Практическое занятие 3

Решить следующую задачу:

Истинные значения параметров регрессии равны соответственно 2 и (0,5 + к):

y = 2 + (0,5 + к)x + u

x принимает значения от 1 до 20; u - случайные числа.

Оценить регрессионную зависимость

Здесь к – порядковый номер студента по журналу.

Таблица 7.1

X	*u*	*Y*	*X*	*u*	*Y*
1	**-0,59**		**11**	**1,59**	
2	**-0,24**		**12**	**-0,92**	
3	**-0,83**		**13**	**-0,71**	
4	**0,03**		**14**	**-0,25**	
5	**-0,38**		**15**	**1,69**	
6	**-2,19**		**16**	**0,15**	
7	**1,03**		**17**	**0,02**	
8	**0,24**		**18**	**-0,11**	
9	**2,53**		**19**	**-0,91**	
10	**-0,13**		**20**	**1,42**	

Практическое занятие 3

Решить следующую задачу:

Истинные значения параметров регрессии равны соответственно 2 и (0,5 + к):

y = 2 + (0,5 + к)x + u

x принимает значения от 1 до 20; u - случайные числа.

Оценить регрессионную зависимость

Здесь к – порядковый номер студента по журналу.

Таблица 7.1

X	*u*	*Y*	*X*	*u*	*Y*
1	**-0,59**		**11**	**1,59**	
2	**-0,24**		**12**	**-0,92**	
3	**-0,83**		**13**	**-0,71**	
4	**0,03**		**14**	**-0,25**	
5	**-0,38**		**15**	**1,69**	
6	**-2,19**		**16**	**0,15**	
7	**1,03**		**17**	**0,02**	
8	**0,24**		**18**	**-0,11**	
9	**2,53**		**19**	**-0,91**	

10	**-0,13**		**20**	**1,42**	

Методические указания к выполнению заданий

С помощью регрессионного анализа мы можем получить оценки параметров зависимости. Однако они являются лишь ***оценками.*** Поэтому возникает вопрос о том, насколько они надежны. Дадим сначала общий ответ, изучив условия несмещенности и факторы, определяющие дисперсию оценок. Основываясь на этом, мы будем совершенствовать способы проверки совместимости регрессионной оценки с конкретной априорной гипотезой об истинном значении оцениваемого параметра. И следовательно, мы будем строить доверительный интервал для истинного значения, который представляет собой множество всех возможных гипотетических значений, не противоречащих результатам экспериментов. Будет также показано, каким образом можно проверить, является ли качество подбора кривой более высоким, чем при чисто случайном подборе.

Коэффициент регрессии, вычисленный методом наименьших квадратов, – это особая форма случайной величины, свойства которой зависят от свойств случайной составляющей в уравнении. Мы продемонстрируем это сначала теоретически, а затем посредством контролируемого эксперимента. В частности, мы увидим, какое значение для оценки коэффициентов регрессии имеют некоторые конкретные предположения, касающиеся остаточного члена.

В ходе рассмотрения мы постоянно будем иметь дело с моделью парной регрессии, в которой ***y*** связан с ***x*** следующей зависимостью:

$$y = \alpha + \beta x + u \quad (3.1)$$

и на основе ***n*** выборочных наблюдений будем оценивать уравнение регрессии

$$\hat{y} = a + bx \quad (3.2)$$

Мы также будем предполагать, что ***x*** — это неслучайная экзогенная переменная. Иными словами, ее значения во всех наблюдениях можно считать заранее заданными и никак не связанными с исследуемой зависимостью.

