Повторить лекцию, параграф 1-6, страницы 206-218 учебник В.Е. Гмурман «Руководство к решению задач по теории вероятности и математической статистики»

**ВЫПОЛНИТЬ: кратко законспектировать лекцию, решить задачу**

1. С целью анализа взаимного влияния зарплаты и текучести рабочей силы на пяти однотипных фирмах с одинаковым числом работников проведены измерения уровня месячной зарплаты Х и числа уволившихся за год рабочих Y

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| X | 100 | 150 | 200 | 250 | 300 |
| Y | 60 | 35 | 20 | 20 | 15 |

Найти линейную регрессию Y на X, выборочный коэффициент корреляции.

1. На основании 18 наблюдений установлено, что на 64% вес ( ) X кондитерских изделий зависит от их объема ( ) Y . Можно ли на уровне значимости α = 0,05 утверждать, что между X и Y существует зависимость?

ВЫПОЛНЕННОЕ ДОМАШНЕЕ ЗАДАНИЕОТСЫЛАТЬ НА МОЮ ПОЧТУ

В ВИДЕ ФОТОГРАФИИ:

[**furkalo25@yandex.ua**](mailto:furkalo25@yandex.ua)

**СРОКИ ВЫПОЛНЕНИЕ ЗАДАНИЕ 23.10.2020 ДО 10:00**

|  |
| --- |
| **Практическое занятие №7**  **по теме: «Уравнения регрессии»** |
| Условимся обозначить через *Х* независимую переменную, а через *Y* зависимую переменную.      В экономике в большинстве случаев между переменными величинами существуют зависимости, когда каждому значению одной переменной соответствует не какое-то определенное, а множество значений другой переменной, причем сказать заранее, какое именно значение примет зависимая величина *Y* , нельзя. Такая зависимость получила название **статистической (или стохастической, вероятностной).** Более часто появление такой зависимости объясняется действием на результирующую переменную не только контролируемого или контролируемых факторов (в данном случае таким контролируемым фактором является переменная *Х*), а и многочисленных неконтролируемых случайных факторов. Примером статистической связи является зависимость урожайности от количества внесенных удобрений, стоимость одного экземпляра книги от тиража, выработки рабочего за смену от его квалификации и т.д.      Допустим, что существует стохастическая зависимость случайной переменной *Y* от *Х*. Зафиксируем некоторое значение *х* переменной *Х*. При *Х=х* переменная *Y* в силу ее стохастической зависимости от *Х* может принять любое значение из некоторого множества, причем какое именно – заранее не известно. Поэтому, прежде всего, стараются выяснить, изменяются или нет при изменении *х* условные математические ожидания *М*(*Y*/*Х=х*). Если при изменении *х*условные математические ожидания *М*(*Y*/*Х=х*) изменяются, то говорят, что имеет место **корреляционная зависимость** величины *Y* от *Х*.      Функция φ(*х*)=*М*(*Y*/*Х=х*), описывающая изменение условного математического ожидания случайной переменной *Y*при изменении значений *х* переменной *Х*, называется **функцией регрессии***,*а ее график – **линией регрессии***.*  Для отыскания функции регрессии, вообще говоря, необходимо знать закон распределения случайной двумерной величины (*Х*,*Y*). В нашем распоряжении лишь выборка ограниченного объема. Поэтому в этом случае речь может идти об оценке (приближенном выражении) функции.  В качестве оценок условных математических ожиданий принимают условные средние, которые находят по данным наблюдений (по выборке).      **Условным средним***`ухт*называют среднее арифметическое наблюдавшихся значений *Y*, соответствующих *Х=х*.  Условное математическое ожидание *М(Y*/*х*) является функцией от *х*, следовательно, его оценка, т.е. условное среднее `*ух*, также функция от *х*; обозначив эту функцию через φ\*(*х*), получим уравнение  `*ух* = φ\*(*х*).     Это уравнение называют **выборочным уравнением регрессии**; функцию φ\*(*х*) называют выборочной регрессией, а ее график – **выборочной линией регрессии.**      Как найти по данным наблюдений параметры функции φ\*(*х*), если вид ее известен? Как оценить силу (тесноту) связи между величинами *Х* и *Y* и установить, коррелированы ли эти величины? Ответы на эти вопросы изложены ниже.  **4.1. Линейная парная регрессия**       Пусть функция регрессии линейная, т.