Лабораторная работа № 3

Тема: КОРРЕЛЯЦИИ

Цель: Изучение способов анализа парных зависимостей между признаками с применением коэффициента корреляции Пирсона и метода наименьших квадратов для расчета коэффициентов регрессионной прямой.

I. Краткие теоретические сведения

Говорят, что ***переменные зависимы***, если их значения каким-то образом согласованы друг с другом в имеющихся наблюдениях.

Мера зависимости между двумя переменными называется парной корреляцией или просто ***корреляцией***.

Говорят, что две переменные *положительно* коррелированны, если при *увеличении* значений одной переменной *увеличиваются* значения другой переменной. Две переменные *отрицательно* коррелированны, если при *увеличении* одной переменной другая переменная *уменьшается.*

Если исследуется зависимость между двумя переменными, измеренными в *интервальной* шкале, наиболее подходящим коэффициентом будет коэффициент корреляции Пирсона , называемый также ***линейной корреляцией,*** так как он отражает степень *линейных* связей между переменными.

Коэффициент *парной* корреляции изменяется в пределах от -1 до +1. *Крайние значения имеют особенный смысл.* Значение -1 означает полную отрицательную зависимость, значение +1 означает полную положительную зависимость. Значение 0,00 интерпретируется как отсутствие корреляции.

Говорят, что корреляция высокая, если на графике зависимость между переменными можно с большой точностью представить прямой линией (с положительным или отрицательным наклоном). Совокупность точек графика при этом называется ***диаграммой рассеяния***.

Проведенная прямая, вокруг которой группируются значения переменных, называется ***прямой регрессии***, или прямой, построенной методом наименьших квадратов.

Формально коэффициент корреляции r_{xy} Пирсона между переменными X и Y вычисляется следующим образом:

$$r_{XY} = r(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X}) \times (Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \times (Y_i - \overline{Y})^2}}$$

где $\overline{X}$ - среднее переменной X,

$\overline{Y}$ - среднее переменной Y.

Даная формула предполагает, что из каждого значения x_i переменной X, должно вычитаться ее среднее значение $\overline{x}$. Это не удобно, поэтому для расчета коэффициента корреляции используют не данную формулу, а ее аналог, получаемый с помощью преобразований:

$$rx = \frac{n \times \sum (x_i \times y_i) - (\sum x_i \times \sum y_i)}{\sqrt{\left[n \times \sum x_i^2 - (\sum x_i)^2\right] \times \left[n \times \sum y_i^2 - (\sum y_i)^2\right]}}$$

II. Расчет коэффициентов линии регрессии методом наименьших квадратов (МНК)

МНК используется для вычисления коэффициентов линии регрессии, которая строится на диаграмме рассеяния.

Пусть в качестве исходных данных имеем таблицу,

X	x_1	x_2	…	x_n
Y	y_1	y_2	…	y_n

содержащую статистические данные, или данные экспериментов. Если в качестве X выступает время, то будем иметь динамический ряд (т.е. тогда x_i будут размещены в возрастающем порядке). Необходимо получить аналитическую зависимость вида

$$\hat{Y} = f(x), \qquad \ldots(1)$$

которая наилучшим образом описывает начальные данные. Словосочетание «наилучшим образом», будем понимать в смысле минимума суммы квадратов отклонений значений y_i, данных в таблице, от $\hat{y}_i$, рассчитанных по формуле …(1).

$$E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 . \qquad \ldots(2)$$

Определение зависимости …(1) необходимо, в том числе, и для нахождения $\hat{Y}_{n+1} = f(x_{n+1})$, что уже представляет собой задачу прогнозирования.

Построим диаграмму рассеяния (т.е. нанесем точки из таблицы на координатную плоскость) и сделаем предположение, что зависимость …(1) является линейной (причем $\hat{Y} = a + bx$), а отклонения от прямой вызваны случайными факторами.

