Лабораторная работа

«Обработка и оценка результатов исследования»

Цель работы – научиться использовать возможности MS Excel для проведения корреляционного и регрессионного анализа исследовательских данных, планирования и обработки результатов факторного эксперимента.

Учебные вопросы:

1. Возможности прикладного программного обеспечения на этапах обработки и оценки результатов исследования.

Изучив данную тему, студент должен:

знать:

- назначение существующих современных средств компьютеризации научных исследований, их функциональные возможности и особенности применения;

уметь:

- производить обработку и оценку результатов исследования.

1.1. Краткое изложение основных теоретических и методических аспектов работы

Параметрический корреляционный анализ

Одна из наиболее распространенных задач статистического исследования состоит в изучении связи между выборками (наборами числовых данных каких-либо величин). Обычно связь между выборками носит не функциональный, а вероятностный (или стохастический) характер. В этом случае нет строгой, однозначной зависимости между величинами. При изучении стохастических зависимостей различают *корреляцию* и *регрессию*.

Корреляционный анализ состоит в определении степени связи между двумя случайными величинами *X* и *Y*. В качестве меры такой связи используется коэффициент корреляции. Коэффициент корреляции оценивается по выборке объема *n* связанных пар наблюдений (x_i, y_i) из совместной генеральной совокупности *X* и *Y*. Существует несколько типов коэффициентов корреляции, применение которых зависит от измерения (способа шкалирования) величин *X* и *Y*.

Для оценки степени взаимосвязи величин *X* и *Y*, измеренных в количественных шкалах, используется коэффициент линейной корреляции (коэффициент Пирсона), предполагающий, что выборки *X* и *Y* распределены по нормальному закону.

Линейный коэффициент корреляции – параметр, который характеризует степень линейной взаимосвязи между двумя выборками, рассчитывается по формуле:

$$r_{xy} = \frac{\sum(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\cdot\sum(y_i-\bar{y})^2}}, \qquad (1)$$

где x_i – значения, принимаемые в выборке *X*,

y_i – значения, принимаемые в выборке *Y*;

$\bar{x}$ – средняя по *X*, $\bar{y}$ – средняя по *Y*.

Коэффициент корреляции изменяется от –1 до 1. Когда при расчете получается величина большая +1 или меньшая –1 – следовательно, произошла ошибка в вычислениях. При значении 0 линейной зависимости между двумя выборками нет.

Знак коэффициента корреляции очень важен для интерпретации полученной связи (таблица 1). Если знак коэффициента линейной корреляции «+», то связь между коррелирующими признаками такова, что большей величине одного признака (переменной) соответствует большая величина другого признака (другой переменной). Иными словами, если один показатель (переменная) увеличивается, то соответственно увеличивается и другой показатель (переменная). Такая зависимость носит название прямо пропорциональной зависимости.

Таблица 1.

Теснота связи и величина коэффициента корреляции.

Коэффициент корреляции r_{xy}	**Теснота связи**
+(0,91 … 1,00)	Очень сильная

±(0,81… 0,90)	Весьма сильная
± (0,65… 0,80)	Сильная
± (0,45… 0,64)	Умеренная
± (0,25… 0,44)	Слабая
До ± 0,25	Очень слабая
«+» – прямая зависимость, «–» – обратная зависимость	

Если же получен знак «–», то большей величине одного признака соответствует меньшая величина другого. Иначе говоря, при наличии знака минус, увеличению одной переменной (признака, значения) соответствует уменьшение другой переменной. Такая зависимость носит название обратно пропорциональной зависимости.

t-статистика Стьюдента

Для того чтобы оценить наличие связи между двумя переменными, также можно использовать *t-статистику Стьюдента*, которая оценивает отношение величины линейного коэффициента корреляции к среднему квадратическому отклонению и рассчитывается по формуле

$$t_{\text{расч}} = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} \qquad (2)$$

Полученную величину $t_{расч}$ сравнивают с табличным значением $t_{табл}$ критерия Стьюдента с $n-2$ степенями свободы. Если $t_{расч} > t_{табл}$, то практически невероятно, что найденное значение обусловлено только случайными совпадениями величин X и Y в выборке из генеральной совокупности, т.е. существует зависимость между X и Y. И наоборот, если $t_{расч} < t_{табл}$, то величины X и Y независимы.

Исследование связей между двумя переменными в Excel

Условие задачи: По 10 интернет-магазинам были определены затраты на рекламную раскрутку сайтов и количество покупателей, воспользовавшихся после ее проведения услугами каждого магазина. Определить коэффициент корреляции между исследуемыми признаками.

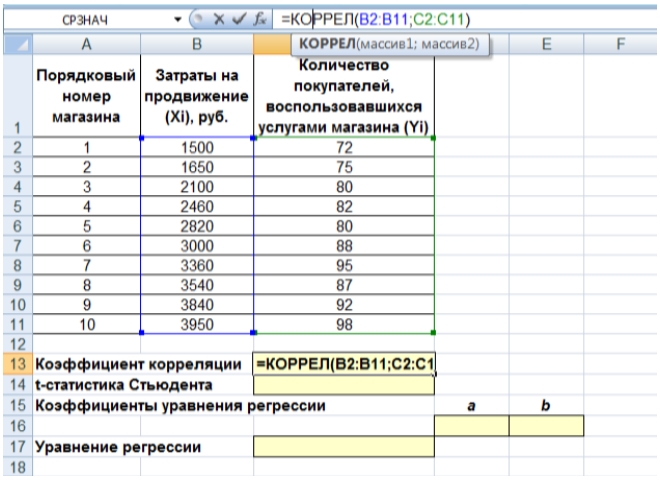
Ход выполнения:

1. Открываем новую книгу MS Excel и создаем таблицу согласно рис. 2.
2. Рассчитываем в ячейке C12 коэффициент корреляции, используя функцию КОРРЕЛ из категории Статистические.