Во-первых, заметим, что величина ***y*** состоит из двух составляющих. Она включает неслучайную составляющую ($\alpha + \beta x$), которая не имеет ничего общего с законами вероятности (**α** и **β** могут быть неизвестными, но тем не менее это постоянные величины), и случайную составляющую ***и.***

Отсюда следует, что, когда мы вычисляем ***b*** по обычной формуле:

$$b = \frac{Cov(x,y)}{Var(x)} \qquad (3.3)$$

b также содержит случайную составляющую. $Cov(x,y)$ зависит от значений ***y,*** а ***y*** зависит от значений ***u.***

Если случайная составляющая принимает разные значения в ***n*** наблюдениях, то мы получаем различные значения ***y*** и, следовательно, разные величины $Cov(x,y)$ и ***b.***

Теоретически мы можем разложить ***b*** на случайную и неслучайную составляющие. Воспользовавшись соотношением (3.1), а также правилом ***1*** расчета ковариации из раздела 1.2, получим:

$$Cov(x,y) = Cov(x,[\alpha + \beta x + u]) = Cov(x,\alpha) + Cov(x,\beta x) + Cov(x,u) \qquad (3.4)$$

По ковариационному правилу ***3,*** ковариация $Cov(x,\alpha)$ равна нулю. По ковариационному правилу ***2,*** ковариация $Cov(x,\beta x)$ равна $Cov(x,u)$. Причем $Cov(x,x)$ это тоже, что и $Var(x)$. Следовательно, мы можем записать:

$$Cov(x,y) = \beta Var(x) + Cov(x,u) \qquad (3.5)$$

и, таким образом,

$$b = \frac{Cov(x,y)}{Var(x)} = \beta + \frac{Cov(x,u)}{Var(x)} \qquad (3.6)$$

Итак, мы показали, что коэффициент регрессии *b,* полученный по любой выборке, представляется в виде суммы двух слагаемых: 1) постоянной величины, равной истинному значению коэффициента β ; 2) случайной составляющей, зависящей от $Cov(x,u)$ **,** которой обусловлены

отклонения коэффициента b от константы β. Аналогичным образом можно показать, что a имеет постоянную составляющую, равную истинному значению α, плюс случайную составляющую, которая зависит от случайного фактора u.

Следует заметить, что на практике мы не можем разложить коэффициенты регрессии на составляющие, так как не знаем истинных значений α и β или фактических значений u в выборке. Они интересуют нас потому, что при определенных предположениях позволяют получить некоторую информацию о теоретических свойствах a и b.

Эксперимент по методу Монте-Карло

По-видимому, никто точно не знает, почему ***эксперимент по методу Монте-Карло*** называется именно так. Возможно, это название имеет какое-то отношение к известному казино как символу действия законов случайности.

Основное понятие будет объяснено посредством аналогии. Предположим, что свинья обучена находить трюфели. Это дикорастущие земляные грибы, встречающиеся во Франции и Италии и считающиеся деликатесом. Они дороги, так как их трудно найти, и хорошая свинья, обученная поиску трюфелей, стоит дорого. Проблема состоит в том, чтобы узнать, насколько хорошо свинья ищет трюфели. Она может находить их время от времени, но возможно также, что большое количество трюфелей она пропускает. В случае действительной заинтересованности вы могли бы выбрать участок земли, закопать трюфели в нескольких местах, отпустить свинью и посмотреть, сколько грибов она обнаружит. Посредством такого контролируемого эксперимента можно было бы непосредственно оценить степень успешности поиска.

Какое отношение это имеет к регрессионному анализу? Проблема в том, что мы никогда не знаем истинных значений α и β (иначе зачем бы мы использовали регрессионный анализ для их оценки?). Поэтому мы не можем сказать, хорошие или плохие оценки дает наш метод. Эксперимент по методу Монте-Карло – это искусственный контролируемый эксперимент, дающий возможность такой проверки. Простейший возможный эксперимент по методу Монте-Карло состоит из трех частей. Во-первых:

1) выбираются истинные значения α и β;
2) в каждом наблюдении выбирается значение x;
3) используется некоторый процесс генерации случайных чисел (или берется последовательность из таблицы случайных чисел) для получения

значений случайного фактора u в каждом из наблюдений.

Во-вторых, в каждом наблюдении генерируется значение ***y*** с использованием соотношения (3.1) и значений α, β, x **и** u**.**

В-третьих, применяется регрессионный анализ для оценивания параметров a и b с использованием только полученных указанным образом значений ***y*** и данных для ***x.*** При этом вы можете видеть, являются ли a и b хорошими оценками α и β, и это позволит почувствовать пригодность метода построения регрессии.