Определим уравнение регрессионной прямой (найдем значения коэффициентов a и b), так, чтобы получить решение задачи $E \to \min$, т.е. необходимо найти минимум функции

$$E = \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

Функция E = E(a,b). Продифференцируем E по a и b. Получим:

$$\frac{\partial E}{\partial a} = -2\sum_{i=1}^{n}(y_i - (a + bx_i)),$$

$$\frac{\partial E}{\partial b} = -2\sum_{i=1}^{n}(y_i - (a + bx_i))x_i .$$

Для того, чтобы найти минимум функции E(a,b), приравняем нулю производные и упростим систему:

$$\begin{cases} -2\sum_{i=1}^{n}(y_i - (a + bx_i)) = 0, \\ -2\sum_{i=1}^{n}(y_i - (a + bx_i))x_i \end{cases} \Rightarrow \begin{cases} \sum_{i=1}^{n} y_i - na - b\sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 = 0, \end{cases} \Rightarrow \begin{cases} na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i , \\ a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

Последнюю систему можно представить в матричном виде:

$$\begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix}$$

Решая её, получим:

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\left(\sum_{i=1}^{n} x_i\right)^2 + n\sum_{i=1}^{n} x_i^2}, \quad a = \bar{y} - b\overline{x} .$$

Вычислив a и b, получим функцию $\hat{Y} = a + bx$, которая в классе линейных функций наилучшим образом описывает табличную зависимость в смысле минимума суммы квадратов отклонений. На основании этой функции можно рассчитывать прогноз

$\hat{Y}_{n+1} = a + bx_{n+1}$

III. Пример корреляционного анализа

Задача: По выборочным данным, приведенным в табл.1, требуется установить наличие взаимосвязи между указанными показателями в центральном регионе России.

Таблица 1.

Область	***Уровень образования***	***Отнош-е числа безр-х к числу вакансий***	***Ур-нь преступ-ти***
Брянская	735	22,3	908
Владимирская	788	10,8	791
Ивановская	779	52,9	804
Калужская	795	2,2	701
Костромская	740	10,4	685
г Москва	902	0,4	496
Московская	838	2,4	536

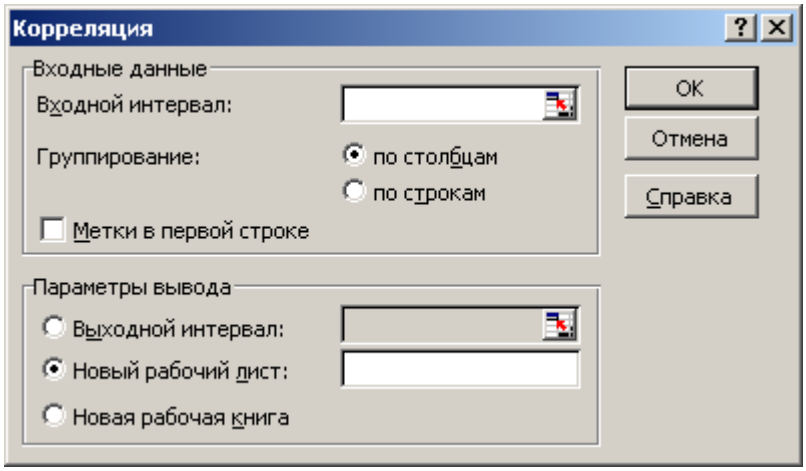
Нижегородская	763	5,4	936
Орловская	762	4,1	662
Рязанская	757	4,1	671
Смоленская	772	1	920
Тверская	764	4,2	1040
Тульская	764	2,1	809
Ярославская	755	25,1	882

Примечания: 1. Уровень образования рассчитывается как численность лиц с высшим и средним образованием на 1000 жителей области.

2. Уровень преступности рассчитывается как число совершенных преступлений на 100 тыс. жителей области.

Для решения задачи воспользуемся режимом «Корреляция».

В диалоговом окне данного режима задаются следующие параметры:



1. Входной интервал.
2. Группирование (переключение по столбцам/по строкам)
3. Метки
4. Выходной интервал/ новый рабочий лист/ новая рабочая книга.

Рассчитанные в этом режиме данные приведены в табл.2.

Таблица 2.

	Уровень образования	*Отнош-е числа безр-х к числу вакансий*	*Ур-нь преступ-ти*
Уровень образования	1		
Отнош-е числа безр-х к числу вакансий	-0,2	1	
Ур-нь преступ-ти	-0,66	0,19	1

Как видно из таблицы 2, между парами всех исследуемых показателей имеются стохастические связи. Причем характер всех выявленных связей различен и состоит в следующем:

- связь «Уровень образования» - «Отнош-е числа безр-х к числу вакансий» является слабой и обратной (r_{xy}=-0,2), т.е. с повышением уровня образования отношение числа безработных к числу вакансий уменьшается;

- связь «Уровень образования» - «Уровень преступности» является заметной и обратной (r_{xy}=-0,66), т.е. с повышением уровня образования уровень преступности уменьшается;

- связь «Отнош-е числа безр-х к числу вакансий» - «Уровень преступности» является слабой и прямой (r_{xy}=0,19).