Синтаксис функции: КОРРЕЛ (<массив 1>;<массив 2>),
где <массив 1> – ссылка на диапазон ячеек первой выборки (X);
<массив 2> – ссылка на диапазон ячеек второй выборки (Y).
В нашей задаче формула будет иметь вид: =КОРРЕЛ(B2:B11;C2:C11) – см. рис. 3.

	A	B	C	D	E	F
1	Порядковый номер магазина	Затраты на продвижение (Xi), руб.	Количество покупателей, воспользовавшихся услугами магазина (Yi)			
2	1	1500	72			
3	2	1650	75			
4	3	2100	80			
5	4	2460	82			
6	5	2820	80			
7	6	3000	88			
8	7	3360	95			
9	8	3540	87			
10	9	3840	92			
11	10	3950	98			
12						
13	Коэффициент корреляции					
14	t-статистика Стьюдента					
15	Коэффициенты уравнения регрессии			a	b	
16						
17	Уравнение регрессии					
18						

Рис. 2. Исходные данные для исследования связей между двумя переменными



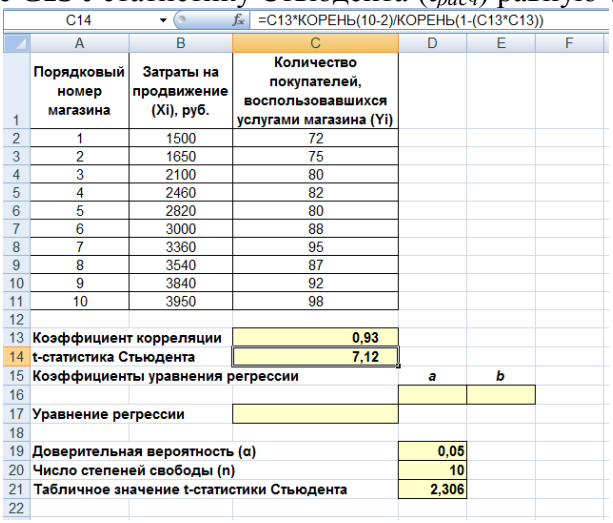
	A	B	C	D	E	F
1	Порядковый номер магазина	Затраты на продвижение (Xi), руб.	Количество покупателей, воспользовавшихся услугами магазина (Yi)			
2	1	1500	72			
3	2	1650	75			
4	3	2100	80			
5	4	2460	82			
6	5	2820	80			
7	6	3000	88			
8	7	3360	95			
9	8	3540	87			
10	9	3840	92			
11	10	3950	98			
12						
13	Коэффициент корреляции	=КОРРЕЛ(B2:B11;C2:C1				
14	t-статистика Стьюдента					
15	Коэффициенты уравнения регрессии			a	b	
16						
17	Уравнение регрессии					
18						

Рис. 3. Вычисление коэффициента корреляции

3. Сделаем вывод о тесноте связи между затратами на рекламную раскрутку сайтов и количество покупателей.

   После ввода формулы получаем в ячейке C13 значение коэффициента корреляции равное 0,93. По таблице 2 делаем вывод, что связь между переменными очень сильная, т.е. имеет место линейная зависимость (прямая пропорциональность).

4. Оценим значимость коэффициента корреляции. С этой целью рассмотрим две гипотезы. Основную H_0: $r_{xy}=0$ и альтернативную H_1: $r_{xy}\neq 0$. Для проверки гипотезы H_0 рассчитаем в ячейке C14 t-статистику Стьюдента по формуле, указанной в 3.1.2. В нашем случае число степеней свободы $\nu = n - 2 = 10 - 2 = 8$ и формула будет следующей: =C13*КОРЕНЬ(10-2)/КОРЕНЬ(1-(C13*C13)). После ввода формулы получаем в ячейке C13 t-статистику Стьюдента ($t_{расч}$) равную 7,12 (рис. 4).



	A	B	C	D	E	F
1	Порядковый номер магазина	Затраты на продвижение (Xi), руб.	Количество покупателей, воспользовавшихся услугами магазина (Yi)			
2	1	1500	72			
3	2	1650	75			
4	3	2100	80			
5	4	2460	82			
6	5	2820	80			
7	6	3000	88			
8	7	3360	95			
9	8	3540	87			
10	9	3840	92			
11	10	3950	98			
12						
13	Коэффициент корреляции		0,93			
14	t-статистика Стьюдента		7,12			
15	Коэффициенты уравнения регрессии			a	b	
16						
17	Уравнение регрессии					
18						
19	Доверительная вероятность (α)			0,05		
20	Число степеней свободы (n)			10		
21	Табличное значение t-статистики Стьюдента			2,306		
22						

Рис. 4. Вычисление t-статистики Стьюдента ($t_{расч}$)

5. Сравним полученное значение с критическим значением $t_{\nu,\alpha,табл}$ распределения Стьюдента (при $\nu = 8$ и доверительной вероятности $\alpha = 0,05$, $t_{\nu,\alpha,табл} = 2,306$). $t_{\nu,\alpha,табл}$ можно найти либо в специальной таблице (приложение 1), либо воспользовавшись встроенной статистической функцией СТЬЮДРАСПОБР(вероятность;степени_свободы). В нашем случае это будет формула: =СТЬЮДРАСПОБР(D19;D20-2).
6. Сделаем вывод о наличии связи между исследуемыми величинами – так как $t_{расч} > t_{\nu,\alpha,табл}$ (7,12 > 2,306), то между переменными существует зависимость и найденный коэффициент корреляции значим.

Регрессионный анализ

Цель регрессионного анализа – определить количественные связи между зависимыми случайными величинами. Одна из этих величин полагается зависимой и называется откликом, другие – независимые, называются факторами. Для установления степени зависимости между

откликом и факторами используются вычисляемые величины ковариации и коэффициент корреляции. Если коэффициент корреляции по абсолютной величине близок к единице, то для построения зависимости используется линейная модель. Для других случаев используются более сложные нелинейные модели (например, полиномиальные и экспоненциальные). В данной работе изучим линейную модель.

Уравнение линейной регрессии имеет вид:

$$Y = a_1X_1 + a_2X_2 + \ldots + a_kX_k,$$

где $a_1, a_2, \ldots, ak$ – параметры, подлежащие определению методом наименьших квадратов (МНК).