е. М(*Y*/*Х=х*)=α+β*х*. Найдем оценки *а* и *b* параметров α и β.      Предположим, что в результате *n* независимых опытов получены *n* пар чисел (*х*1,*у*1), (*х*2,*у*2),…, (*х* *n*, *yn*). Рассмотрим случай, когда различные значения *х* признака *Х* и соответствующие им значения у признака *Y* наблюдались по одному разу. Тогда выборочное уравнение можно записать так: .      Для нахождения оценок *а* и *b* применим метод наименьших квадратов. Суть этого метода в том, что отыскиваются такие значения *а* и *b*, которые обеспечивают минимум суммы квадратов отклонений измеренных значений *уi* от прямой линии, задаваемой параметрами *а* и *b*, т.е.       Для отыскания минимума приравняем нулю соответствующие частные производные:            .      Выполнив элементарные преобразования, получим систему двух линейных уравнений относительно *а* и *b*:                                     ( 4.1)       Решения этой системы уравнений можно записать в следующем, удобном для расчетов виде:                                     ( 4.2)       Обычно *b* называют коэффициентом регрессии. Коэффициент регрессии показывает, на сколько единиц в среднем изменяется переменная *Y* при увеличении переменной *Х* на одну единицу.       **Пример 4.1**. Найти выборочное уравнение прямой линии регрессии по данным *n*=8 наблюдений, которые получены при изучении зависимости количества продаж товара *у* от затрат на рекламу этого товара *х*:   |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | *х* | 1,5 | 4,0 | 5,0 | 7,0 | 8,5 | 10,0 | 11,0 | 12,5 | | *y* | 5,0 | 4,5 | 7,0 | 6,5 | 9,5 | 9,0 | 11,0 | 9,0 |        **Решение**. Экспериментальные данные изобразим в виде точек в системе декартовых координат. Ломаная линия, соединяющая эти точки, называется эмпирической линией регрессии. По виду ломанной можно предположить наличие корреляционной зависимости *Y* по *Х* между двумя рассматриваемыми переменными, которая графически выражается тем точнее, чем больше объем выборки (рис.4.1).       Рис. 4.1       Составим расчетную таблицу 4.1.    Таблица 4.1     |  |  |  |  |  | | --- | --- | --- | --- | --- | | № | *хi* | *yi* | *x* *i*2 | *xiyi* | | 1 2 3 4 5 6 7 8 | 1,5 4,0 5,0 7,0 8,5 10,0 11,0 12,5 | 5,0 4,5 7,0 6,5 9,5 9,0 11,0 9,0 | 2,25 16,00 25,00 49,00 72,25 100,00 121,00 156,25 | 7,50 18,00 35,00 45,50 80,75 90,00 121,00 112,50 | | Σ | 59,5 | 61,5 | 541,75 | 510,25 |        `*х* =7,4375, `*у* =7,6875      Найдем искомые параметры, для чего подставим вычисленные по таблице суммы в соотношения (4.2):      *а* = (61,5 × 541,75 – 510,25 ×59,50)/ (8 ×541,75 – 3540,25) = 3,73,      *b* = (8 × 510,25 – 59,50 × 61,50)/ (8 ×541,75 – 3540,25) = 0,53.      Таким образом, уравнение регрессии имеет вид      .      Прямая, построенная по этому уравнению, показана на рис. 4.2 вместе с исходными данными. Эта прямая является наилучшей линейной оценкой уравнения регрессии, полученной по имеющимся данным. Но это не означает, что нельзя построить оценку регрессии в виде какой-то другой зависимости (нелинейной), которая будет лучше соответствовать экспериментальным данным, чем прямая линия.            Рис. 4.2      Построенная таким образом линия регрессии позволяет с некоторой вероятностью не только предсказать в интервале от *х*=1,5 до *х*=12,5 любые значения функции *у* при отсутствующих в табл. 4.1 значениях фактора *х*, но и за пределами данного интервала.      Составленное уравнение регрессии можно проверить на точность зависимости между переменными (*х, у*) по коэффициенту точности выравнивания линии*r*1, отражающему степень приближения расчетных данных к фактическим значениям эмпирического ряда. Этот коэффициент определяется следующим образом:      ,                        ( 4.3)      где  – отклонение индивидуальных вариант от общего среднего арифметического по *y*;  – отклонение индивидуальных экспериментальных вариант по *y* от расчетных по уравнению.      Составим таблицу расчета данных для определения коэффициента точности выравнивания линии.     