СТАТИСТИЧЕСКИЙ АНАЛИЗ СВЯЗЕЙ МЕЖДУ ВЕЛИЧИНАМИ: КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ

1 Корреляционный анализ

Одна из наиболее распространенных задач корреляционного анализа состоит в том, чтобы выяснить, имеется ли значимая связь между исследуемыми величинами, т.е. можно ли утверждать, что изменение одной или нескольких величин, как правило, сопровождается изменением каких-либо других величин (одной или нескольких).

1.1 Анализ связи между двумя величинами. Коэффициент парной корреляции

Во многих случаях требуется выяснить, имеется ли значимая *линейная* связь между *двумя* исследуемыми величинами (т.е. связь, которую можно приближенно описать формулой $y = a_0 + a_1x$, где a_0, a_1 – некоторые коэффициенты). Для выявления и анализа такой связи вычисляется коэффициент парной корреляции (коэффициент корреляции Пирсона) между исследуемыми величинами.

Примечание – Если оказывается, что связь между величинами существует, то возникает задача построения модели этой связи, т.е. определения коэффициентов уравнения $y = a_0 + a_1x$, Это задача регрессионного анализа, рассматриваемая во второй части данной работы.

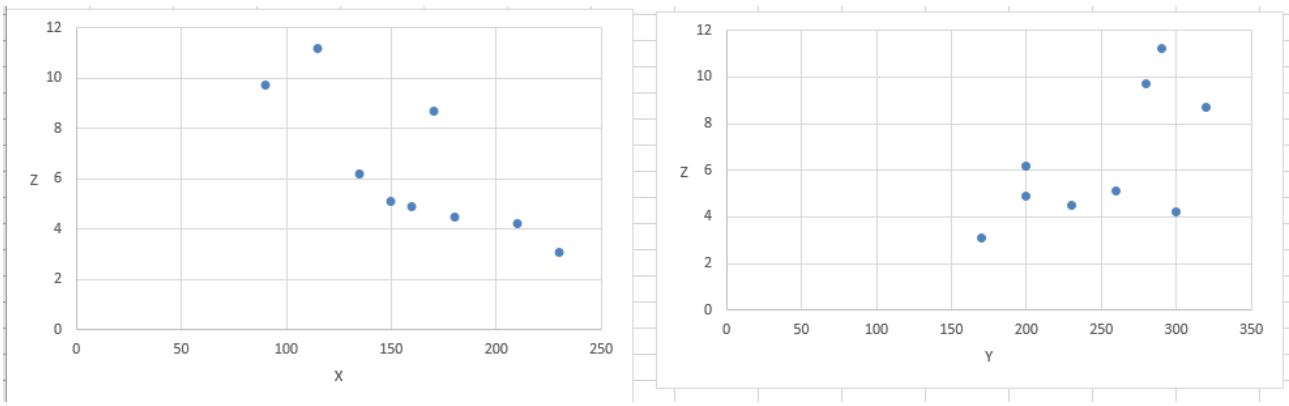
Пример 1 – Исследуются факторы, влияющие на текучесть кадров на предприятиях. Анализируется зависимость текучести кадров от заработной платы и размера предприятия (численности работающих). Получены данные по девяти предприятиям.

Средняя заработная плата, ден.ед.	135	170	115	160	230	180	90	210	150
Численность работающих	200	320	290	200	170	230	280	300	260
Текучесть кадров, %	6,2	8,7	11,2	4,9	3,1	4,5	9,7	4,2	5,1

Найти коэффициенты парной корреляции: а) между текучестью кадров и средней заработной платой; б) между текучестью кадров и численностью работающих. Выполнить анализ полученных результатов.

1 Перейти на свободный рабочий лист. В ячейки A1, B1, C1 ввести заголовки "Заработная плата", "Численность работающих", "Текучесть кадров". В столбцы A, B, C ввести исходные данные.

2 Для иллюстрации связей между величинами построить две диаграммы (тип диаграммы – **Точечная**):



а) **б)**

Рисунок 1 – Диаграммы, описывающие связь текучести кадров со средней зарплатой (а) и с численностью работающих (б)

3 В любой свободной ячейке вычислить коэффициент парной корреляции между текучестью кадров и средней заработной платой, используя функцию **КОРРЕЛ** со следующими аргументами: **Массив1:** A2:A10, **Массив2:** C2:C10.

Примечание – Формула для вычисления коэффициента парной корреляции (например, между некоторыми величинами *X* и *Y*) следующая:

$$r_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{X}\,\overline{Y}}{(n-1)\sigma_x \sigma_y},$$

где $\overline{X}$, $\overline{Y}$ – средние;

σ_x, σ_y – среднеквадратические (стандартные) отклонения;

n – объем выборки.

4 Выполнить проверку статистической значимости найденного коэффициента парной корреляции. Для этого найти расчетный уровень значимости (обычно обозначаемый *P*): упрощенно говоря, это вероятность того, что коэффициент парной корреляции *статистически незначим*. Эта величина определяется следующим образом:

– вычислить *t*-критерий Стьюдента по формуле:

$$t = r\frac{\sqrt{n-2}}{\sqrt{1-r^2}},$$

где *r* – коэффициент парной корреляции, найденный на шаге 3; *n* – объем анализируемой выборки (в данном примере *n*=9);

– в ячейку E4 ввести обозначение "P". В ячейке F4 найти расчетный уровень значимости. Для этого используется функция **СТЬЮДЕНТ.РАСП.2Х** с аргументами: **X –** абсолютное значение *t*-критерия, **Степени свободы –** *n*-2.