Обычно находят первые два параметра, которые принято обозначать a и b. В этом случае уравнение линейной регрессии имеет вид $Y = a \cdot X + b$.

Коэффициенты a и b вычисляются следующим образом (формулы 3 – 4):

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}, \quad (3)$$

$$b = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}, \quad (4)$$

где i – номер измерения, x_i и y_i – значения переменных при i-том измерении, n – число измерений при моделировании системы.

В среде MS Excel для нахождения модели регрессии (т.е., фактически коэффициентов a и b) можно использовать несколько способов:

- использовать встроенную функцию ЛИНЕЙН;
- графический способ – построение линии тренда на диаграмме с показом уравнения регрессии;
- инструмент Регрессия из Пакета анализа;
- использовать встроенную функцию СУММКВРАЗН и инструмент Поиск решения;
- использовать встроенные функции НАКЛОН (вычисляет коэффициент a) и ОТРЕЗОК (вычисляет коэффициент b).

Построение регрессионной модели средствами Excel

Рассмотрим на примере первые три из перечисленных способов нахождения модели регрессии.

1-й способ. Функция ЛИНЕЙН.

В первом способе для получения коэффициентов a и b линейного уравнения регрессии $Y = a \cdot X + b$, описывающего зависимость количества привлеченных покупателей от затрат на рекламную раскрутку сайтов, воспользуемся статистической функцией ЛИНЕЙН. Для этого выделите две ячейки D16:E16 и выполните вставку функции ЛИНЕЙН с аргументами согласно рис. 5.

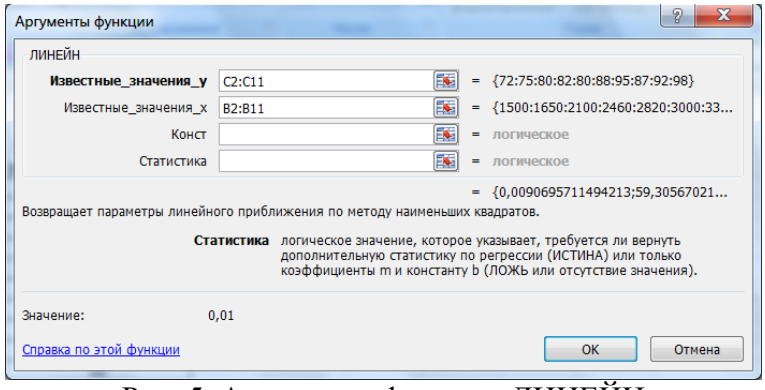


Рис. 5. Аргументы функции ЛИНЕЙН

Здесь *«Известные_значения_y»* – диапазон значений *«Количество покупателей»*, *«Известные_значения_x»* – диапазон значений *«Затраты на продвижение»*. Нажмите комбинацию клавиш SHIFT+CTRL+ENTER.

Получаем следующие значения коэффициентов регрессии – *a = 0,01* (ячейка D16), *b = 59,32* (ячейка E16). В ячейку D17 введем уравнение *Y = 0,01·X + 59,31*, чтобы продемонстрировать уравнение регрессии:

15	**Коэффициенты уравнения регрессии**		***a***	***b***
16			**0,01**	**59,31**
17	**Уравнение регрессии**	***Y=0,01·X+59,31***		

2-й способ (графический). Построение линии тренда

1. Для получения уравнения регрессии построим корреляционное поле переменных *X* (затраты на продвижение) и *Y* (количество покупателей).
2. Выделим диапазон ячеек B2:C11, запустим мастер диаграмм и выберем тип диаграммы – Точечная (в Excel 2007 выберем на панели инструментов «Вставка» кнопку «Точечная» и выберем подтип «Точечная с маркерами», после этого диаграмма будет создана и помещена на текущий лист, после чего ее можно будет дооформить). Задаем для диаграммы имя – *«Корреляционное поле»*, название оси *X* – *«Затраты на продвижение, руб.»*, оси *Y* – *«Количество покупателей»* (в Excel 2007 данные действия выполняются на вкладке «Макет» после выделения диаграммы – команды «Название диаграммы» и «Названия осей»). На последнем шаге мастера указываем место расположения – текущий лист.
3. Добавим линию тренда на точечный график (рис. 6). Для этого необходимо выделить диаграмму и выполнить команду меню «Диаграмма/Добавить линию тренда» (в Excel 2007 на вкладке «Макет» выберите команду «Анализ» и далее «Линия тренда» и «Линейное приближение»), либо выполнить данную команду из контекстного меню «Добавить линию тренда…», щелкнув по любой точке графика правой кнопкой мыши. Линия тренда – графическое представление направления изменения ряда данных.
4. Выбираем тип тренда «Линейный», который используется для аппроксимации данных по методу наименьших квадратов в соответствии с уравнением: *Y = a·X + b*, где *a* – угол наклона (в радианах) и *b* – координата пересечения оси абсцисс (оси Y).

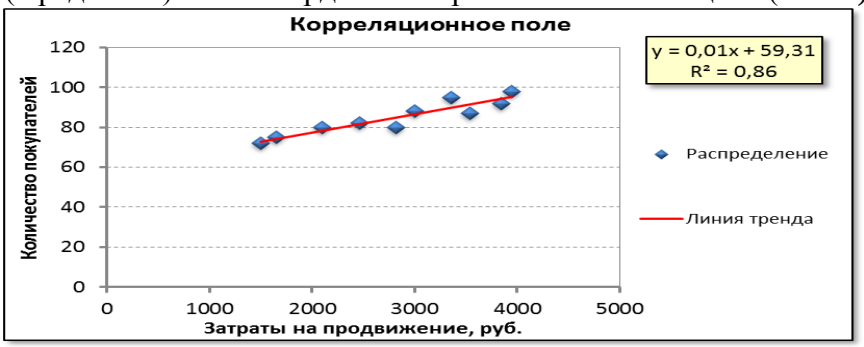


Рис. 6. Диаграмма с линией и уравнением тренда

5. На вкладке Параметры устанавливаем флажки «Показать уравнение на диаграмме» и «Поместить на диаграмму величину достоверности аппроксимации R^2». Щелкаем по кнопке ОК. Далее можно отформатировать эти уравнения, выделив их и в контекстном меню выбрав «Формат подписи линии тренда». R^2 – это число от 0 до 1, которое отражает близость линии тренда к фактическим данным. Линия тренда наиболее соответствует действительности, когда значение близко к 1.
6. Сравниваем уравнение регрессии, полученное графическим методом, с уравнением, рассчитанным с помощью функции ЛИНЕЙН. Как видим, эти уравнения одинаковые.

