходимости усложнять форму уравнения регрессии и можно использовать линейное уравнение. Практически, если $\left(R^2_{yx} - r^2_{yx}\right)$ не превышает 0,1, то предположение о *линейной связ и*считается *оправданным*.

В противном случае проводится оценка существенности различия через t– критерий Стьюдента:

$$t = \frac{R^2_{yx} - r^2_{yx}}{m_{|R-r|}},$$

где ошибка $m_{|R-r|} = 2\sqrt{\frac{\left(R^2_{yx} - r^2_{yx}\right) - \left(R^2_{yx} - r^2_{yx}\right)^2 \cdot \left(2 - \left(R^2_{yx} + r^2_{yx}\right)\right)}{k}}$ разности между и .

Если $t_{факт} > t_{табл}$, то различия между рассматриваемыми показателями корреляции существенны и замена нелинейной регрессии уравнением линейной функции невозможна. Практически, если величина $t < 2$, то различия между R_{yx} и r_{yx} несущественны, и возможно применение линейной регрессии, даже если есть предположения о некоторой нелинейности рассматриваемых соотношений признаков фактора и результата.

Пример

Предположим, что найдено уравнение регрессии $\hat{y} = 9{,}876 + 5{,}129 \ln x$.

Была использована линейная функция $\hat{y} = 9{,}28 + 1{,}777x$ и коэффициент корреляции для нее составил 0,97416. Индекс корреляции для нелинейной зависимости *R*=0,99581.

Тогда $R_{yx}^2 - r_{yx}^2 = (0{,}99581)^2 - (0{,}97416)^2 = 0{,}04265$, то есть применение нелинейной функции увеличивает долю объясненной вариации на 4,3%.

$R_{yx}^2 + r_{yx}^2 = (0{,}99581)^2 + (0{,}97416)^2 = 1{,}94063$;

$$m_{|R-r|} = 2\sqrt{\frac{0{,}04265 - (0{,}04265)^2 \cdot (2 - 1{,}94063)}{6}} = 0{,}16841;$$

$$t = \frac{0{,}04265}{0{,}16841} = 0{,}25 < 2.$$

Следовательно, если нет уверенности в правильности выбора полулогарифмической функции, то она может быть заменена линейной функцией.

Средняя ошибка аппроксимации.

Фактические значения результативного признака y_i отличаются от теоретических, рассчитанных по уравнению регрессии $\hat{y}_i$. Величина отклонений $(y_i - \hat{y}_i)$ по каждому наблюдению представляет собой ***ошибку аппроксимации*** .Их число соответствует *объему совокупности*(*k*). В отдельных случаях ошибка аппроксимации может оказаться равной нулю. Для сравнения используются величины отклонений, выраженные в процентах к фактическим значениям.

Отклонения $(y_i - \hat{y}_i)$ можно рассматривать как ***абсолютную ошибку*** аппроксимации, а $\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\%$ –как ***относительную*** ошибку аппроксимации.

Чтобы иметь общее суждение о качестве модели из относительных отклонений по каждому наблюдению, определяют среднюю ошибку аппроксимации как среднюю арифметическую простую:

$$A = \frac{1}{k}\sum_{i=1}^{k}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\%.$$

Ошибка аппроксимации в пределах 5 – 7% свидетельствует о хорошем подборе модели к исходным данным. *Допустимый* предел значений *A*− не более 8− 10% (допускается 8− 15%).

Возможно и иное определение средней ошибки аппроксимации:

$$A = \frac{100\%}{\bar{y}}\sqrt{\frac{\sum_{i=1}^{k}(y_i - \hat{y}_i)^2}{k}}.$$

В стандартных программах чаще используется первая формула.

Сущность и причины автокорреляции в остатках

Автокорреляция в остатках обычно встречается при регрессионном анализе временных рядов, и почти не встречается при анализе пространственных выборок. Чаще встречается положительная автокорреляция. Она в большинстве случаев вызывается направленным постоянным воздействием некоторых неучтенных в модели факторов. При положительной автокорреляции остатки изменяются монотонно с течением времени наблюдения, а при отрицательной – следует частое изменение знака остатка.

Среди основных причин автокорреляции можно выделить следующие:

а) ошибки спецификации – неучет в модели какой-то важной объясняющей переменной или неверный выбор вида функции, что ведет к систематическим отклонениям точек наблюдения от линии регрессии,

б) инерция – запаздывание реакции экономической системы на изменение факторов,

в) сглаживание данных.

Последствия автокорреляции в остатках такие же, как и в случае гетероскедастичности (потеря эффективности, смещение дисперсий оценок параметров, занижение стандартных ошибок и завышение *t*–статистик параметров), а это может повлечь признание незначимых факторов значимыми. Вследствие перечисленных обстоятельств, прогнозные качества модели ухудшаются.

При анализе временных рядов вместо индекса *i* часто будем использовать время *t*, а вместо числа наблюдений *n* будем писать T – продолжительность интервала наблюдения временного ряда.