По найденным значениям коэффициента парной корреляции r и его расчетного уровня значимости P выполняется анализ связи между исследуемыми величинами (в данном примере – между средней заработной платой и текучестью кадров). Если $P < \alpha$ (обычно используется α=0,05), то коэффициент парной корреляции между исследуемыми величинами статистически значим. Это означает, что имеется достаточно существенная линейная связь между исследуемыми величинами. При этом если r>0, то связь между исследуемыми величинами положительная: с ростом одной величины увеличивается и другая. Значение r<0 указывает на отрицательную связь между величинами: с ростом одной величины другая уменьшается.

Если коэффициент парной корреляции статистически незначим, это означает, что *линейная* связь между исследуемыми величинами выражена слабо. Это может означать, что связь между величинами является более сложной (нелинейной) или вообще отсутствует.

Примечание – Приближенную оценку связи между исследуемыми величинами можно выполнить по значению коэффициента парной корреляции r (даже не проверяя его статистическую значимость). Значение r, близкое к 1, указывает на положительную линейную связь между исследуемыми величинами. Значение r, близкое к –1, указывает на отрицательную линейную связь между ними. Значение r, близкое к 0, является признаком отсутствия линейной связи. Однако такая оценка является лишь самой приближенной, поэтому практически всегда следует выполнять проверку статистической значимости.

В рассматриваемой задаче результаты должны быть примерно следующими: r = –0,796; t = –3,48; P = 0,01. Так как P < 0,05, коэффициент парной корреляции значим. Это означает, что между заработной платой и текучестью кадров имеется значимая линейная связь. Так как коэффициент парной корреляции r<0 (и при этом достаточно близок к –1), можно утверждать, что связь отрицательная. Таким образом, результаты исследования показывают, что заработная плата является значимым фактором, влияющим на текучесть кадров. Рост заработной платы сопровождается снижением текучести кадров, а снижение заработной платы – ростом текучести кадров.

5 Аналогично рассмотренному примеру найти коэффициент парной корреляции между численностью работающих и текучестью кадров, проверить его статистическую значимость и сделать выводы о связи между этими величинами.

Пример 2 – Решить задачу из примера 1, используя инструмент **Корреляция**.

Для решения этой задачи перейти на рабочий лист, где введены данные для примера 1. Выбрать элемент меню **Данные – Анализ данных – Корреляция**. В появившемся окне указать необходимые параметры: **Входной интервал:** A1:C10, **Группирование: по столбцам**. Установить флажок **Метки в первой строке**. В области **Параметры вывода** указать, куда требуется вывести результаты. Нажать **ОК**.

В этом примере вычисляются *три* коэффициента парной корреляции: между заработной платой и текучестью кадров; между численностью работающих и текучестью кадров; между заработной платой и численностью работаю-

щих. Проверку значимости коэффициентов парной корреляции и их анализ необходимо выполнять так же, как показано в примере 1. Результаты, полученные обоими методами, конечно же, должны совпадать.

1.2 Анализ связи между несколькими величинами. Коэффициент множественной корреляции

Если требуется выяснить, имеется ли *линейная* связь некоторой величины y с несколькими величинами x_1, x_2,…, x_k, то используется коэффициент множественной корреляции, обозначаемый как $r_{Y \cdot X_1, X_2, \ldots, X_k}$. Например, если исследуется связь некоторой величины y с пятью величинами x_1, x_2,…, x_5, то коэффициент множественной корреляции обозначается как $r_{Y \cdot X_1, X_2, X_3, X_4, X_5}$.

Будем называть величины x_1, x_2,…, x_k входными (независимыми), а величину y – выходной (зависимой).

Вычисление коэффициента множественной корреляции рассмотрим на следующем примере.

Пример 3 – Исследуется связь прочности меха некоторых животных, оцениваемой по 100-балльной шкале (чем прочнее мех, тем выше балл), с длиной шерсти и с весом животного. Имеются данные для 12 животных (см. таблицу).

Длина шерсти, мм	14	25	11	17	29	10	15	27	19	11	18	28
Вес животного, кг	6,7	5,8	3,9	4,7	5,2	3,5	7,1	6,6	7,1	4,2	7,4	5,1
Показатель прочности меха, баллы	48	30	38	42	37	42	49	41	53	52	43	37

Найти коэффициент множественной корреляции между показателем прочности меха, длиной шерсти и весом животного. Сделать вывод о связи между исследуемыми величинами.

Обозначим длину шерсти как x_1, вес животного - как x_2, показатель прочности меха - как Y.

1 Получить матрицу коэффициентов парной корреляции.

Для этого в MS Excel сначала получить все коэффициенты парной корреляции, используя инструмент **Корреляция** из набора инструментов **Анализ данных** (результат показан на рисунке 2). Затем заполнить матрицу коэффициентов парной корреляции, как показано на рисунке 3. Матрица коэффициентов парной корреляции симметрична.

	A	B	C	D	E	F	G	H	I	J
1	Длина шерсти, мм	Вес животного, кг	Показатель прочности меха, баллы							
2	14	6,7	48				*Длина шерсти, мм*	*Вес животного, кг*	*Показатель прочности меха, баллы*	
3	25	5,8	30			Длина шерсти, мм	1			
4	11	3,9	38			Вес животного, кг	0,314806763	1		
5	17	4,7	42			Показатель прочност	-0,533535927	0,285371578	1	
6	29	5,2	37							
7	10	3,5	42							
8	15	7,1	49							
9	27	6,6	41							
10	19	7,1	53							
11	11	4,2	52							
12	18	7,4	43							
13	28	5,1	37							
14										

Рисунок 2 – Коэффициенты парной корреляции

	A	B	C	D	E	F	G	H	I	J
1	Длина шерсти, мм	Вес животного, кг	Показатель прочности меха, баллы							
2	14	6,7	48				*Длина шерсти, мм*	*Вес животного, кг*	*Показатель прочности меха, баллы*	
3	25	5,8	30			Длина шерсти, мм	1	0,314806763	-0,533535927	
4	11	3,9	38			Вес животного, кг	0,314806763	1	0,285371578	
5	17	4,7	42			Показатель прочност	-0,533535927	0,285371578	1	
6	29	5,2	37							
7	10	3,5	42							
8	15	7,1	49							
9	27	6,6	41							
10	19	7,1	53							
11	11	4,2	52							
12	18	7,4	43							
13	28	5,1	37							
14										

Рисунок 3 – Матрица коэффициентов парной корреляции

2 Найти определитель матрицы коэффициентов парной корреляции (обозначим его как R).