3-й способ. Инструмент анализа Регрессия.

1. Прежде чем мы начнем использовать этот инструмент, нужно убедится, что был активизирован Пакет анализа (меню «Сервис» есть команда «Анализ данных»). Если нет, то выполните команду «Сервис/Надстройки». В диалоговом окне «Надстройки» установите флажок «Пакет анализа» и щелкните по кнопке ОК (в Excel 2007 этот инструмент находится на вкладке «Данные» – «Анализ данных»).
2. Далее выполните команду «Сервис/Анализ данных». Выберите инструмент анализа «Регрессия» из списка «Инструменты анализа». Щелкните по кнопке ОК.

3. На экране появится диалоговое окно «Регрессия» (рис. 7):
   - в текстовом поле «Входной интервал Y» введите диапазон со значениями зависимой переменной $C$2:$C$211.
   - в текстовом поле «Входной интервал X» введите диапазон со значениями независимых переменных $B$2:$B$11.
   - Убедитесь, что в поле Уровень надежности введено 95% и переключатель «Параметры вывода» установлен в положении «Новый рабочий лист».
   - Щелкните по кнопке ОК.

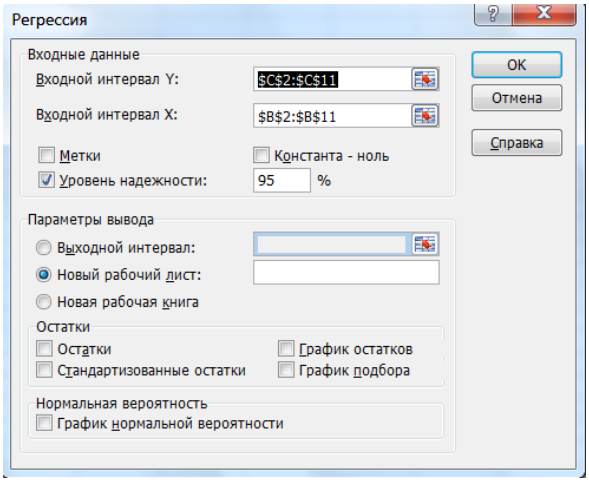


Рис. 7. Диалоговое окно инструмента анализа «Регрессия»

4. В результате на новом листе будет отображены результаты использования инструмента «Регрессия» (рис. 8).

	A	B	C	D	E	F	G	H	I
1	ВЫВОД ИТОГОВ								
2									
3	*Регрессионная статистика*								
4	Множественный R	0,93							
5	R-квадрат	0,86							
6	Нормированный R-квадрат	0,85							
7	Стандартная ошибка	3,35							
8	Наблюдения	10							
9									
10	Дисперсионный анализ								
11		*df*	*SS*	*MS*	*F*	*Значимость F*			
12	Регрессия	1	569,1337	569,1337	50,7214	0,0001			
13	Остаток	8	89,7663	11,2208					
14	Итого	9	658,9000						
15									
16		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
17	Y-пересечение	59,31	3,747	15,829	0,000	50,666	67,945	50,666	67,945
18	Переменная X 1	0,01	0,001	7,122	0,000	0,006	0,012	0,006	0,012

Рис. 8. Вывод итогов инструмента «Регрессия»

5. Среди полученных результатов после применения инструмента Регрессия есть столбец «Коэффициенты», содержащий значение *b* в строке «Y-пересечение», *a* – в строке «Переменная X1».
6. Сравним полученные результаты с ранее рассчитанными коэффициентами *a* и *b* – результаты полностью совпадают.
7. Следует обратить также внимание на следующие показатели:

а) Столбец «*df*» – число степеней свободы (используется при проверке адекватности модели по статистическим таблицам):
   - в строке «Регрессия» находится k_1 – количество коэффициентов уравнения, не считая свободного члена *b*;
   - в строке «Остаток» находится $k_2 = n - k_1 - 1$, где *n* – количество исходных данных.

б) Столбец «*SS*» (сумма квадратов):
   - в строке Регрессия: $SS_{reg} = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$, где $\hat{Y}_i$ – модельные значения Y, полученные путем подстановки значений X в построенную модель; $\bar{Y}$ – среднее значение Y;

– в строке Остаток: $SS_{resid} = \sum_{i=1}^{n}\left(\widehat{Y}_i - Y_i\right)^2$.

в) Столбец «*MS*» – вспомогательные величины:

– в строке Регрессия: $S_r^2 = SS_{reg}/k_1$;

– в строке Остаток: $S_b^2 = SS_{resid}/k_2$.

г) Столбец «*F*» – критерий Фишера. Используется для проверки адекватности модели: $F = S_r^2/S_b^2$.

д) Столбец «*Значимость F*» – оценка адекватности построенной модели. Находится по значениям F, и с помощью функции FРАСП. Если значимость F меньше 0,05, то модель может считаться адекватной с вероятностью 0,95.

е) «*Стандартная ошибка*», «*t-статистика*» – это вспомогательные величины, используемые для проверки значимости коэффициентов модели.

ж) «*P-Значение*» – оценка значимости коэффициентов модели. Если «*P-Значение*» меньше 0,05, то с вероятностью 0,95 можно считать, что соответствующий коэффициент модели значим (т.е. его нельзя считать равным нулю и Y значимо зависит от соответствующего X).

и) Нижние и верхние 95% – доверительные интервалы для коэффициентов модели.