Для облегчения выводов относительно практической значимости коэффициенту корреляции дается качественная оценка. Это осуществляется на основе шкалы Чеддока

Коэффициент r_{xy}	0,1 – 0,3	0,3 – 0,5	0,5 – 0,7	0,7 – 0,9	0,9 – 0,99
Характеристика силы связи	слабая	умеренная	заметная	значительная	очень значительная

Пользоваться режимом «Корреляция» удобно, если необходимо найти коэффициент корреляции между несколькими парами случайных величин.

Для расчета коэффициента корреляции между парой переменных можно использовать одну из приведённых в лекционном курсе формул расчёта коэффициента корреляции Пирсона.

IV. ЗАДАНИЯ для аудиторной работы

Вариант 1. Требуется на основе выборочных данных о деловой активности однотипных коммерческих структур оценить тесноту связи между прибылью Y (млн. руб.) и затратами X (руб) на производство единицы продукции. Исходные данные приведены в таблице В1

Таблица В1

№ пп	Y	X
1	221	96
2	1070	77
3	1001	77
4	606	89
5	779	82
6	789	81

Вариант 2. Имеется выборка из генеральной совокупности системы двух случайных величин (X,Y), приведенная в таблице В2. Требуется оценить тесноту связи между этими величинами.

Таблица В2

x_i	12.1	14.7	20.5	11.2	16.6	10.0	13.0	14.9	16.3	15.1
y_i	53.2	44.2	51.4	57.7	45.5	42.0	53.5	68.9	57.7	63.3

Вариант 3. С целью анализа взаимосвязи показателей эффективности производства продукции была отобрана группа из десяти однотипных предприятий. Оцените тесноту взаимосвязи производительности труда (x1) и фондоотдачи (x2) по данным, приведенным в таблице В3.

Таблица В3

№ предприятия	X1	X2
1.	6	2
2.	4,9	0,8
3.	7	2,7
4.	6,7	3
5.	5,8	1
6.	6,1	2,1
7.	5	0,9
8.	6,9	2,6
9.	6,8	3
10.	5,9	1,1

Вариант 4. С целью анализа взаимосвязи показателей эффективности производства продукции была отобрана группа из десяти однотипных предприятий. Оцените тесноту взаимосвязи производительности труда (x1) и материалоемкости производства (x2) по данным, приведенным в таблице В4.

Таблица В4

№ предприятия	X1	X2
1.	6	25
2.	4,9	30
3.	7	20
4.	6,7	21
5.	5,8	28
6.	6,1	26
7.	5	30
8.	6,9	22
9.	6,8	20
10.	5,9	29

Вариант 5. С целью анализа взаимосвязи показателей эффективности производства продукции была отобрана группа из десяти однотипных предприятий. Оцените тесноту взаимосвязи фондоотдачи (x1) и материалоемкости производства (x2) по данным, приведенным в таблице В5.

Таблица В5

№ предприятия	X1	X2

1.	2	25
2.	0,8	30
3.	2,7	20
4.	3	21
5.	1	28
6.	2,1	26
7.	0,9	30
8.	2,6	22
9.	3	20
10.	1,1	29

Вариант 6. Требуется оценить тесноту связи между двумя рядами данных, приведенных в таблице В6

Таблица В6

	1935	1936	1937	1938	1939	1940	1941	1942	1943	1944	1945	1946
Детская смертность (до 1 года, на тыс. человек)	60	62	61	55	53	60	63	53	52	48	49	43
Потребление пива (в объемных баррелях)	23	23	25	25	26	26	29	30	30	32	33	31

Вариант 7. Требуется оценить тесноту связи между показателями смертности мужчин в ряде стран за рассматриваемый период, приведенными в таблице В7

Таблица В7

Смертность мужчин от зависящих от алкоголя причин
(хронический алкоголизм, алкогольный психоз, цирроз печени,
случайные отравления алкоголем) в некоторых странах мира
(на 100 000 лиц мужского пола)

Страна	1970- 1974гг.	1985- 1989гг.
Латвия	42,8	29,3
Россия	57,7	49,4
Франция	70,7	42,5
Венгрия	26,2	81,8
Германия	44,2	38,4
Бывшая ГДР	23,8	42,0
Польша	27,1	34,2
Финляндия	25,9	37,1
Великобритания	6,7	9,1
Швеция	20,8	20,3
США	34,3	24,8

Япония	29,5	22,7
Норвегия	11,6	21,0

Вариант 8. Требуется оценить тесноту связи между среднемесячной заработной платой работников разных отраслей и выплатами социального характера по данным, приведенным в таблице В8

Таблица В8

Размер среднемесячной заработной платы и выплат социального характера по отраслям народного хозяйства по состоянию на декабрь 1995г.