В данном примере (см. рисунок 4) он найден в ячейке G9 с использованием функции **МОПРЕД** с параметром **Массив: G3:I5**.

	A	B	C	D	E	F	G	H	I	J
1	Длина шерсти, мм	Вес животного, кг	Показатель прочности меха, баллы							
2	14	6,7	48				*Длина шерсти, мм*	*Вес животного, кг*	*Показатель прочности меха, баллы*	
3	25	5,8	30			Длина шерсти, м	1	0,314806763	-0,533535927	
4	11	3,9	38			Вес животного, к	0,314806763	1	0,285371578	
5	17	4,7	42			Показатель прочн	-0,533535927	0,285371578	1	
6	29	5,2	37							
7	10	3,5	42							
8	15	7,1	49							
9	27	6,6	41			R	0,438936748			
10	19	7,1	53			R*	0,900896702			
11	11	4,2	52			$r_{y\cdot x1,x2}$	0,716085147			
12	18	7,4	43							
13	28	5,1	37			F	7,346252633			
14						P	0,012833404			
15						$F_{a,k,n-k-1}$	4,256494729			

Рисунок 4 – Вычисление и проверка значимости коэффициента множественной корреляции

3 Найти определитель матрицы, полученной из матрицы коэффициентов парной корреляции вычеркиванием строки и столбца, содержащих коэффициенты корреляции величины y с другими величинами (обозначим этот определитель как R^*).

В данном примере (см. рисунок 4) он найден в ячейке G10 с использованием функции **МОПРЕД** с параметром **Массив: G3:H4**.

4 Найти коэффициент множественной корреляции по формуле:

$$r_{Y\cdot X_1,X_2} = \sqrt{1-\frac{R}{R*}}$$

где R и $R*$ – определители, вычисленные на шагах 2 и 3.

В данном примере он вычислен в ячейке G11 по формуле: **=КОРЕНЬ(1-G9/G10)**.

Коэффициент множественной корреляции может принимать значения в диапазоне от нуля до единицы. Чем ближе коэффициент множественной корреляции к *единице*, тем более выражена *линейная* связь между величинами Y и X_1, X_2,…, X_k. Если коэффициент множественной корреляции близок к *нулю*, это значит, что анализируемые величины не связаны друг с другом, или связь между ними нелинейная.

5 Выполнить проверку статистической значимости коэффициента множественной корреляции. Для этого вычисляется F-критерий Фишера:

$$F = \frac{(n-k)\cdot r^2_{Y\cdot X_1,X_2}}{(k-1)\cdot\sqrt{1-r^2_{Y\cdot X_1,X_2}}}.$$

где k – количество независимых величин (в данной задаче k=2).

В данном примере он вычислен в ячейке G13.

По значению критерия F и числу степеней свободы k и n-k-1 вычисляется расчетный уровень значимости P. Если $P < \alpha$, где α – заданный уровень значимости (обычно α =0,05), то коэффициент множественной корреляции статистически значим.

В данном примере значение P найдено в ячейке G14 с помощью функции **=F.РАСП.ПХ** с параметрами **X**: G13, **Степени свободы 1**: 2, **Степени свободы 2**: 9. Так как $P < \alpha$, коэффициент множественной корреляции статистически значим, т.е. имеется значимая связь прочности меха с длиной шерсти и весом животного.

Примечание – В MS Excel нет специальных средств для вычисления коэффициента множественной корреляции. Он вычисляется при построении линейных регрессионных моделей связей между величинами с использованием инструмента **Регрессия** и выводится под обозначением **Множественный R** (см. пункт 2.3.2).

Задание 1 – Анализируются факторы, влияющие на заработную плату на предприятиях по добыче некоторого стройматериала. Имеются данные по 11 предприятиям:

Производительность, т/день	115	96	124	88	117	106	92	105	120	92	125
Глубина добычи, м	4,3	5,6	3,4	9,7	6,5	3,1	4,1	7,1	3,7	2,8	5,1
Доля материала высшего сорта, %	46	34	48	32	46	36	34	35	47	32	47
Средняя зарплата, ден.ед.	1200	1000	1300	900	1250	1080	950	1000	1240	950	1300

Требуется найти коэффициент множественной корреляции заработной платы с исследуемыми факторами. Проверить его статистическую значимость. Сделать выводы о связи между исследуемыми величинами.

1.3 Анализ связи между несколькими величинами. Коэффициенты частной корреляции

Коэффициент частной корреляции отражает связь между *двумя* величинами при *исключении* (элиминировании) влияния *остальных величин*. Коэффициент частной корреляции между величинами y и x_j обозначается как $r_{YX_j \cdot X_1, X_2, ..., X_k}$. Например, если исследуется связь некоторой величины y с пятью величинами x_1, x_2,…, x_5, то коэффициент частной корреляции между y и x_4 (при исключении влияния x_1, x_2, x_3, x_5) обозначается как $r_{YX_4 \cdot X_1, X_2, X_3, X_5}$.

Примечание – При вычислении коэффициентов частной корреляции используются *алгебраические дополнения* к элементам матрицы коэффициентов парной корреляции. Алгебраическое дополнение к элементу матрицы - это определитель матрицы, полученной *вычеркиванием* данного элемента из исходной матрицы, умноженный на $(-1)^{a+b}$, где a и b - номера строки и столбца элемента матрицы (т.е. номера вычеркиваемых строки и столбца).

