МИНИСТЕРСТВО СЕЛЬСКОГО ХОЗЯЙСТВА РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«БАШКИРСКИЙ ГОСУДАРСТВЕННЫЙ АГРАРНЫЙ УНИВЕРСИТЕТ»

Кафедра безопасности жизнедеятельности
и технологического оборудования

Оптимизация технологических процессов общественного питания

МЕТОДИЧЕСКИЕ УКАЗАНИЯ

к лабораторной работе

КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ В EXCEL

Направление подготовки магистра

2.19.04.04 Технология продукции и организация общественного питания

Квалификация выпускника
магистр

УФА 2018

Рекомендовано к изданию методической комиссией факультета пищевых технологий (протокол №9 от 29.03.2018 г.)

Составитель: докт.техн.наук, профессор Мартынов В.М.

Ответственная за выпуск: заведующий кафедрой безопасности жизнедеятельности и технологического оборудования, к.б.н. Латыпова Г.Ф.

Г.Уфа, ФГБОУ ВО Башкирский ГАУ, кафедра БЖД и ТО

Цель работы. Освоение метода корреляционно-регрессионного анализа в MS Excel.

Задача работы. Найти уравнение регрессии, произвести его статистический анализ.

Корреляционно-регрессионный анализ

Этот анализ содержит две свои составляющие части. **Корреляционный анализ** – это количественный метод определения тесноты и направления взаимосвязи между выборочными переменными величинами. **Регрессионный анализ** – показывает влияние независимых переменных на зависимую переменную.

Регрессия одной переменной чаще всего бывает:

- линейной $y = a + bx$;
- параболической $y = a + bx + cx^2$;
- экспоненциальной $y = a \cdot exp(bx)$;
- степенной $y = a \cdot x^b$;
- гиперболической $y = a + b/x$;
- логарифмической $y = a + b \cdot ln(x)$;
- показательной $y = a \cdot b^x$.

Рассмотрим пример построения регрессионной модели в MS Excel и интерпретацию результатов.

Задача. На шести предприятиях общественного питания была проанализирована среднемесячная заработная плата и процент уволившихся в течение полугодия сотрудников. Необходимо определить зависимость числа уволившихся сотрудников от их средней зарплаты.

	A	B	C
1	Номер предприятия	Количество уволившихся, %	Зарплата, тыс. руб.
2		Y	X
3	1	6	10
4	2	3,5	15
5	3	2,2	20
6	4	2	25
7	5	1,5	30
8	6	1,5	35

Первоначально возьмем линейный тип регрессии. В общем случае он имеет вид:

$$y = b_0 + b_1x_1 + \ldots + b_kx_f.$$

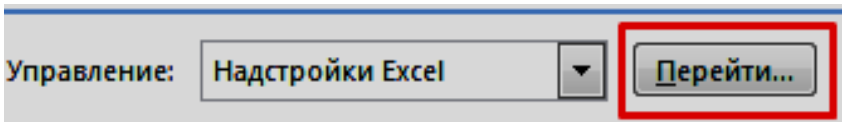
где b – коэффициенты регрессии, x – влияющие переменные (факторы), f – число независимых переменных (факторов).

В нашем примере в качестве отклика y выступает количество уволившихся сотрудников, а в качестве фактора – одна переменная x (заработная плата).

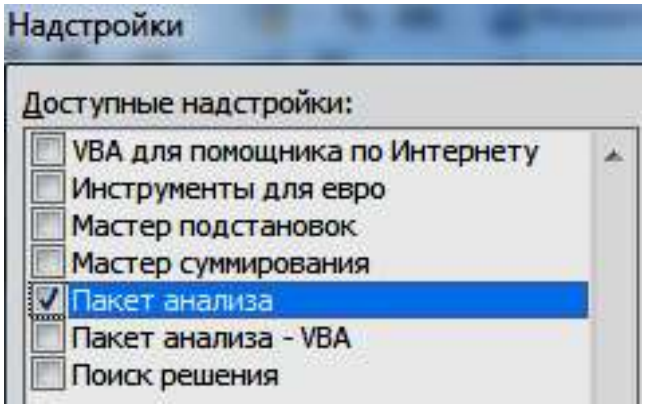
В Excel существуют встроенная функция «**ЛИНЕЙН**», с помощью которой можно рассчитать параметры модели линейной регрессии. Но быстрее это сделает надстройка **«Пакет анализа».** Для этого её нужно активировать:

1. Нажимаем кнопку «**Офис**» и переходим на вкладку «**Параметры Excel**» и выбираем «**Надстройки**».