Мы будем рассматривать автокорреляцию первого порядка, так как в большинстве практических случаев автокорреляционная функция быстро убывает.

Коэффициент автокорреляции 1-го порядка в остатках:

$$r_\varepsilon(1) = \frac{\operatorname{cov}(\varepsilon_{t-1}, \varepsilon_t)}{\sigma(\varepsilon_{t-1}) \cdot \sigma(\varepsilon_t)} = \frac{\overline{\varepsilon_{t-1} \cdot \varepsilon_t} - \overline{\varepsilon_{t-1}} \cdot \overline{\varepsilon_t}}{\sigma(\varepsilon_{t-1}) \cdot \sigma(\varepsilon_t)} \approx \frac{\overline{\varepsilon_{t-1} \cdot \varepsilon_t} - \overline{\varepsilon_{t-1}} \cdot \overline{\varepsilon_t}}{\sigma^2(\varepsilon_t)}.$$

Если этот коэффициент корреляции существенно отличен от 0, то можно говорить о наличии автокорреляции.

Обнаружение автокорреляции в остатках

1. Графический метод – при использовании этого метода строится график: ε_t есть функция от ε_{t-1}. Если в графике прослеживается отчетливая положительная или отрицательная тенденция, то, скорее всего, имеет место соответствующая автокорреляция в остатках.

2. Метод рядов

В моменты времени $t = 1,2,3 \ldots T$, определяются знаки отклонений, например: $(-----)_5 (+++++++)_7 (---)_3 (++++)_4 (-)_1$ – для 20-ти наблюдений.

Рядом называют непрерывную последовательность одинаковых знаков (ряд ограничен скобками, в примере приведено 5 рядов). Количество знаков называют длиной ряда. Если рядов мало по сравнению с числом наблюдений, то вполне вероятна положительная автокорреляция, если рядов много, – то отрицательная.

3. Тест Дарбина-Уотсона (DW). Это – самый популярный

$$DW = \frac{\sum_{t=2}^{T} (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{T} \varepsilon_t^2}$$

тест: — критерий Дарбина – Уотсона.

Установим связь между этим критерием и коэффициентом корреляции:

$$r_\varepsilon(1) = \frac{\sum_{t=2}^{T} \left(\varepsilon_t - \overline{\varepsilon_t}\right) \cdot \left(\varepsilon_{t-1} - \overline{\varepsilon_{t-1}}\right)}{\sqrt{\sum_{t=1}^{T} \left(\varepsilon_t - \overline{\varepsilon_t}\right)^2} \cdot \sqrt{\sum_{t=2}^{T} \left(\varepsilon_{t-1} - \overline{\varepsilon_{t-1}}\right)^2}} \approx \frac{\sum_{t=2}^{T} \left(\varepsilon_t \cdot \varepsilon_{t-1}\right)}{\sum_{t=1}^{T} \varepsilon_t^2};$$

учитывая, что $\overline{\varepsilon_t} = \overline{\varepsilon_{t-1}} = 0$ и $\sum_{t=1}^{T} \varepsilon_t^2 \approx \sum_{t=2}^{T} \varepsilon^2{}_{t-1}$, получим:

$$r_\varepsilon(1) \approx 1 - \frac{DW}{2},$$

$$DW \approx 2(1 - r_\varepsilon(1)),$$

$$-1 \le r_\varepsilon(1) \le 1,$$

$$0 \le DW \le 4.$$

Процедура обнаружения автокорреляции по критерию DW такова:

1. Вычисляется критерий DW, для чего должна быть выполнена регрессия y на x и определены остатки. Затем выдвигается гипотеза H_0 об отсутствии автокорреляции в остатках.

2. По таблице критических значений теста Дарбина–Уотсона для назначенного уровня значимости γ, числа наблюдений n и числа факторов p определяются верхняя du и нижняя dl критические точки $du,\ dl(\gamma,\ n,\ p).$

3. Строятся области: I–от 0 до dl; II–от dl до du; III–от du до 4–du; IV– от 4–ul до 4–dl и V–от 4–dl до 4.

Это поясняется табл. 9.1.

таблица 9.1

I (+) автокорреляция	II неопределенность	III нет автокорреляции	IV неопределенность	V (–) автокорреляция
0 … dl	dl … du	du … (4 – du)	4–du …4– dl	4–dl … 4

При использовании критерия DW следует учитывать следующие ограничения:

а) он применим лишь для модели с ненулевым свободным членом,

б) остатки должны описываться авторегрессионной моделью первого порядка $AR(1):\ \varepsilon_t = \rho\varepsilon_{t-1} + u_t,$

в) временной ряд должен иметь одинаковую периодичность, то есть не должно быть пропусков наблюдений,

г) DW нельзя применять для моделей авторегрессионных относительно объясняемой переменной y_t, так как в этом случае окажется, что регрессор будет коррелировать с остатком.

3. МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ ПО ПОДГОТОВКЕ К ЗАНЯТИЯМ

3.1 Практическое занятие 1 - 4 (ПЗ-1-4) Случайные события, их вероятность. Основные теоремы теории вероятностей. Условная вероятность. Следствия основных теорем теории вероятностей

При подготовке к занятию необходимо обратить внимание на:

- классификацию случайных событий;
- различные подходы к определению вероятности случайного события;
- комбинаторные формулы, методы непосредственного вычисления вероятности случайного события, основные теоремы.
- понятие условной вероятности;
- модели задач на формулу полной вероятности и формулу Байеса;

3.2 Практическое занятие 5-7 (ПЗ-5-7) Схема повторных испытаний. Простейший поток событий

При подготовке к занятию необходимо обратить внимание на:
- определение схемы повторных испытаний;
- формулы Бернулли, Пуассона, Лапласа, условия их применения;
- вычисление наивероятнейшего числа наступлений события в схеме повторных испытаний;
- свойства простейшего потока, формулу для вычисления вероятности события с заданной интенсивностью.

3.3 Практическое занятие 8-10 (ПЗ-8-10): Случайные величины. Функция и плотность распределения СВ. Числовые характеристики случайной величины

При подготовке к занятию необходимо обратить внимание на:
- классификацию СВ, определение закона распределения вероятностей;
- построение функции распределения ДСВ, вычисление плотности распределения НСВ, вероятности попадания СВ в заданный интервал;
- вычисление числовых характеристик СВ.

3.4 Практическое занятие 11-14 (ПЗ-11-14): Некоторые распределения ДСВ. Некоторые распределения НСВ

При подготовке к занятию необходимо обратить внимание на:
- особенности законов распределения ДСВ: биномиального, Пуассона, геометрического, гипергеометрического;
- особенности законов распределения НСВ: равномерного, показательного, нормального;
- алгоритм применения формул для нахождения вероятности попадания в интервал нормально распределенной СВ, вероятности ее отклонения от м.о., ее частоты от вероятности, правило трех сигм.

3.5 Практическое занятие 15-18 (ПЗ-15-18): Случайный вектор. Распределение многомерной СВ. Условные законы распределения. Числовые характеристики случайного вектора.

При подготовке к занятию необходимо обратить внимание на:
- определение закона распределения МСВ, частные законы распределения МСВ;
- зависимость СВ, условные законы распределения МСВ, способы их получения;
- вычисление числовых характеристик МСВ, условные числовые характеристики МСВ, их свойства;
- ковариацию. Формулу ее вычисления.

3.6 Практическое занятие 19 -21 (ПЗ-19-21): Статистическое распределение

При подготовке к занятию необходимо обратить внимание на:
- определение основных понятий статистики
- первичную обработку статистических данных;
- точечные и интервальные оценки параметров распределения;
- метод моментов, метод доверительных интервалов.

3.7 Практическое занятие 22-24 (ПЗ-22-23): Статистические критерии, их виды

При подготовке к занятию необходимо обратить внимание на:
- определение статистического критерия, ошибок первого и второго рода;
- классификацию статистических критериев, их мощность;
- применение критериев согласия;
- методы выравнивания рядов.

3.8 Практическое занятие 25 -26 (ПЗ-25-26): Стохастическая зависимость межу величинами. Показатели стохастической зависимости

При подготовке к занятию необходимо обратить внимание на:
- понятие стохастической зависимости величин;
- функцию регрессии;
- установление корреляционной зависимости между величинами;
- определение, вычисление, коэффициента корреляции;
- признаки корреляционной зависимости, коэффициент детерминации;

3.9 Практическое занятие 27-28 (ПЗ-27-28): Линейная парная регрессия

При подготовке к занятию необходимо обратить внимание на:
- основные термины и формулы, необходимые для построения парной линейной регрессии;
- формулы для вычисления коэффициента корреляции, его интерпретацию;
-выработку навыков по проверки значимости выборочных коэффициентов.

3.10 Практическое занятие 29-30 (ПЗ-29-30): Основные понятия теории марковских процессов. Простейший поток. Классификация марковских процессов

При подготовке к занятию необходимо обратить внимание на:
- основные понятия и теоремы теории марковских процессов, их классификацию;
- методы работы с моделями простейшего потока СС;
- моделирование по схеме дискретных и непрерывных марковских процессов;
- алгоритмы работы по схеме гибели и размножения;
- определение надежностных характеристик системы, путем обсчета соответствующей модели марковских процессов.

3.11 Практическое занятие 31-34 (ПЗ-31-34): Основные понятия теории систем массового обслуживания. СМО с отказами и СМО с ожиданием (очередью). Предельные вероятности состояний. Модели систем массового обслуживания при пуассоновских потоках заявок.

При подготовке к занятию необходимо обратить внимание на:
- основные термины и формулы, необходимые для работы с марковскими цепями, СМО, их классификацию;
- методы моделирования по схеме марковских цепей, применение теории СМО к решению инженерных задач;
- выработку навыков по составлению уравнений Колмогорова по размеченному графу непрерывной марковской цепи;
- основные методы моделирования СМО при пуассоновском потоке заявок.