Коэффициент частной корреляции между величинами y и x_j находится по следующей формуле:

$$r_{YX_j \cdot X_1, X_2, ..., X_k} = \frac{-A_{YX_j}}{\sqrt{A_{X_jX_j} \cdot A_{YY}}}, \qquad (1)$$

где A_{YX_j}, $A_{X_jX_j}$, A_{YY} – алгебраические дополнения к элементам матрицы парных корреляций r_{YX_j}, $r_{X_jX_j}$, r_{YY} соответственно.

Коэффициент частной корреляции принимает значения от -1 до 1. Его смысл такой же, как и для коэффициента парной корреляции (см. раздел 1.1).

Для проверки статистической значимости коэффициента частной корреляции $r_{YX_j \cdot X_1, X_2, ..., X_k}$ используется t-критерий Стьюдента:

$$t_j = \frac{r_{YX_j \cdot X_1, X_2, ..., X_k} \sqrt{n-k-1}}{\sqrt{1 - r^2_{YX_j \cdot X_1, X_2, ..., X_k}}}. \qquad (2)$$

Пример 4 – По данным примера 3 вычислить коэффициенты частной корреляции между: а) показателем прочности меха и длиной шерсти; б) показателем прочности меха и весом животного.

а) Коэффициент частной корреляции между показателем прочности меха и длиной шерсти ($r_{YX_1 \cdot X_2}$) вычисляется по формуле (1):

$$r_{YX_1 \cdot X_2} = \frac{-A_{YX_1}}{\sqrt{A_{X_1X_1} \cdot A_{YY}}},$$

Вычисление в MS Excel показано на рисунке 5.

	A	B	C	D	E	F	G	H	I
1	Длина шерсти, мм	Вес животного, кг	Показатель прочности меха, баллы						
2	14	6,7	48				*Длина шерсти, мм*	*Вес животного, кг*	*Показатель прочности меха, баллы*
3	25	5,8	30			Длина шерсти, мм	1	0,314806763	-0,533535927
4	11	3,9	38			Вес животного, кг	0,314806763	1	0,285371578
5	17	4,7	42			Показатель прочности	-0,533535927	0,285371578	1
6	29	5,2	37						
7	10	3,5	42						
8	15	7,1	49			A_{YX1}	0,314806763	-0,533535927	0,62337283
9	27	6,6	41				1	0,285371578	
10	19	7,1	53						
11	11	4,2	52			A_{X1X1}	1	0,285371578	0,918563062
12	18	7,4	43				0,285371578	1	
13	28	5,1	37						
14						A_{YY}	1	0,314806763	0,900896702
15							0,314806763	1	
16									
17						$r_{YX1 \cdot X2}$	-0,685260789		
18									
19						t1	-2,822722291		
20						P	0,019958107		

Рисунок 5 – Вычисление коэффициента частной корреляции $r_{YX_1 \cdot X_2}$

Здесь в ячейках G8:H9 размещена матрица для вычисления алгебраического дополнения A_{YX_1}. Матрица получена вычеркиванием коэффициента парной корреляции r_{YX_1} из матрицы этих коэффициентов. Вычеркнутый элемент находится в третьей строке, первом столбце. Алгебраическое дополнение вычислено в ячейке I8 по формуле: **=(-1)^(3+1)*МОПРЕД(G8:H9)**.

Аналогично в ячейках I11 и I14 вычислены алгебраические дополнения $A_{X_1X_1}$ и A_{YY}.

В ячейке G17 вычислен коэффициент частной корреляции $r_{YX_1 \cdot X_2}$ по формуле (1): **=-I8/КОРЕНЬ(I11*I14)**.

В ячейке G19 по формуле (2) вычислен *t*-критерий (t_1) для проверки значимости найденного коэффициента частной корреляции. Здесь n=12 (объем выборки), k = 2 (количество независимых переменных).

В ячейке G20 по значению *t*-критерия и числу степеней свободы n-k-1 найден расчетный уровень значимости P. Для этого использована функция **СТЬЮДЕНТ.РАСП.2Х** с параметрами **X: ABS(G19); Степени свободы: 9**.

При заданном уровне значимости α=0,05, выполняется условие $P < \alpha$. Значит, коэффициент частной корреляции $r_{YX_1 \cdot X_2}$ статистически значим, т.е. имеется значимая линейная связь прочности меха с длиной шерсти. Эта связь обратная, так как коэффициент частной корреляции – отрицательный.

б) Коэффициент частной корреляции между показателем прочности меха и весом животного ($r_{YX_2 \cdot X_1}$) вычисляется по формуле (1):

$$r_{YX_2 \cdot X_1} = \frac{-A_{YX_2}}{\sqrt{A_{X_2X_2} \cdot A_{YY}}},$$

Вычисление в MS Excel показано на рисунке 6.

	A	B	C	D	E	F	G	H	I
1	Длина шерсти, мм	Вес животного, кг	Показатель прочности меха, баллы						
2	14	6,7	48				*Длина шерсти, мм*	*Вес животного, кг*	*Показатель прочности меха, баллы*
3	25	5,8	30			Длина шерсти, мм	1	0,314806763	-0,533535927
4	11	3,9	38			Вес животного, кг	0,314806763	1	0,285371578
5	17	4,7	42			Показатель прочности	-0,533535927	0,285371578	1
6	29	5,2	37						
7	10	3,5	42						
8	15	7,1	49			A_{YX2}	1	-0,533535927	-0,453332297
9	27	6,6	41				0,314806763	0,285371578	
10	19	7,1	53						
11	11	4,2	52			A_{X2X2}	1	-0,533535927	0,715339415
12	18	7,4	43				-0,533535927	1	
13	28	5,1	37						
14						A_{YY}	1	0,314806763	0,900896702
15							0,314806763	1	
16									
17						$r_{YX2 \cdot X1}$	0,564706887		
18									
19						t2	2,052754175		
20						P	0,070299779		

Рисунок 6 – Вычисление коэффициента частной корреляции $r_{YX_2 \cdot X_1}$

Здесь, например, алгебраическое дополнение A_{YX_2} вычислено в ячейке I8 по формуле: **=(-1)^(3+2)*МОПРЕД(G8:H9)** (так как коэффициент парной корреляции r_{YX_2} находится в третьей строке, втором столбце исходной матрицы коэффициентов парной корреляции).