На первых двух шагах проводится подготовка к применению регрессионного метода. Мы полностью контролируем модель, которую создаем, и знаем истинные значения параметров, потому что сами их определили. На третьем этапе мы определяем, может ли поставленная нами задача решаться с помощью метода регрессии, т. е. могут ли быть получены хорошие оценки для α и β при использовании только данных об ***y*** и ***x.*** Заметим, что проблема возникает вследствие включения случайного фактора в процесс получения ***y.*** Если бы этот фактор отсутствовал, то точки, соответствующие значениям каждого наблюдения, лежали бы точно на прямой (3.1) и точные значения α и β можно было бы очень просто определить по значениям ***y*** и ***x***.

Произвольно положим $\alpha = 2$ и $\beta = 0{,}5$, так что истинная зависимость имеет вид:

$$\boldsymbol{y = 2 + 0,5x + u.} \qquad (3.7)$$

Предположим для простоты, что имеется 20 наблюдений и что ***x*** принимает значения от 1 до 20. Для случайной остаточной составляющей ***u*** будем использовать случайные числа, взятые из нормально распределенной совокупности с нулевым средним и единичной дисперсией. Нам потребуется набор из 20 значений, обозначим их rn_1 ..., rn_{20}**.** Случайный член u_1, в первом наблюдении просто равен rn_1 ***x*** и т.д.

Зная значения ***x*** и ***u*** в каждом наблюдении, можно вычислить значения ***y,*** используя уравнение (3.7); это сделано в табл. 3.1. Теперь при оценивании регрессионной зависимости - y от ***x*** получим:

$$\hat{y} = 1{,}63 + 0{,}54x. \qquad (3.8)$$

В данном случае оценка ***a*** приняла меньшее значение (1,63) по сравнению с α (2,00), а b немного выше β (0,54 по сравнению с 0,50). Расхождения вызваны совместным влиянием случайных членов в 20 наблюдениях.

Очевидно, что одного эксперимента такого типа едва ли достаточно для оценки качества метода регрессии. Он дал довольно хорошие результаты, но, возможно, это лишь счастливый случай. Для дальнейшей проверки повторим эксперимент с ***тем же*** истинным уравнением (3.7) и с ***теми же*** значениями ***x,*** но с ***новым*** набором случайных чисел для остаточного члена, взятых из того же распределения (нулевое среднее и единичная дисперсия). Используя эти значения и значения ***x,*** получим новый набор значений ***y.***

В целях экономии места таблица с новыми значениями ***u*** и ***y*** не приводится. Вот результат оценивания регрессии между новыми значениями ***y*** и ***x***:

$$\hat{y} = 2{,}52 + 0{,}48x \qquad \textbf{(3.9)}$$

Таблица 3.1

X	u	У	X	u	У
1	-0,59	1,91	11	1,59	9,09
2	-0,24	2,76	12	-0,92	7,08
3	-0,83	2,67	13	-0,71	7,79
4	0,03	4,03	14	-0,25	8,75
5	-0,38	4,12	15	1,69	11,19
6	-2,19	2,81	16	0,15	10,15
7	1,03	6,53	17	0,02	10,52
8	0,24	6,24	18	-0,11	10,89
9	2,53	9,03	19	-0,91	10,59
10	-0,13	6,87	20	1,42	13,42

Второй эксперимент также был успешным. Теперь a оказалось больше α, а b — несколько меньше β. В табл. 3.2 приведены оценки a и b при 10-кратном повторении эксперимента с использованием разных наборов случайных чисел в каждом варианте.

Можно заметить, что, несмотря на то что в одних случаях оценки принимают заниженные значения, а в других — завышенные, в целом значения a и b группируются вокруг истинных значений α и β, равных соответственно 2,00 и 0,50. При этом хороших оценок получено больше, чем плохих. Например, фиксируя значения b при очень большом числе повторений эксперимента, можно построить таблицу частот и получить аппроксимацию функции плотности вероятности, показанную на рис. 3.1. Это нормальное распределение со средним 0,50 и стандартным отклонением 0,0388.