Таблица 4.2   |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | № | *xi* | *yi* |  |  |  |  |  | | 1 2 3 4 5 6 7 8 | 1,5 1,0 5,0 7,0 8,5 10,0 11,0 12,5 | 5,0 4,5 7,0 6,5 9,5 9,0 11,0 9,0 | 4,53 5,85 6,38 7,44 8,24 9,03 9,56 10,35 | –2,6875 –3,1875 –0,6875 –1,1875 1,8125 1,3125 3,3125 1,3125 | 7,2227 10,160 0,4727 1,4102 3,2852 1,7227 10,9727 1,7227 | 0,47 –1,35 0,62 –0,94 1,26 –0,03 1,44 –1,35 | 0,2209 1,8225 0,3844 0,8836 1,5876 0,0009 2,0736 1,8225 | | Σ |  |  |  |  | 36,9691 |  | 8,7956 |        `*у* = 7,6875      На основании исходных данных, полученных в табл. 4.2, используя формулу (4.3), имеем             Принято считать: если *r*1>0,95, то уравнение регрессии адекватно отражает существующую связь. При *r*1<0,95 необходимо найти другую математическую зависимость между признаками. В приведенном примере *r*1=0,87<0,95, поэтому следует подобрать другую математическую зависимость. Критерий оценки *r*1на точность выравнивания линии уравнения регрессии используется и для других форм регрессионной зависимости.      Проверку адекватности линейной модели можно провести по графику остатков:      ,      где *уi* – измеренные значения, соответствующие значениям *xi*; *ỹi* – значения функции регрессии при *х*=*хi*  Если остатки *di*сконцентрированы в горизонтальной полосе вдоль оси абсцисс, то линейную модель можно считать адекватной. Если зона, где расположены остатки, расширяется, это означает, что дисперсии неодинаковы при различных значениях *хi*. Это требует изменения регрессионной модели. Если остатки имеют тенденцию закономерно изменяться, то не учтены какие-то факторы, существенно влияющие на связь между величинами *Y* и *х*. В этом случае также нужно изменить модель и ввести неучтенные факторы.      В заключение построим график остатков для предыдущего примера. Для этого используем столбцы *уi*и *yi*–*ỹi*табл.4.2. Этот график приведен на рис.4.3.            Рис. 4.3      Как следует из рис. 4.3, зона, где расположены остатки, расширяется, поэтому следует подобрать другую математическую зависимость. Такие же выводы получены при проверке на точность зависимости между переменными по коэффициенту точности выравнивания линии *r*1.  **4.2. Выборочный коэффициент корреляции**       Если зависимость между признаками на графике указывает на линейную корреляцию, рассчитывают коэффициент корреляции *r*, который позволяет оценить тесноту связи переменных величин, а также выяснить, какая доля изменений признака обусловлена влиянием основного признака, какая – влиянием других факторов. Коэффициент варьирует в пределах от –1 до +1. Если *r*=0, то связь между признаками отсутствует. Равенство *r*=0 говорит лишь об отсутствии линейной корреляционной зависимости, но не вообще об отсутствии корреляционной, а тем более статистической зависимости. Если *r* = ±1, то это означает наличие полной (функциональной) связи. При этом все наблюдаемые значения располагаются на линии регрессии, которая представляет собой прямую.      Практическая значимость коэффициента корреляции определяется его величиной, возведенной в квадрат, получившая название коэффициента детерминации.       Например, если *r* = 0,8, то *r*2 = 0,64, т.е. 64% всех изменений одного признака связано с изменением другого.       Выборочный коэффициент корреляции определяется равенством      ,              (4.4)      где *хi*, *уi* – варианты (наблюдавшиеся значения) признаков *Х* и *Y*; *n* – объем выборки;  – выборочные средние.     Чтобы получить исходные данные для формулы (4.4), сопряженные варианты обрабатывают по рекомендуемой форме (табл.4.3). Приведем расчет показателей для вычисления коэффициента корреляции *r* с использованием данных примера предыдущего параграфа.       Таблица 4.3   |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | № | *xi* |  | ()2 | *yi* |  | ()2 | ()() | | 1 2 3 4 5 6 7 8 | 1,5 4,0 5,0 7,0 8,5 10,0 11,0 12,5 | –5,9375 –3,4375 –2,4375 –0,4375 1,0625 2,5625 3,5625 5,0625 | 35,2539 11,8164 5,9414 0,1914 1,1280 6,5664 12,6914 25,6289 | 5,0 4,5 7,0 6,5 9,5 9,0 11,0 9,0 | –2,6875 –3,1875 –0,6875 –1,1875 1,8125 1,3125 3,3125 1,3125 | 7,2227 10,1602 0,4727 1,4102 3,2852 1,7227 10,9727 1,7297 | 15,9570 10,9570 1,6758 0,5195 1,9258 3,3633 11,8008 6,6445 | | Σ | 59,5 | 0 | 99,2187 | 61,5 | 0 | 36,9691 | 52,8437 |        *x*= 7,4375,  `*у* = 7,6875  .      