Так как в данном случае $P > \alpha$, коэффициент частной корреляции $r_{YX_2 \cdot X_1}$ статистически незначим, т.е. линейная связь прочности меха с весом животного выражена слабо.

Примечание – В MS Excel нет специальных средств для вычисления коэффициентов частной корреляции. Для их вычисления необходимо использовать специальные программные средства, например, программу Statistica.

Задание 2 – По данным задания 1 найти коэффициенты частной корреляции заработной платы с каждым из анализируемых факторов. Проверить их статистическую значимость и сделать выводы.

2 Регрессионный анализ

Методы регрессионного анализа позволяют по выборкам значений двух величин (x и y) или нескольких величин (x_1, x_2, …,x_m, y) построить модель (уравнение), описывающую связь между этими величинами. Основной математический метод, применяемый для построения таких моделей – метод наименьших квадратов. Основная идея этого метода состоит в построении уравнения связи между исследуемыми величинами, *максимально соответствующего* фактическим данным, т.е. наблюдаемым значениям этих величин.

В MS Excel основным средством решения задач регрессионного анализа является инструмент **Регрессия** из пакета **Анализ данных**. Кроме того, модели связи между двумя величинами могут строиться на основе диаграмм.

2.1 Линейные модели связи двух величин

Пример 5 – Исследуется зависимость между некоторыми двумя величинами (x и y). Имеются значения этих величин, полученные в десяти наблюдениях.

x	2,2	2,8	3,5	2,6	2	2,9	3,1	3,4	1,9	3,5
y	21,4	28,6	35,7	21,7	18,5	27,1	34,1	31,4	18,1	21,2

Требуется построить *линейную* модель связи между исследуемыми величинами: $y = a_0 + a_1 x$, где a_0, a_1 – коэффициенты модели, которые требуется определить, используя метод наименьших квадратов. Выполнить проверку адекватности модели и статистической значимости ее коэффициентов.

2.1.1 Построение модели

1 Перейти на свободный рабочий лист. В ячейку A1 ввести заголовок “X”, в ячейку B1 – заголовок “Y”. В ячейки A2:A11 и B2:B11 ввести исходные данные.

2 Из меню **Данные – Анализ данных** выбрать инструмент **Регрессия**. Установить параметры: **Входной интервал Y:** B1:B11, **Входной интервал X:** A1:A11. Установить флажок **Метки**. В области **Параметры вывода** указать, куда требуется вывести результаты. Для получения результатов нажать **ОК**. Результаты будут иметь примерно такой вид, как показано на рисунке 7.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,735357212
R-квадрат	0,540750229
Нормированный R-квадра	0,483344008
Стандартная ошибка	4,656404212
Наблюдения	10

Дисперсионный анализ

	df	*SS*	*MS*	*F*	***Значимость F***
Регрессия	1	204,2391985	204,2391985	9,419714733	**0,015367589**
Остаток	8	173,4568015	21,68210018		
Итого	9	377,696			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	***P-Значение***	*Нижние 95%*	*Верхние 95%*
Y-пересечение	**3,794223168**	7,313234083	0,518816043	**0,617931139**	-13,07012486	20,65857119
X	**7,88020675**	2,567548977	3,069155378	**0,015367589**	1,959428197	13,8009853

Рисунок 7 – Построение линейной модели связи между двумя величинами с использованием инструмента Регрессия

Искомые коэффициенты модели, вычисленные по методу наименьших квадратов, находятся в столбце **Коэффициенты**. Сначала указывается коэффициент a_0, затем – a_1. В данном примере a_0=3,79, a_1=7,88 (значения коэффициен-

тов указаны с округлениями). Таким образом, связь между величинами x и y может быть описана уравнением $y = 3{,}79 + 7{,}88\,x$.

Здесь для вычисления коэффициентов модели $a_0 = 3{,}79$ и $a_1 = 7{,}88$ применен метод наименьших квадратов. Это означает, что при подстановке в построенную модель имеющихся значений x (т.е. 2,2, 2,8, …, 3,5) получаются модельные значения y (т.е. $\hat{y}_1 = 3{,}79 + 7{,}88 \cdot 2{,}2 = 21{,}13$, $\hat{y}_2 = 3{,}79 + 7{,}88 \cdot 2{,}8 = 25{,}86$,…, $\hat{y}_{10} = 3{,}79 + 7{,}88 \cdot 3{,}5 = 31{,}38$), максимально близкие к фактическим значениям (т.е. $y_1 = 21{,}4$, $y_2 = 28{,}6$, …, $y_{10} = 21{,}2$). Более точно, коэффициенты a_0 и a_1 вычисляются таким образом, чтобы минимизировать *среднеквадратическую ошибку* $Q_e = \sum_{j=1}^{n} (\hat{y}_j - y_j)^2$, т.е. минимизировать разности между модельными и фактическими значениями.