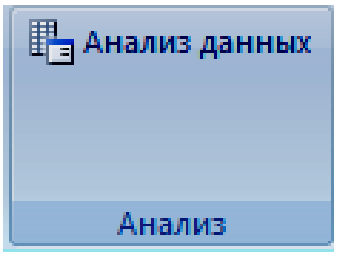
2. Внизу, под выпадающим списком, в поле «**Управление**» будет надпись «**Надстройки Excel**» (если ее нет, нажмите на флажок справа и выберите). Нажимаем кнопку «**Перейти**».



3. Открывается список доступных надстроек. Выбираем «**Пакет анализа**» и нажимаем **ОК**.

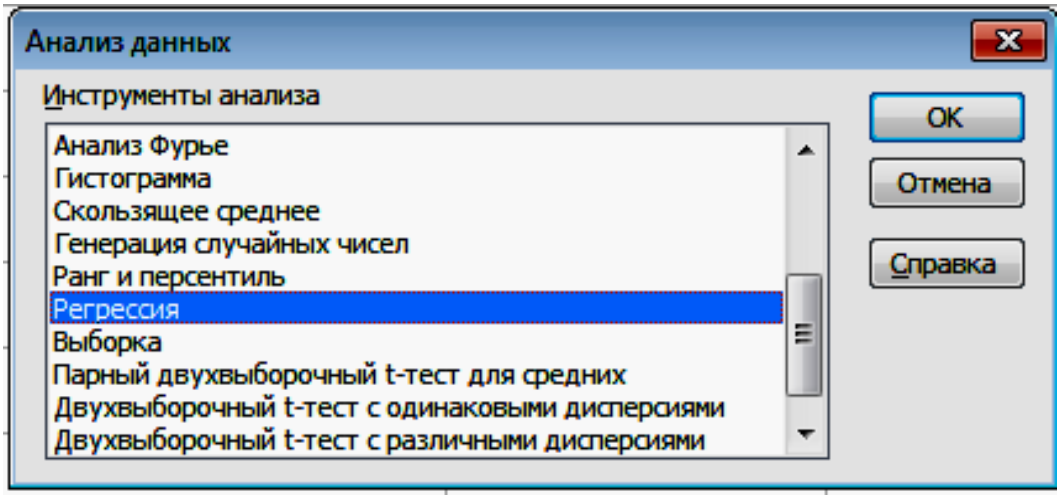


После активации надстройка будет доступна на вкладке «**Данные**».

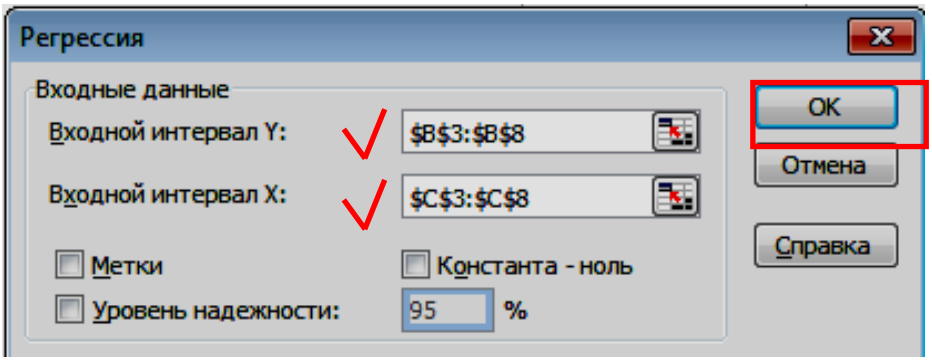


Теперь займемся непосредственно регрессионным анализом.