Тыс. руб.

Отрасли	Среднемесячная заработная плата	Выплаты социального характера на 1 работника в среднем за мес
Промышленность	682,5	78,2
Сельское хоз-во	268,6	21,5
Строительство	806,4	106,6
Транспорт	936,7	55,5
Связь	718,4	64,0
ЖКХ	676,3	71,3
Здравоохр-ние, соц. обеспечение	516,8	25,3
Образование	472,9	27,5
Культура и искусство	412,6	51,3
Наука и научное обслуживание	483,3	37,3
Финансы и кредит	808,5	291,1

Вариант 9. Требуется оценить тесноту связи между мнениями работников издательства и журналистов об условиях работы коллектива редакции (в %) по данным, приведенным в таблице В9

Таблица В9

Трудности в работе редакции	Мнение работников издательства	Мнение журналистов
Недостаток транспорта	46	74
Неукомплектованность кадрами	30	37
Недостаточная квалифицированность кадров	29	39
Плохое техническое оснащение типографии	19	46
Малопригодные для работы помещения	13	32

Отсутствие внимания к бытовым нуждам работников	5	43
Плохая связь	3	16
Сложная обстановка в коллективе	4	8
Плохое техническое состояние редакции	2	45

Вариант 10. Требуется оценить тесноту связи между высказываниями респондентов в 1994 и 1996 годах по данным, приведенным в таблице В10.

Таблица В10

Доверие к действующим в РФ общественным структурам и институтам власти

(в % к числу опрошенных)

Варианты	1994г.	1996г.
Президент РФ	20,1	20,0
Правительство РФ	14.0	14.0
Совет Федерации	11.1	11.0
Государственная Дума	16.4	15.0
Руководители регионов	13.3	17.0
Милиция	13.2	13.0
Суд	14.0	15.0
Прокуратура	14.0	14.0
Армия	38.2	31.0
Профсоюзы	16.1	17.0
Политические партии	5.4	9.0
Средства массовой инфор-	18.5	13.0
Директора, руководители	15.1	11.0
Банковские, предпринима-	9,5	8,0

V. Порядок выполнения работы

1. Определить коэффициент линейной корреляции следующими способами:

а) с помощью режима «Корреляция» Пакета Анализа MS Excel;

б) с помощью расчётных формул коэффициента корреляции Пирсона.

Сравнить полученные результаты.

2. На основании таблицы Чеддока дать качественную оценку коэффициенту Пирсона.

3. Построить диаграмму рассеяния исходных данных. Сделать вывод о соответствии данных диаграммы рассчитанному коэффициенту линейной корреляции.

4.Разработать и реализовать в табличном процессоре Excel алгоритм расчета коэффициентов прямой регрессии.

5. Построить прямую регрессии на диаграмме рассеяния. Один из способов построения такой прямой заключается в расчете любых двух точек, принадлежащих этой прямой и нанесении их на диаграмму рассеяния. Прямая, проведенная, через эти две точки – прямая регрессии.

6. Ответить на вопрос: подтверждает ли визуальный характер построенной прямой выводы, сделанные в п.2?

7. Сделать выводы по работе.

VI. Отчет о выполнении лабораторной работы

Отчет о выполнении лабораторной работы должен содержать

1) формулировку темы и цели работы;
2) формулировку заданий на лабораторную работу (постановка задачи и задания из пункта V)
3) комментарии и выводы ко всем выполненным заданиям
4) печатные формы исходных и полученных в результате анализа таблиц и диаграмм.

VII. Контрольные вопросы

1. Понятие корреляции между парой переменных.
2. Линейная корреляция. Расчёт коэффициента линейной корреляции
3. Расчет коэффициентов линии регрессии методом наименьших квадратов
4. Качественная оценка коэффициента корреляции.
5. Нелинейные зависимости между переменными.
6. Ложные корреляции.