2.1.2 Анализ и проверка модели

а) Проверка адекватности модели. Для этого используется расчетный уровень значимости **Значимость F**. Эта величина представляет собой вероятность того, что построенная модель является неадекватной (недостаточно точной). Если расчетный уровень значимости *меньше* заданного уровня значимости α (обычно используется α=0,05), то модель адекватна исходным данным (т.е. достаточно точная). В данном примере величина З**начимость F** равна (округленно) 0,015, что меньше, чем 0,05. Таким образом, построенная модель связи между величинами x и y является адекватной. Упрощенно говоря, это означает следующее: если в построенную модель ($y = 3{,}79 + 7{,}88\,x$) подставить известные значения x (т.е. значения 2,2; 2,8; …;3,5), то будут получены модельные значения y (т.е. $\hat{y}_1 = 3{,}79 + 7{,}88 \cdot 2{,}2 = 21{,}13$, $\hat{y}_2 = 3{,}79 + 7{,}88 \cdot 2{,}8 = 25{,}86$,…, $\hat{y}_{10} = 3{,}79 + 7{,}88 \cdot 3{,}5 = 31{,}38$), достаточно близкие к фактическим (21,4; 28,6;…; 21,2).

Примечание – Если построенная линейная модель оказывается неадекватной (величина **Значимость F** превышает заданный уровень значимости), это означает, что связь между исследуемыми величинами x и y не может быть достаточно точно описана линейным уравнением $y = a_0 + a_1 x$, т.е. является более сложной (нелинейной) или вообще отсутствует.

б) Проверка значимости коэффициентов модели. Эта проверка выполняется по расчетным уровням значимости, указанным в столбце **P-Значение**. Если расчетный уровень значимости *меньше* заданного уровня значимости α, то соответствующий коэффициент модели статистически значим. В данном примере **P-значение** для коэффициента a_0 составляет $0{,}618 > 0{,}05$, для коэффициента a_1 – $0{,}015 < 0{,}05$. Таким образом, коэффициент a_0 статистически незначим, коэффициент a_1 – статистически значим. Значимость коэффициента модели при входной переменной (в рассматриваемом примере – значимость коэффициента a_1 при переменной x) означает, что имеется статистически значимая (т.е. достаточно сильно выраженная) линейная связь величин y и x.

в) Коэффициент детерминации. Из других величин, вычисляемых с помощью инструмента **Регрессия**, интерес представляет коэффициент детерминации, обозначаемый как **R-квадрат**. Эта величина характеризует точность построенной модели: чем ближе коэффициент детерминации к единице, тем точнее модель. В данном примере коэффициент детерминации равен 0,54. Это значит, что построенная модель позволяет объяснить примерно 54% различий в наблюдаемых значениях величины y.

Основные выводы по результатам решения задачи таковы. Связь между величинами x и y может быть описана следующей моделью: $y = 3{,}79 + 7{,}88\ x$. эта модель является адекватной, т.е. описывает связь между x и y достаточно точно. Линейная связь между величинами x и y статистически значима, т.е. выражена достаточно сильно.

2.1.3 Решение задачи с использованием диаграммы (построение тренда)

Пусть требуется построить точечную диаграмму, отражающую значения исследуемых величин, и нанести на нее график линейного уравнения связи между этими величинами: $y = a_0 + a_1 x$. Эта задача решается следующим образом.

1 Построить точечную диаграмму, отражающую связь между исследуемыми величинами (тип диаграммы – **Точечная**).

2 Нанести на диаграмму линейный тренд, т.е. график линейного уравнения (линейной модели), описывающего связь между исследуемыми величинами. Тренд строится в следующем порядке:

- выделить диаграмму щелчком мыши;
- из меню **Диаграмма** выбрать команду **Добавить линию тренда**;
- в окне **Линия тренда,** на вкладке **Тип,** выбрать вид линии тренда – **Линейная**. На вкладке **Параметры** выбрать переключатель **Название аппроксимирующей сглаживающей кривой: другое**. Установить флажки **Показывать уравнение на диаграмме** и **Поместить на диаграмму величину достоверности аппроксимации (R^2)**. Нажать кнопку **ОК**. На диаграмме будет построен линейный тренд, указано его уравнение, а также коэффициент детерминации (R^2). Диаграмма, описывающая связь между величинами x и y, приведена на рисунке 8. Легко убедиться, что построенная модель связи между исследуемыми величинами совпадает с моделью, полученной с помощью инструмента **Регрессия**.

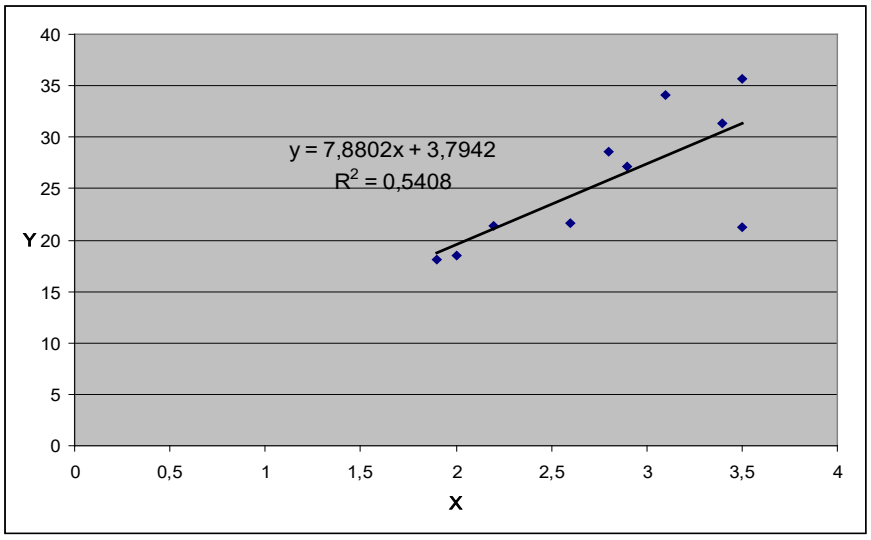


Рисунок 8 – Построение линейной модели связи между величинами *X* и *Y* с использованием диаграммы

2.1.4 Прогнозирование на основе регрессионной модели

Регрессионные модели могут применяться для прогнозирования *среднего значения* выходной величины y при известном значении входной величины x. Для этого достаточно подставить входную величину в уравнение модели.