1. Открываем меню инструмента «**Анализ данных**». Выбираем «**Регрессия**».



2. Откроется меню для выбора входных значений и параметров вывода (где отобразить результат). В полях для исходных данных указываем диапазон описываемого параметра (Y) и влияющего на него фактора (X).



Также в этом диалоговом окне задаются следующие параметры:

• **Метки** – отмечается только тогда, когда первые ячейки содержат пояснительный текст (подписи данных). В примере метки не используются;

• **Уровень надежности** – указывается желаемый уровень надежности (обычно флажок активизируется, что означает уровень равный 95%);

• **Константа-ноль** – установите данный флажок в активное состояние, если требуется, чтобы линия регрессии прошла через начало координат;

• **Параметры вывода** – определяют, куда должны быть помещены результаты. По умолчанию стоит режим «**Новый рабочий лист**». Можно указать «**Выходной интервал**», для этого вводятся номера ячеек для помещения результатов (достаточно указать левую верхнюю ячейку будущего диапазона). Если требуется создания новой книги, в которой результаты будут добавлены в новый лист, установите переключатель в положение «**Новая рабочая книга**»;

• **Остатки** – активизировать флажок;

• **Стандартизованные остатки** – активизировать флажок;

• **График остатков** – команда не обязательна;

• **График подбора** – команда не обязательна;

• **График нормальной вероятности** – команда не обязательна.

3. После нажатия **ОК**, программа отобразит расчеты в ранее указанном месте.

	A	B	C	D	E	F	G	H	I
1	ВЫВОД ИТОГОВ								
2									
3	*Регрессионная статистика*								
4	Множественный R	0,882520084							
5	R-квадрат	0,778841699							
6	Нормированный R-квадрат	0,723552124							
7	Стандартная ошибка	0,913965718							
8	Наблюдения	6							
9									
10	Дисперсионный анализ								
11		*df*	*SS*	*MS*	*F*	*Значимость F*			
12	Регрессия	1	11,767	11,767	14,08659218	0,019891595			
13	Остаток	4	3,341333333	0,835333333					
14	Итого	5	15,10833333						
15									
16		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
17	Y-пересечение	6,473333333	1,051580821	6,155811524	0,003533637	3,553676911	9,392989755	3,553676911	9,392989755
18	Переменная X 1	-0,164	0,043695919	-3,753210916	0,019891595	-0,28531932	-0,04268068	-0,28531932	-0,04268068
19									
20									
21									
22	ВЫВОД ОСТАТКА								
23									
24	*Наблюдение*	*Предсказанное Y*	*Остатки*	*Стандартные остатки*					
25	1	4,833333333	1,166666667	1,427157454					
26	2	4,013333333	-0,513333333	-0,62794928					
27	3	3,193333333	-0,993333333	-1,215122632					
28	4	2,373333333	-0,373333333	-0,456690385					
29	5	1,553333333	-0,053333333	-0,065241484					
30	6	0,733333333	0,766666667	0,937846327					

В первую очередь обращаем внимание на множественный коэффициент корреляции R (0,883) и численное значение «**Значимость F**», определяющее значимость R и значимость влияния всех переменных X на отклик Y. «**Значимость F**» менее 0,05 свидетельствует о существенной корреляции между зависимой переменной Y и всеми независимыми переменными X.

«**R-квадрат**» – коэффициент детерминации. В нашем примере – 0,779, т.е. расчетные параметры модели на 77,9% объясняют зависимость между изучаемыми параметрами. Чем выше коэффициент детерминации, тем качественнее модель. Хорошо – выше 0,8, плохо – меньше 0,5 (такой анализ вряд ли можно считать резонным). В нашем примере – «неплохо».

Ввод в математическую модель дополнительных факторных переменных приводит к увеличению значения «**R-квадрат**», однако это не всегда может свидетельствовать об улучшении модели. «**Нормированный R-квадрат**» – это тот же коэффициент детерминации, но скорректированный на величину выборки в случае ее малого размера. Он вычисляется по формуле

$$\overline{R}^2 = 1 - \left(1 - R^2\right)\left(m - 1\right)/\left(m - k\right),$$

где m – общее число наблюдений;

k – число коэффициентов регрессии, включая свободный член.