Прогнозирование данных

Кроме нахождения уравнения регрессии, часто необходимо на основании этого уравнения предсказать теоретические значения *Y* при известных значениях *X*.

Это можно сделать тремя способами (рис. 9).

	A	B	C	D	E
1	Порядковый номер магазина	Затраты на продвижение (Xi), руб.	Количество покупателей, воспользовавшихся услугами магазина (Yi)		
2			Способ 1	Способ 2	Способ 3
3	1	1500	72	72	72
4	2	1650	75	75	75
5	3	2100	80	80	80
6	4	2460	82	82	82
7	5	2820	80	80	80
8	6	3000	88	88	88
9	7	3360	95	95	95
10	8	3540	87	87	87
11	9	3840	92	92	92
12	10	3950	98	98	98
13	**11**	**5000**			
14	**12**	**10000**			
15	**13**	**50000**			
16					
17	**Коэффициенты уравнения регрессии:**				
18	***a***	***b***			
19	**0,01**	**59,31**			

Рис. 9. Исходные данные для прогнозирования

1. *Способ 1*. Создать в Excel обычную формулу, основанную на уравнении регрессии $Y = a \cdot X + b$, типа C13=$A$19*B13+$B$19, где C13 – адрес ячейки с прогнозным значением функции *Y*, B13 – адрес ячейки со значением переменной *X*, для которого мы хотим спрогнозировать значение *Y*, $A$19 – абсолютный адрес ячейки со значением коэффициента *a*, $B$19 – абсолютный адрес ячейки со значением коэффициента *b*. В нашем случае нужно округлить до целого с помощью функции ОКРУГЛ($A$19*B13+$B$19;0). После чего скопируем формулу в ячейки C14 и C15.
2. *Способ 2*. Также можно вычислить теоретическое значение *Y* при *X* из ячейки B13 с помощью функции ПРЕДСКАЗ. Ее синтаксис – ПРЕДСКАЗ(X_i;<массив Y>;<массив X>). Аргумент X_i – это точка данных из массива X, для которой предсказывается теоретическое значение Y_i. Теоретическое значение в ячейке D13 вычислим по формуле = ПРЕДСКАЗ(B13;$D$3:$D$12;$B$3:$B$12). После чего скопируем формулу в ячейки D14 и D15.
3. *Способ 3*. Еще один способ прогнозирования – вычислить значения уравнения линейной регрессии Y для целого диапазона значений независимой переменной X с помощью функции ТЕНДЕНЦИЯ. Ее синтаксис – ТЕНДЕНЦИЯ(<массив Y>;<массив

X>;<новые значения X>;[<константа>]). Аргумент <новые значения X > – это массив значений X, для которых функция ТЕНДЕНЦИЯ возвращает соответствующие значения Y. Новые значения зависимой переменной вычислим в ячейках E13:B15 по формуле =ТЕНДЕНЦИЯ(E3:E12;B3:B12;B13:B15). **Важно** оформить эту функцию в ячейках E13:E15 как массив, для чего после ввода формулы в ячейку B12 нажать клавишу ENTER, выделить ячейки E13:E15, нажать клавишу F2, после этого нажать комбинацию клавиш SHIFT+CTRL+ENTER.

4. Сравним полученные результаты для всех трех способов (рис. 10). Видим, что все три способа дают одинаковые результаты, что не удивительно, так как во всех случаях используются линейная регрессия.

C13 =ОКРУГЛ($A$19*B13+$B$19;0)

	A	B	C	D	E
1	Порядковый номер магазина	Затраты на продвижение (Xi), руб.	Количество покупателей, воспользовавшихся услугами магазина (Yi)		
2			Способ 1	Способ 2	Способ 3
3	1	1500	72	72	72
4	2	1650	75	75	75
5	3	2100	80	80	80
6	4	2460	82	82	82
7	5	2820	80	80	80
8	6	3000	88	88	88
9	7	3360	95	95	95
10	8	3540	87	87	87
11	9	3840	92	92	92
12	10	3950	98	98	98
13	**11**	**5000**	**105**	**105**	**105**
14	**12**	**10000**	**150**	**150**	**150**
15	**13**	**50000**	**513**	**513**	**513**
16					
17	**Коэффициенты уравнения регрессии:**		=ОКРУГЛ(ПРЕДСКАЗ(B13;$D$3:$D$12;$B$3:$B$12);0)		
18	***a***	***b***			
19	**0,01**	**59,31**			
20			{=ОКРУГЛ(ТЕНДЕНЦИЯ(E3:E12;B3:B12;B13:B15);0)}		

Рис. 10. Результаты прогнозирования тремя способами

1.2.1. Введение в теорию факторного планирования эксперимента

Если необходимо изучить влияние, например, количества углерода X на прочность стали Y проводят однофакторный эксперимент. И чем больше различных значений примет X, тем более полно мы узнаем изучаемую зависимость Y(X).

Допустим, что исследуем влияние на прочность стали Y количества углерода X_1 и количества хрома X_2. Последовательно проведя 2 серии однофакторных экспериментов получим всего лишь 2 линии на двумерном экспериментальном поле – основная область возможных сочетаний факторов останется неисследованной (рис. 19).

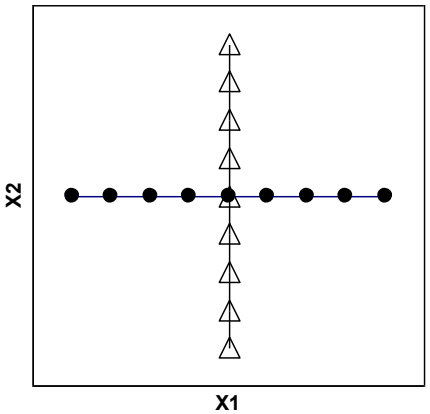


Рис. 19. Схема эксперимента «крест»

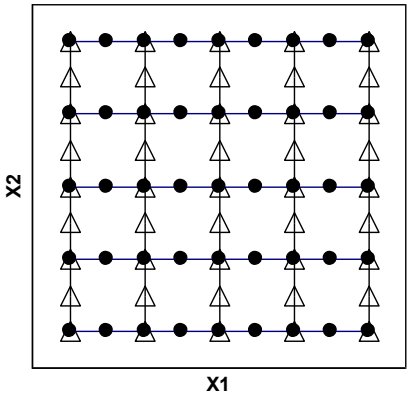


Рис. 20. Схема эксперимента «решетка»

Попытка «заштриховать» всё поле эксперимента экспериментальными линиями (рис. 20) приведет к недопустимо высоким затратам по времени и по средствам. На рисунках 19 и 20 треугольниками обозначены серии с варьированием X_2 при постоянном X_1, точками – серии с варьированием X_1 при постоянном X_2.