Воспользуемся построенной в примере 5 моделью связи величин y и x, чтобы найти ожидаемое значение y (обозначим его как y_0) при x=1,8: $y_0 = 3{,}79 + 7{,}88 \cdot 1{,}8 = 17{,}98$. Таким образом, среднее значение y составит 17,98.

Очевидно, что этот прогноз не может быть абсолютно точным, и в любом конкретном наблюдении при заданном значении x могут быть получены разные значения y. Поэтому представляют интерес *доверительные интервалы* для среднего значения выходной величины и для ее отдельных значений, найденные с заданной вероятностью.

Доверительный интервал для *среднего значения* выходной величины y при заданном значении входной величины $x=x_0$ с заданной доверительной вероятностью P (или с уровнем значимости $\alpha=1-P$) определяется по следующей формуле:

$$y_0 \pm t_{\alpha;n-2}\sqrt{\frac{Q_e}{n-2}}\sqrt{\frac{1}{n}+\frac{(x_0-\overline{X})^2}{\sum_{i=1}^{n}x_i^2-n\overline{X}^2}}, \qquad (*)$$

а доверительный интервал для *отдельных значений* выходной величины y – по следующей формуле:

$$y_0 \pm t_{\alpha;n-2}\sqrt{\frac{Q_e}{n-2}}\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{X})^2}{\sum_{i=1}^{n}x_i^2-n\overline{X}^2}}, \qquad (**)$$

где y_0 – прогноз среднего значения выходной величины (см. выше).

Для рассмотренного примера найдем 95-процентные доверительные интервалы для среднего значения величины y и ее отдельных значений при x=1,8.

Здесь x_0=1,8, y_0=17,98. Так как требуется определить 95-процентный доверительный интервал, доверительная вероятность равна 0,95, а уровень значимости α=1–0,95=0,05. Квантиль распределения Стьюдента $t_{\alpha;n-2}=t_{0,05;8}$=2,306 (определяется с помощью функции **СТЬЮДЕНТ.ОБР.2Х**). Найдем остальные величины: n=10, $\overline{X}$=2,79, $\sum_{i=1}^{n} x_i^2$=81,13, $Q_e = \sum_{j=1}^{n} (\hat{y}_j - y_j)^2$ = (21,13-21,4)^2 + (25,86-28,6)^2 + … + (31,38-21,2)^2 = 173,46.

Доверительный интервал для среднего значения y, найденный по формуле (*), составляет (11,2; 24,75). Доверительный интервал для отдельных значений Y, определяемый по формуле (**), составляет (5,28; 30,67). Таким образом, можно ожидать, что при x = 1,8 с вероятностью 95% среднее значение y составит от 11,2 до 24,75. Конкретные значения y с вероятностью 95% будут составлять от 5,28 до 30,67.

Задание 3 – Имеются данные о производительности труда и годовой прибыли на 11 предприятиях:

Производительность, изделий/день	115	96	124	88	117	106	92	105	120	92	125
Прибыль, млн ден.ед.	1,20	1,00	1,30	0,90	1,25	1,08	0,95	1,00	1,24	0,95	1,30

Построить линейную модель зависимости прибыли от производительности: $y = a_0 + a_1 x$, где x – производительность, y – прибыль. Выполнить анализ и проверку модели. Построить диаграмму с трендом, иллюстрирующую построенную модель. Найти прогноз средней прибыли и отдельных значений прибыли при производительности труда 90 изделий/день.

2.2 Нелинейные модели связи двух величин

Пример 6 – Исследуется зависимость между некоторыми двумя величинами (x и y). Имеются значения этих величин, полученные в семи наблюдениях.

x	5	10	20	30	40	50	60
y	48	32	17	14	9	5	2

Требуется построить модели зависимости величины Y от X: логарифмическую, степенную, экспоненциальную, линейную модель. Определить наиболее точную модель.

1 Перейти на свободный рабочий лист. В ячейку A1 ввести заголовок “X”, в ячейку B1 – “Y”. В ячейках A2:A8 и B2:B8 ввести исходные данные.

2 Построить точечную диаграмму, отражающую связь между исследуемыми величинами (см. пример 5, пункт 2.1.3 – построение модели с использованием диаграммы, шаг 1).

3 Сделать три копии построенной диаграммы. Таким образом, всего на рабочем листе должны находиться *четыре* диаграммы, так как будут строиться четыре модели.

4 Построить на диаграммах степенной, линейный, логарифмический и экспоненциальный тренды (по одному тренду на каждой диаграмме), как показано в примере 5 (пункт 2.1.3, шаг 2).

На рисунке 9 показана *степенная* модель связи между величинами x и y. Видно, что эта связь описывается уравнением $y = 391.4x^{-1,1125}$. Коэффициент детерминации для этой модели равен 0,8628.

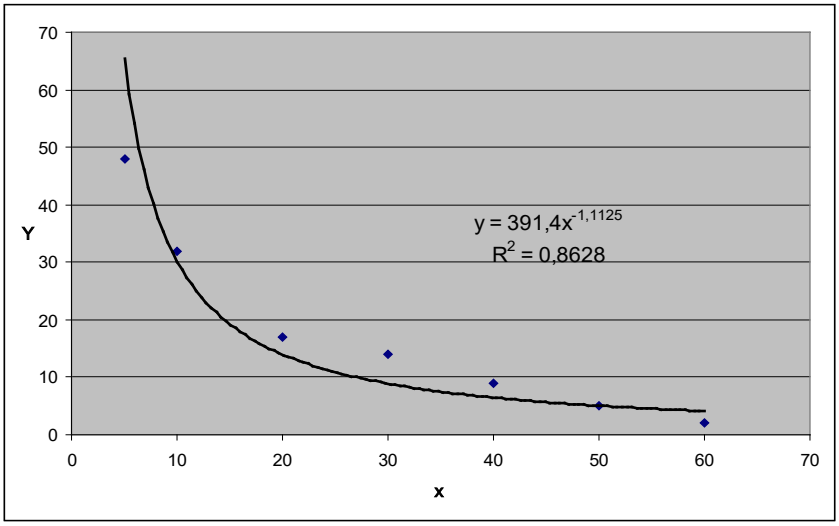


Рисунок 9 – Построение модели связи между величинами *x* и *y* (степенная модель)

По максимальному значению коэффициента детерминации (R^2) определяется наиболее точная (т.е. наиболее близкая к фактическим данным) модель связи между x и y.