Коэффициент регрессии 6,4733 (свободный член) показывает, каким будет Y, если все переменные в рассматриваемой модели будут равны 0. Коэффициент регрессии -0,164 показывает весомость переменной X на Y. Знак «–» указывает на отрицательное влияние: чем больше зарплата, тем меньше уволившихся, что справедливо.

В столбце «**Стандартная ошибка**» вычислены среднеквадратические отклонения коэффициентов регрессии.

Значимость коэффициентов регрессии устанавливается с помощью критерия Стьюдента «**t-статистика**» и оценивается вероятностью «**P-Значение**». Если она меньше принятого уровня значимости 0,05, то с вероятностью 0,95 можно считать, что соответствующий коэффициент регрессии модели значим (т.е. его нельзя считать равным нулю, и Y значимо зависит от переменной X). Так как в нашем примере всего одна переменная X, то «**P-Значение**» для X1 равняется вероятности «**Значимость F**».

«**Нижние 95%**» и «**Верхние 95%**» – это нижние и верхние границы 95-процентных доверительных интервалов для коэффициентов регрессии. Если в блоке ввода данных значение доверительной вероятности было оставлено по умолчанию, то последние два столбца будут дублировать предыдущие. Если задать другое значение доверительной вероятности, то последние два столбца содержат значения нижней и верхней границы для указанной доверительной вероятности.

В столбце «**Предсказанное Y**» вычисляются значения функции регрессии $Y_i = f(X_i)$ для тех значений переменных X, которым соответствует порядковый номер i в столбце «**Наблюдение**». В столбце «**Остатки**» содержатся разности между заданным и расчетным (предсказанным уравнением регрессии) значениями Y.

Для построения корреляционного поля зависимости y от x в командной

строке выбираем меню **Вставка** / **Диаграмма**. В появившемся диалоговом окне выбираем тип диаграммы: **Точечная**; вид: **Точечная диаграмма**, позволяющая сравнить пары значений x и y (рисунок 1).

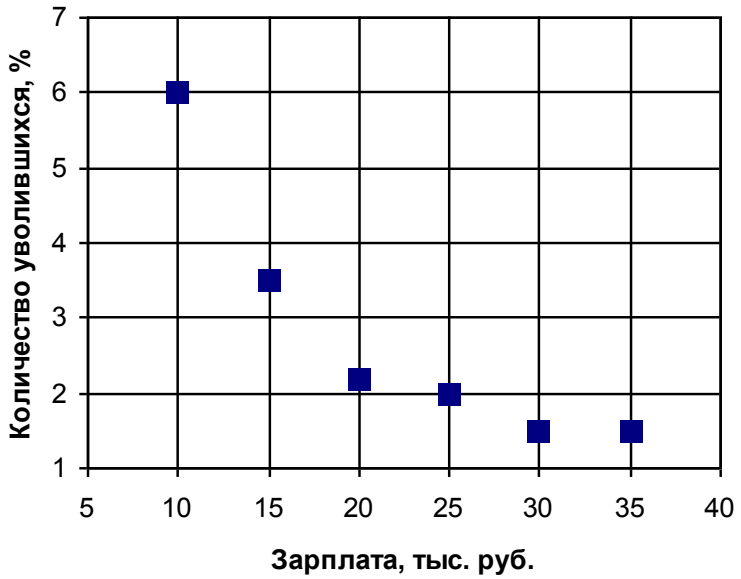


Рисунок 1 – Корреляционное поле

Учитывая, что R и R-квадрат при небольшом числе наблюдений, равном 6, в нашем примере имеют не столь высокие значения, а также с учетом вида графической зависимости Y = F(X), представленной на рисунке 9, применим в качестве математической модели уравнение регрессии второго порядка

$$y = b_0 + b_1 x + b_2 x^2,$$

которое эквивалентно

$$y = b_0 + b_1 x_1 + b_2 x_2.$$

Для этого в таблицу исходных данных добавим еще один столбец X^2, образованный возведением в квадрат численных значений столбца переменной X.