Решение – провести отдельные эксперименты в точках, расположенных на границах, в углах и в центре исследуемой области. Это пример факторного планирования эксперимента (рис. 21).

Факторное планирование эксперимента имеет цель: за минимальное количество экспериментов описать исследуемую область с достаточной для экспериментатора точностью.

Факторный эксперимент – мощное средство эмпирического изучения процессов, обеспечивающее точное математическое описание отклика системы при минимальном количестве экспериментов.

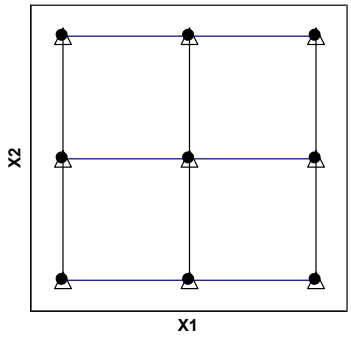


Рис. 21. Схема полного факторного эксперимента (ПФЭ) 3^2

Не приводя строгих определений терминов, связанных с факторным планированием, опишем их упрощенно.

Фактор, X – величина, которую экспериментатор меняет (варьирует).

Отклик Y – величина, которую экспериментатор измеряет.

Факторное пространство – служит для мысленного расположения в нем экспериментальных точек. Количество измерений равно количеству факторов.

План полного факторного эксперимента ПФЭ обозначается m^k где m – число уровней варьирования факторов, k – число факторов. Например, если 3 фактора варьируются на 2-х уровнях, то план ПФЭ обозначится 2^3 и будет состоять из 8 опытов на различных сочетаниях факторов. Очевидно, что план 3^2 состоит из 9 опытов. Планы 2^k называют планами первого порядка, планы 3^k – планы второго порядка. Планы больших порядков используют редко – для повышения точности выгоднее повторить эксперимент, сузив диапазоны варьирования.

Пример

Гипотеза: Чтобы корова меньше ела и давала больше молока – ее надо меньше кормить и чаще доить.

1. Постановка задачи.

Корова (рис. 22) представляет собой систему. Система имеет на входе контролируемые воздействия (варьируемые факторы) X_1 и X_2 и неконтролируемые воздействия (случайные факторы), например X_3, X_4, … . Случайные факторы не учитываем – полагаем систему детерминированной. Из выходных характеристик системы, Y, Y_1, Y_2, … в соответствии с целями исследования для контроля выбираем Y.

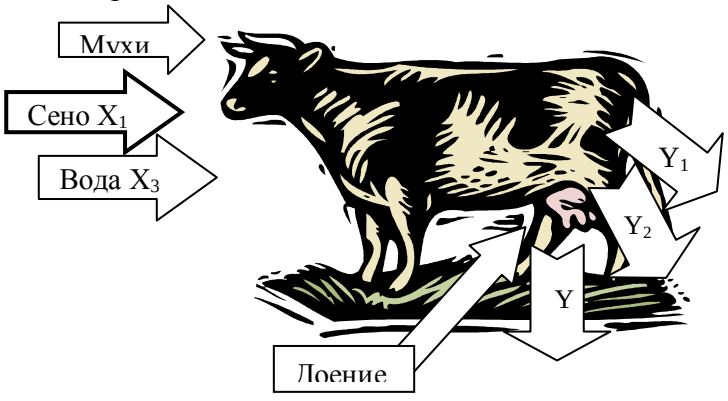


Рис. 22. Исследуемая система

В рамках принятой модели (рис. 23) исследуем зависимость количества молока в сутки Y от количества корма X_1 и числа доений X_2.

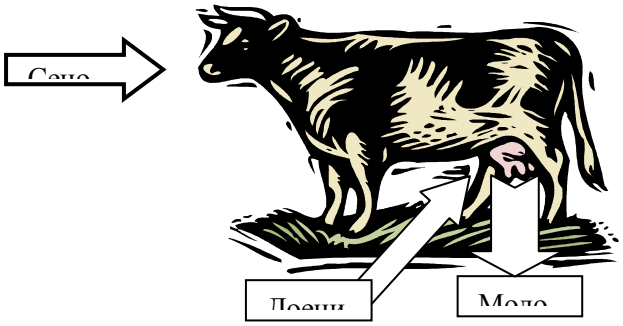


Рис. 23. Принятая модель исследуемой системы

2. Планирование и обработка результатов эксперимента

План эксперимента 2^2.

Технически возможные пределы изменения факторов. Количество корма от 0 до 100 кг. Количество доений от 1 до 10.

Пределы варьирования факторов не должны превышать технически возможных и выбираются на усмотрение экспериментатора. С учетом гуманного отношения к животным принимаем пределы варьирования: $X_1 = 10…70$ кг, $X_2 = 2…5$ шт. (табл. 10).

В планах первого порядка два уровня варьирования факторов, верхний, обозначаемый «+» или «1» и нижний, обозначаемый «–» или «–1». При этом от натуральных значений факторов (X) переходят к кодированным (x) и оформляют в виде таблицы.

Следующая таблица (табл. 11) – матрица эксперимента – состоит из уровней варьирования факторов, взаимодействий и отклика.

Столбцы взаимодействий получаются перемножением соответствующих кодированных значений факторов.

Таблица 10.