Выборочный коэффициент корреляции *r* является оценкой коэффициента корреляции *r*ггенеральной совокупности. Допустим, что выборочный коэффициент оказался отличным от нуля. Так как выборка отобрана случайно, то еще нельзя заключить, что коэффициент корреляции генеральной совокупности *r*гтакже отличен отнуля. В конечном счете нас интересует именно этот коэффициент, поэтому возникает необходимость проверить гипотезу о значимости (существенности) выборочного коэффициента корреляции (или, что то же, о равенстве нулю коэффициента корреляции генеральной совокупности).      Для того чтобы при заданном уровне значимости a проверить нулевую гипотезу *H*o: *r*г= 0 о равенстве нулю генерального коэффициента корреляции нормальной двумерной случайной величины при конкурирующей гипотезе *H*1:*r*г≠ 0, надо вычислить наблюдаемое значение критерия:       и по таблице П.6. критических точек распределения Стьюдента, по заданному уровню значимости и числу степеней свободы ν = *n*-2 найти критическую точку *t*кр( a, ν) для двухсторонней критической области. Если ½*t*набл½ < *t*кр– нет оснований отвергнуть нулевую гипотезу. Если ½*t*набл½ >*t*кр – нулевую гипотезу отвергают.      Для данного примера найдем наблюдаемое значение критерия:      .      Поскольку *t*набл=4,32 >*t*кр= 2,45 при ν= 6 и a= 0,05, то нулевую гипотезу отвергаем. Другими словами, выборочный коэффициент корреляции значимо отличается от нуля, т.е. *Х* и *Y* коррелированны.      Подобный способ оценки значимости коэффициента корреляции не является безукоризненным, особенно если оцениваемый коэффициент корреляции по абсолютной величине близок к единице.      Более правильную оценку значимости *r*гможно получить, если воспользоваться преобразованием *Z*, предложенным Р.А. Фишером, где            (*Z*=*f*(*r*) см. в таблице П 9).       Критерий проверки гипотезы сводится к вычислению наблюдаемого значения:      и сравнению полученного *t*набл с *t*кр(α, ∞). При *t*набл ≥ *t*кр можно утверждать (с риском ошибиться в 100 × a % случаев), что связь имеется (*r*г ≠ 0).      К примеру, для *r*= 0,87, согласно таблице П.9, *Z* = 1,3331. При *n*= 8 что больше *t*кр(0,05, ∞)=1,96, поэтому можно считать коэффициент корреляции статистически значимым (т.е. можно утверждать, что *r*г≠ 0).      Использование преобразования *Z* дает возможность корректного получения интервальной оценки *r*г .Для этого сначала находятся доверительные границы для среднего значения  *M( Z*):       (*t*кр берется для ν= ∞). Затем, прибегая к помощи таблицы П.10, можно найти те значения *r*, которые соответствуют нижней и верхней границам для M(*Z*).      Так для нашего примера получим (*n*=8; *r* = 0,87; *Z*= 1,3331; α=0,05):      ,    т.е.       0,40 <*M(Z*)<2,15       Обращаясь к таблице П.10, найдем, что доверительные границы коэффициента корреляции оказываются равными  *r*0,05= 0,38÷ 0,97.       Все операции по проверке значимости коэффициента корреляции можно упростить, заранее вычислив для различных абсолютных значений оценок *r* минимальные объемы корреляционных рядов, обеспечивающих возможность утверждать с уровнем значимости α, что *r*г ≠ 0, т.е. утверждения наличия линейной связи (таблица П.9).      Та же таблица может служить для оценки необходимого и достаточного числа повторности *n*α, чтобы при ожидаемой величине *r* коэффициента корреляции можно было утверждать, что связь есть (*r*г≠ 0) при заданном уровне значимости α. Так, воспользовавшись таблицей П.9, мы обнаружим, что коэффициент корреляции, оценка которого равна 0,87, можно считать статистически значимым с α= 0,05, если *n*, по крайней мере, равно 6. У нас повторяемость*n*=8, что больше 6, следовательно, коэффициент корреляции значим. И минимальная повторяемость, которая может обеспечить значимость коэффициента корреляции при *r* = 0,87, есть *n*0,05 = 6, что следует иметь в виду, если опыт планируется повторить. |
|  |