	A	B	C	D
1	Номер предприятия	Количество уволившихся, %	Зарплата, тыс. руб.	
2		Y	X	X^2
3	1	6	10	100
4	2	3,5	15	225
5	3	2,2	20	400
6	4	2	25	625
7	5	1,5	30	900
8	6	1,5	35	1225

Тогда вновь обращаемся к меню инструмента «**Анализ данных**» и выбираем «**Регрессия**». Увеличиваем интервал фактора X на один столбец ($C$3:$D$8) и переходим к решению. В результате получаем близкие к 1 значения R и R-квадрат: 0,986 и 0,973, а также значимые коэффициенты регрессии при X: -0,669 и 0,0112, для которых «**P-Значение**» меньше принятого уровня значимости 0,05. «**Значимость F**» менее **0,05** свидетельствует о существенной корреляции между зависимой переменной Y и всеми независимыми переменными X (в нашем примере x и x^2).

	A	B	C	D	E	F	G	H	I
1	ВЫВОД ИТОГОВ								
2									
3	*Регрессионная статистика*								
4	Множественный R	0,986441054							
5	R-квадрат	0,973065952							
6	Нормированный R-квадрат	0,95510992							
7	Стандартная ошибка	0,36829724							
8	Наблюдения	6							
9									
10	Дисперсионный анализ								
11		*df*	*SS*	*MS*	*F*	*Значимость F*			
12	Регрессия	2	14,70140476	7,350702381	54,19159207	0,004420307			
13	Остаток	3	0,406928571	0,135642857					
14	Итого	5	15,10833333						
15									
16		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
17	Y-пересечение	11,33285714	1,127460574	10,05166602	0,0020965	7,744774404	14,92093988	7,744774404	14,92093988
18	Переменная X 1	-0,668642857	0,10991771	-6,083122134	0,008920213	-1,018450068	-0,318835646	-1,018450068	-0,318835646
19	Переменная X 2	0,011214286	0,002411071	4,651162842	0,018743668	0,00354118	0,018887391	0,00354118	0,018887391
20									
21									
22									
23	ВЫВОД ОСТАТКА								
24									
25	*Наблюдение*	*Предсказанное Y*	*Остатки*	*Стандартные остатки*					
26	1	5,767857143	0,232142857	0,813731701					
27	2	3,826428571	-0,326428571	-1,144231962					
28	3	2,445714286	-0,245714286	-0,861303709					
29	4	1,625714286	0,374285714	1,311985882					
30	5	1,366428571	0,133571429	0,468208702					
31	6	1,667857143	-0,167857143	-0,588390615					

Замечание. Регрессионный анализ многофакторной математической модели не имеет принципиальных отличий от рассмотренного примера с двумя переменными X1 = x, X2 = x^2, в котором математическая модель по существу является многофакторной.

Корреляционный анализ в Excel

Корреляционный анализ состоит в определении степени связи между двумя случайными величинами, допустим, X и Y (например, между сроком службы и стоимостью оборудования, ценой блюд и их спросом и т.д.) 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. 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.

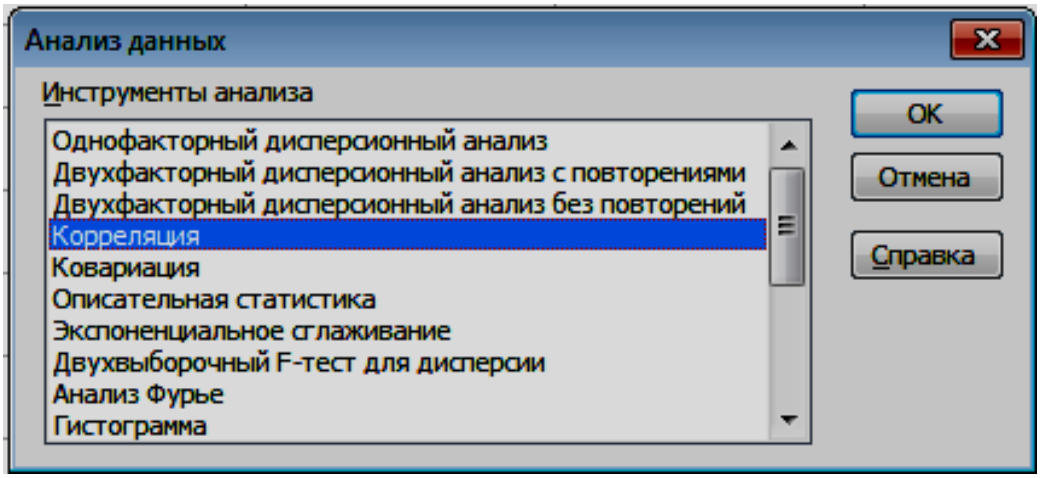
Коэффициент корреляции обозначается r и изменяется от -1 (строгая обратная линейная зависимость) до 1 (строгая прямая пропорциональная зависимость). При значении 0 линейной зависимости между двумя выборками нет.