Уровни и интервалы варьирования факторов

Факторы	**д. изм.**	**Кодовое обозначение**	**Интервал варьирования**	**Натуральные уровни соответствующие кодированным**	
				1	–1
X_1 – количество корма	г	x_1	60	70	10
X_2 – число доений	т.	x_2	3	5	2

Таблица 11.

Матрица эксперимента

№ опыта	x_1	x_2	x_1x_2	**Y**
1	1	1	1	12
2	–1	1	–1	2
3	1	–1	–1	17
4	–1	–1	1	6

Для описания результатов планов первого порядка используют полиномы первого порядка, в данном случае:

$$Y=a_0+a_1x_1+a_2x_2+a_{12}x_1x_2\,. \qquad (5)$$

Коэффициенты при кодированных факторах дают информацию о влиянии факторов или из сочетаний на отклик.

В данном случае применив надстройку Excel «Поиск решения» получаем:

a_0	a_1	a_2	a_{12}
9,249997	5,249999	–2,25	–0,25

Это говорит о том, что повышение количества корма сильно влияет на количество молока, а вот повышения частоты доений – немного снижает количество молока, совместное влияние факторов не выражено.

Таким образом – чтобы было больше молока, корову надо больше кормить и реже доить.

3. Анализ полученных результатов

Вывод несколько противоречит сложившимся представлениям – если корову совсем не доить то, скорее всего, молока не будет, да и бесконечно увеличивать кормление тоже нецелесообразно.

Предполагаем, что существует оптимальное количество корма и числа доений, соответствующее максимуму молока. Проведем уточняющий эксперимент 3^2 сузив диапазоны варьирования и перейдя в предполагаемую оптимальную область. Поместим центр плана в точку с $x_1 = 1$ ($X_1 = 70$ кг) и $x_2 = -1$ ($X_2 = 2$ шт.).

Таблица 12.

Уровни и интервалы варьирования факторов

Факторы	**д. изм.**	**Кодовое обозначение**	**Интервал варьирования**	**Натуральные уровни соответствующие кодированным**		
				1		–1
X_1 – количество корма	г	x_1	20	90	0	50
X_2 – число доений	т.	x_2	1	3		1

Кодированные значения факторов x связаны с натуральными X через диапазон варьирования e и натуральное значение фактора в центре плана X_0:

$$x = \frac{X - X_0}{e} \qquad (6)$$

В факторном планировании часто при переходе от одного плана к другому стараются использовать данные предыдущего плана – в данном случае опыт 5 плана второго порядка можно не проводить, т.к. он соответствует опыту 3 плана второго порядка.

Для описания результатов плана второго порядка применяют полиномы второй степени, в данном случае:

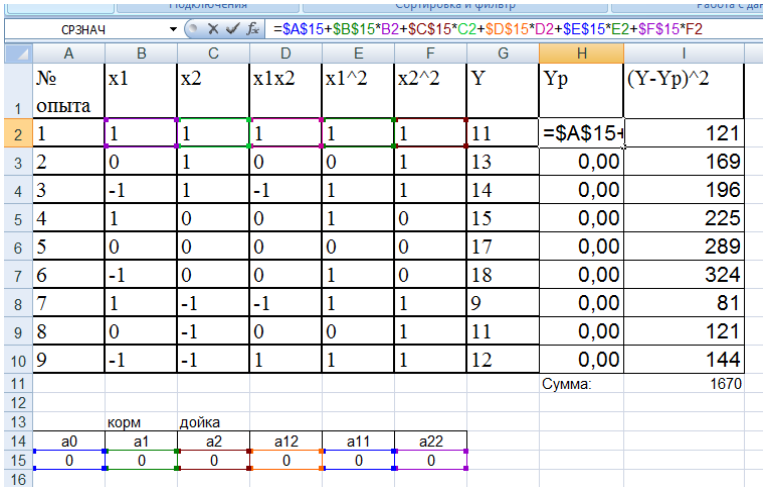
$$Y=a_0+a_1x_1+a_2x_2+a_{12}x_1x_2+a_{11}x_1^2+a_{22}x_2^2 \qquad (7)$$

Таблица 13.

Матрица эксперимента

№ опыта	x_1	x_2	x_1x_2	x_1^2	x_2^2	**Y**
1	1	1	1	1	1	11
2	0	1	0	0	1	13
3	–1	1	–1	1	1	14
4	1	0	0	1	0	15
5	0	0	0	0	0	17
6	–1	0	0	1	0	18
7	1	–1	–1	1	1	9
8	0	–1	0	0	1	11
9	–1	–1	1	1	1	12

Построим в Excel таблицу на основе таблицы 13. Справа от столбца Y с экспериментальными данными располагаем столбец Y_p где с помощью формул строим выражение (7) положив в качестве начальных значений всех коэффициентов ноль (для этого нужно для каждого коэффициента выбрать ячейку в Excel и записать в нее «0» – рис. 24).



	A	B	C	D	E	F	G	H	I
1	№ опыта	x1	x2	x1x2	x1^2	x2^2	Y	Yp	(Y-Yp)^2
2	1	1	1	1	1	1	11	=$A$15+	121
3	2	0	1	0	0	1	13	0,00	169
4	3	-1	1	-1	1	1	14	0,00	196
5	4	1	0	0	1	0	15	0,00	225
6	5	0	0	0	0	0	17	0,00	289
7	6	-1	0	0	1	0	18	0,00	324
8	7	1	-1	-1	1	1	9	0,00	81
9	8	0	-1	0	0	1	11	0,00	121
10	9	-1	-1	1	1	1	12	0,00	144
11								Сумма:	1670
12									
13		корм	дойка						
14	a0	a1	a2	a12	a11	a22			
15	0	0	0	0	0	0			

Рис. 24 Подготовка таблицы эксперимента для применения надстройки «Поиск решения»