2.3 Линейные модели связи нескольких величин

Пример 7 – По данным примера 1 построить линейную модель связи между исследуемыми величинами: $y = a_0 + a_1x_1 + a_2x_2$, где y – текучесть кадров (выходная переменная); x_1 – заработная плата, x_2 – численность работающих (x_1, x_2 – входные переменные); a_0, a_1, a_2 – коэффициенты модели, которые требуется определить, используя метод наименьших квадратов. Выполнить проверку адекватности модели и статистической значимости ее коэффициентов.

2.3.1 Построение модели

1 Перейти на свободный рабочий лист. В ячейки A1, B1, C1 ввести заголовки. В ячейки A2:C10 ввести исходные данные.

2 Из меню **Данные – Анализ данных** выбрать инструмент **Регрессия**. Установить параметры: **Входной интервал Y:** C1:C10, **Входной интервал X:** A1:B10. Установить флажок **Метки**. В области **Параметры вывода** указать, куда требуется вывести результаты. Для получения результатов нажать **ОК**. Результаты будут иметь примерно такой вид, как показано на рисунке 10.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,887070707
R-квадрат	0,786894438
Нормированный R-квадрат	0,715859251
Стандартная ошибка	1,490392668
Наблюдения	9

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2	49,21237818	24,60618909	11,0775303	0,009677972
Остаток	6	13,32762182	2,221270304		
Итого	8	62,54			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	7,723900223	3,804490171	2,030206381	0,088638449	-1,585351846	17,03315229
Средняя заработная плата	-0,042633786	0,012534916	-3,401202279	0,014474486	-0,073305622	-0,011961951
Численность работающих	0,021990022	0,010586646	2,077147216	0,083061963	-0,003914567	0,047894612

Рисунок 10 – Построение линейной модели связи между несколькими величинами с использованием инструмента Регрессия

Из полученных результатов видно, что модель связи текучести кадров (y) со средней заработной платой (x_1) и численностью работающих (x_2) имеет следующий вид: $y = 7{,}72 - 0{,}04x_1 + 0{,}02x_2$. Коэффициенты модели (7,72, -0,04 и 0,02) найдены методом наименьших квадратов.

2.3.2 Анализ и проверка модели

Модель является адекватной: величина **Значимость F** равна 0,0097, т.е. меньше, чем 0,05.

Из величин, указанных в столбце **P-значение**, видно, что коэффициент a_1 статистически значим (**P-значение** < 0,05). Из этого следует, что имеется значимая линейная связь между текучестью кадров и средней заработной платой. Следует также отметить, что эта связь – обратная (чем больше средняя заработная плата, тем меньше текучесть кадров), так как коэффициент a_1 – отрицательный. Коэффициенты a_0 и a_2 статистически незначимы (**P-значение** > 0,05). В то же время, для этих коэффициентов **P-значение** достаточно мало (**P-значение** < 0,1). Для коэффициента a_2 это означает, что исключать линейную связь между текучестью кадров и численностью работающих не следует, но она выражена слабее, чем для средней зарплаты. Коэффициент a_0 также не следует исключать из модели.

Коэффициент детерминации для построенной модели достаточно высок (**R-квадрат** = 0,787), что также подтверждает высокую точность модели.

Результаты построения модели включают также коэффициент множественной корреляции, обозначенный как **Множественный R**. Его вычисление показано в разделе 1.2.

Интерпретация коэффициентов модели. Коэффициенты при входных переменных x_1, x_2 показывают, на сколько, в среднем, изменяется выходная переменная y при изменении соответствующей входной переменной на единицу.

В данном случае a_1 = -0,04 означает, что *повышение* средней заработной платы на одну денежную единицу приводит к *снижению* текучести кадров в среднем на 0,04%. Коэффициент a_2 = 0,02 означает, что *увеличение* численности работающих на одного человека соответствует *увеличению* текучести кадров в среднем на 0,02%.

Коэффициент a_0 представляет собой среднее значение выходной переменной y при $x_1 = x_2 = 0$. В данном примере коэффициент a_0 не имеет конкретного физического смысла, так как, по смыслу задачи, входные переменные здесь не могут быть равны нулю.

2.3.3 Прогнозирование на основе регрессионной модели

Построенная модель может применяться для прогнозирования *среднего* значения выходной переменной при известном значении входных переменных. Например, для предприятия со средней зарплатой 200 ден.ед. и численности работающих 300 чел. ожидаемая текучесть кадров составляет в среднем 7,72 – $0{,}04 \cdot 200 + 0{,}02 \cdot 300 = 5{,}72\%$. Очевидно, что этот прогноз – приближенный, и желательно найти прогноз с заданной точностью. Получение таких прогнозов для моделей с несколькими входными переменными требует сложных расчетов. Эта задача решается с использованием специализированных программных средств (например, программы Statistica).

Задание 4 – По данным задания 1 построить линейную модель зависимости средней зарплаты от анализируемых факторов (производительности, глубины добычи и доли материала высшего сорта). Выполнить анализ и проверку модели. Найти прогноз средней зарплаты для предприятия, где производительность составляет 90 т/день, глубина добычи – 5 м, доля материала высшего сорта – 30%.