Можно придерживаться следующей общей классификации корреляционных связей:

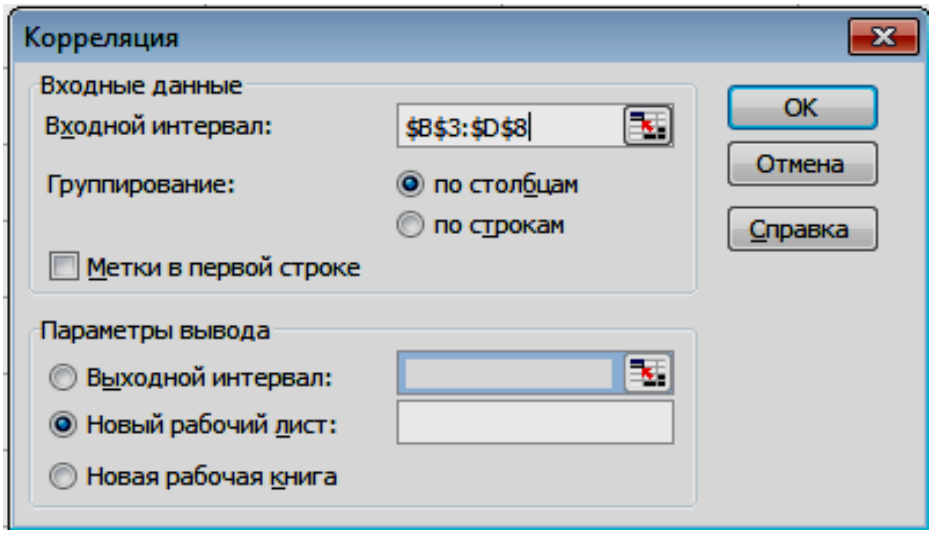
- сильная, или тесная при коэффициенте корреляции $|r| \geq 0{,}80$;
- средняя при $0{,}60 \leq |r| < 0{,}80$;
- умеренная при $0{,}40 \leq |r| < 0{,}60$;
- слабая при $0{,}20 \leq |r| < 0{,}40$;
- очень слабая при $|r| < 0{,}20$.

Рассмотрим, как с помощью MS Excel найти коэффициент корреляции.

Для нахождения парных коэффициентов корреляции применяется функция «**КОРРЕЛ**», но проще это сделать с помощью инструмента **«Анализ данных»**. В диалоговом окне «**Анализ данных**» выбираем «**Корреляция**».



После нажатия **ОК** в появившемся диалоговом окне указываем входной интервал (для последнего примера $B$3:$D$8 с группированием в нашем случае по столбцам) и параметры вывода.