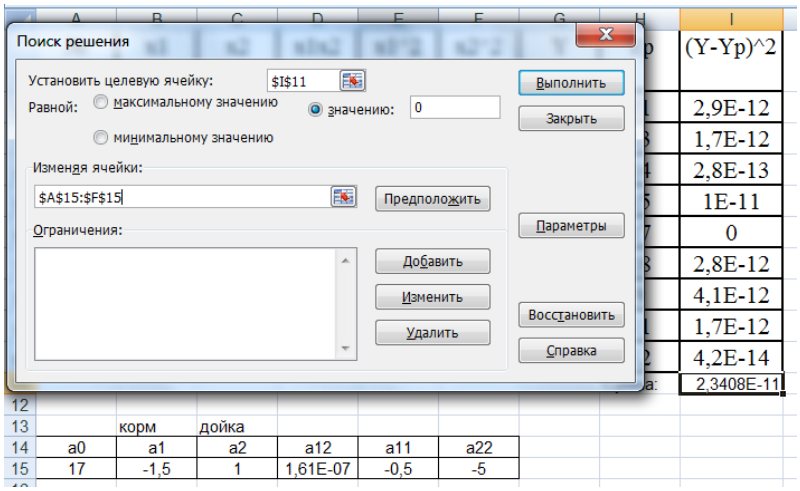


Рис. 25. Применение надстройки «Поиск решения»

Применив надстройку «Поиск решения», положив значение целевой ячейки «0» (см. рис. 25), получаем:

a_0	a_1	a_2	a_{12}	a_{11}	a_{22}
17	–1,5	1	1,61E-07	–0,5	–5

Разные знаки при квадратичных и линейных коэффициентах указывают, что возможно, оптимум лежит внутри исследованной области.

Найти искомый оптимум можно с использованием надстройки «Поиск решения» положив в качестве целевой функции (7) с найденными коэффициентами и изменяя ячейки x_1 и x_2.

Результат:

x_1	x_2
–1,49999	0,1

Перейдя от кодированных значений к натуральным по (6) получаем, что максимальный суточный надой 18,2 л возможен при 2-х разовом доении и кормлении в объеме 40 кг в день. Оптимум по сену лежит за пределами диапазона варьирования и, строго говоря, нуждается в дополнительной экспериментальной проверке.

Вопросы статистической обработки при планировании и обработке результатов факторного эксперимента в данном примере не рассмотрены – для лучшего понимания основных принципов факторного планирования.

Требуемый результат исследования – оптимальное сочетание факторов – достигнут за 12 опытов.

Контрольные вопросы

1. В чем цель корреляционного анализа?

1. Что такое коэффициент корреляции?
2. Для чего используется t-статистика Стьюдента?
3. Какими способами можно определить коэффициент корреляции в MS Excel?
4. В чем цель регрессионного анализа?
5. Опишите уравнение линейной регрессии.
6. Какими способами можно найти модель регрессии в MS Excel? Коротко опишите эти способы.
7. В чем задача прогнозирования данных?
8. Какими способами осуществить прогнозирование в MS Excel?
9. Что обозначает план эксперимента 34?
10. Как подключить надстройку «Поиск решения»?
11. Для чего выполняют кодирование переменных при планировании и обработке результатов эксперимента?
12. Что такое «целевая ячейка»?
13. Для чего используются относительные, абсолютные и смешанные ссылки в формулах?
14. Полный факторный план какого порядка целесообразно применить при 8 факторном эксперименте?
15. Чем отличаются уравнения регрессии в описании планов первого и второго порядка?
16. Для чего выполняется переход от натуральных размерных значений факторов к кодированным безразмерным?

1.2. Порядок выполнения задания

1. Перед выполнением задания_1 изучить теоретическую часть практикума (1.1) и ответить на контрольные вопросы.
2. Открыть новую книгу Excel и сохранить под именем «Статфункции.xls».
3. В книге выполнить задание со следующими условиями:
   Имеются данные по двум экономическим показателям X и Y:

Цена (X)	995	983	1001	1012	1011	1017	978	997	1010	989	900	1100	5000
Спрос (Y)	122	144	114	100	100	90	150	130	95	155	?	?	?

   Необходимо:
   - вычислить коэффициент корреляции;
   - построить корреляционное поле (диаграмму) на отдельном листе;
   - построить регрессионную модель (с использованием функции ЛИНЕЙН);
   - спрогнозировать значение Y для 3-х новых значений X с помощью функции ПРЕДСКАЗ.

   Все действия (в том числе форматирование таблицы) необходимо выполнять, опираясь на образец.
4. На диаграмме разместить линию тренда с уравнением регрессии и оформить их как показано в образце. Дополнить диаграмму спрогнозированными данными (кроме последнего значения цены 5000).
5. Используя инструмент «Регрессия» на отдельном листе построить регрессионную модель с учетом новых спрогнозированных значений. Записать на листе уравнение регрессии на основании данных из «Вывода итогов».
6. Представить файл с выполненной работой преподавателю для проверки.
7. Перед выполнением задания_2 изучите теоретическую часть работы (1.2) и ответьте на контрольные вопросы.
8. Создать книгу MS Excel с названием «Корова». Лист 1 озаглавить «2-2» и воспроизвести на нем пример плана первого порядка. Лист 2 озаглавить «3-2» и воспроизвести на нем пример плана второго порядка и поиск оптимальных значений. Скопировать лист «2-2» на новый лист «доить-не-кормить» и путем подбора результатов эксперимента подтвердить проверяемую гипотезу: «Чтобы корова меньше ела и давала больше молока ее надо меньше кормить и чаще доить».
9. В п.3 примера принято решение, в результате которого для достижения цели понадобилось 12 опытов. Предложить вариант решения с достижением цели за 9

опытов. Создать книгу «Корова2» и провести в ней расчеты по новому варианту с описанием проводимых действий в отчете по работе по аналогии с примером.

10. Оформить отчет по работе, содержащий: цель работы, описание действий, выводы по работе.

1.3. Требования к оформлению, процедура защиты

Отчет по данной работе должен содержать распечатку каждого листа книги «Статфункции.xls». При защите необходимо дать требуемые пояснения к содержанию каждого листа книги, продемонстрировать выполнение работы в файле книги «Статфункции.xls» и ответить на контрольный вопрос.