Результат расчетов отобразится в корреляционной матрице:

	A	B	C	D
1		*Столбец 1*	*Столбец 2*	*Столбец 3*
2	Столбец 1	1		
3	Столбец 2	-0,8825201	1	
4	Столбец 3	-0,8005249	0,98708582	1

Между переменной Y (столбец 1) и переменной X1 (столбец 2, которому соответствует переменная x) коэффициент корреляции составляет -0,8825, между Y и X2 (столбец 3, которому соответствует переменная x^2) – -0.8005, а между X1 и X2 – 0,987. То есть две независимые переменные x и x^2 сильно коррелированны между собой и возникает вопрос об обоснованности включения в математическую модель переменной x^2. Однако если же произвести преобразование первообразной независимой переменной X1 (x) путем центрирования (вычитания из нее средней арифметической 22,5) или кодирования по формуле

$$x = (X0 - X0_0)/I,$$

где x – кодированное значение переменной;

$X0$ – натуральное значение переменной;

$X0_0$ – натуральное значение переменной на нулевом уровне (в центре эксперимента), для нашего примера 22,5;

I – интервал варьирования натуральных значений переменной, для нашего примера 12,5,

	A	B	C	D
1	Номер предприятия	Количество уволившихся, %	Зарплата, тыс. руб.	
2		Y	X	X^2
3	1	6	-1	1
4	2	3,5	-0,6	0,36
5	3	2,2	-0,2	0,04
6	4	2	0,2	0,04
7	5	1,5	0,6	0,36
8	6	1,5	1	1

то получим следующую корреляционную матрицу

	A	B	C	D
1		*Столбец 1*	*Столбец 2*	*Столбец 3*
2	Столбец 1	1		
3	Столбец 2	-0,8825201	1	
4	Столбец 3	0,44070881	0	1

с независимыми переменными x и x^2 (столбцы 2 и 3). В этом случае значения каждого коэффициента регрессии и его доверительные границы определяются независимо от других коэффициентов регрессии. Поэтому, если один или несколько коэффициентов регрессии окажутся незначимыми, их можно отбросить, при этом не изменятся оценки других коэффициентов.

Задание. Произведите корреляционно-регрессионный анализ модели второго порядка для условий рассмотренной задачи с кодированным фактором. Сравните результаты расчета с ранее полученным результатом, раскодируйте полученное уравнение регрессии, пересчитайте его коэффициенты. Убедитесь в равенстве коэффициентов регрессии в обоих случаях.

Библиографический список

1. Адлер Ю. П., Маркова Е. Б., Грановский Ю. В. Планирование эксперимента при поиске оптимальных условий / 2-е изд. – М.: Наука, 1976. – 279 с.

2. Ахтназарова С. Л., Кафаров В. В. Методы оптимизации эксперимента в химической технологии: Учебн. пособие для хим.-технол. спец. вузов. – 2-е изд., перераб. и доп. – М.: Высш. шк., 1985. – 327 с.

3. Джонсон Н., Лион Ф. Статистика и планирование эксперимента в технике и науке: Методы обработки данных / Пер. с англ. – М.: Мир, 1980 – 610 с.

4. Дрейпер Н., Смит Г. Прикладной регрессионный анализ: в 2-х кн. Кн. 2 / Пер. с англ. – 2-е изд., перераб. и доп. – М.: Финансы и статистика, 1987. – 351 с.

5. Леоненков А. В. Решение задач оптимизации в среде MS Excel. – СПб.: БХВ-Петербург, 2005. – 704 с.

6. Мартынов В.М. Оптимизация технологических процессов общественного питания. – Уфа: Башкирский ГАУ, 2018. -137 с.

7. Мельников С. В., Алешкин В. Р., Рощин П. М. Планирование эксперимента в исследованиях сельскохозяйственных процессов. – Л.: Колос, 1980. – 168 с.

8. Митков А. Л., Кардашевский С. В. Статистические методы в сельхозмашиностроении. – М.: Машиностроение, 1978. – 360 с.

9. Митропольский А. К. Техника статистических вычислений. – М.: Наука. Гл. ред. физ.-мат. лит., 1971. – 576 с.

10. Монтгомери Д. К. Планирование эксперимента и анализ данных: Пер. с англ. – Л.: Судостроение, 1980. – 384 с.

11. Налимов В. В., Чернова Н. А. Статистические методы планирования экстремальных экспериментов. – М.: Наука, 1965. – 340 с.

12. Хартман К., Лецкий Э., Шеффер В. Планирование эксперимента в исследованиях технологических процессов. – М.: Мир, 1977. – 552 с.