

Лабораторная работа № 2 «ПАРНАЯ КОРРЕЛЯЦИЯ И ЛИНЕЙНАЯ РЕГРЕССИЯ В PAST»

1 Парная корреляция в PAST

Корреляционный анализ позволяет сделать заключение о наличии (или отсутствии) связи между анализируемыми признаками. При этом он показывает не только то, какова связь между двумя признаками по направлению (прямая или обратная), но и выражает её количественно при помощи коэффициента корреляции (r) – величины, изменяющейся от -1 до +1 (чем ближе коэффициент к 1, тем сильнее связь между признаками, а знак коэффициента корреляции указывает на направлении зависимости).

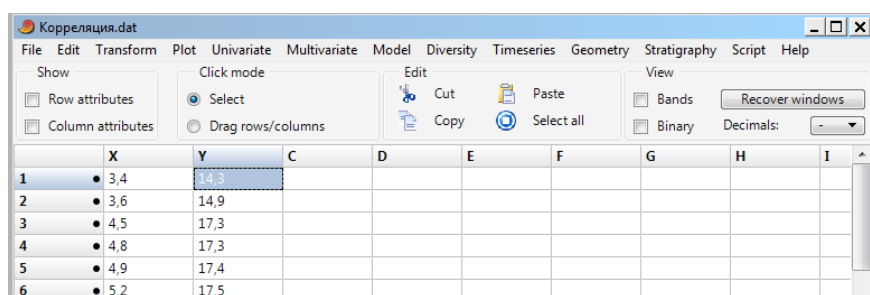
Задание 1. В следующей таблице приведены данные корреляционного анализа двух признаков, обозначенных как X и Y (таблица 1). Необходимо установить в PAST, есть ли корреляция между двумя анализируемыми признаками?

Таблица 1 – Данные для корреляционного анализа

X	Y	X	Y	X	Y
3,4	14,3	8,4	19,8	10,7	21,3
3,6	14,9	8,5	19,9	11,6	21,3
4,5	17,3	8,8	19,9	12	21,8
4,8	17,3	8,9	20,1	12,3	22,0
4,9	17,4	8,9	20,1	12,6	22,1
5,2	17,5	8,9	20,1	12,7	22,4
5,4	17,6	8,9	20,1	13,3	22,7
5,7	17,6	9,0	20,2	13,6	23,5
6,2	17,6	9,0	20,3	13,8	24,2
6,7	17,8	9,1	20,3	14,0	24,4
7,1	18,0	9,3	20,5	15,0	25,2
7,5	18,0	9,4	20,6	15,2	25,2
7,7	18,1	9,7	20,9	15,8	25,3
7,8	18,1	9,7	21,0	15,9	25,7
7,9	18,6	9,9	21,1	16,6	26,8
8,0	19,7	10,1	21,1	17,1	27,5

Решение:

Создадим электронную таблицу с данными (рисунок 1)



	X	Y	C	D	E	F	G	H	I
1	3,4	14,3							
2	3,6	14,9							
3	4,5	17,3							
4	4,8	17,3							
5	4,9	17,4							
6	5,2	17,5							

Рисунок 1 – Таблица с данными в PAST

Для доступа к модулю корреляционного анализа необходимо после выделения данных для анализа в электронной таблице выбрать в разделе **Univariate** (Одномерные данные) параметр **Correlation** (Корреляция) – рисунок 2.

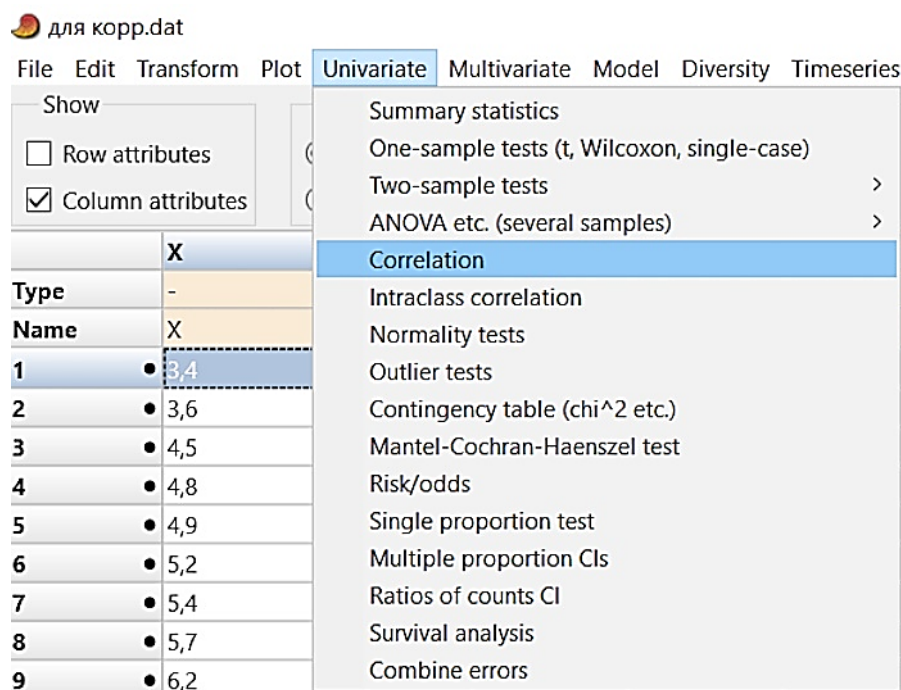


Рисунок 2 – Модуль **Correlation** (Корреляция) в меню PAST

После выбора данного пункта меню на экране будет показано окно с результатами анализа (рисунок 3). В нашем случае коэффициент корреляции составил очень высокое значение 0,98337 (~0,98) при достоверном уровне значимости p .

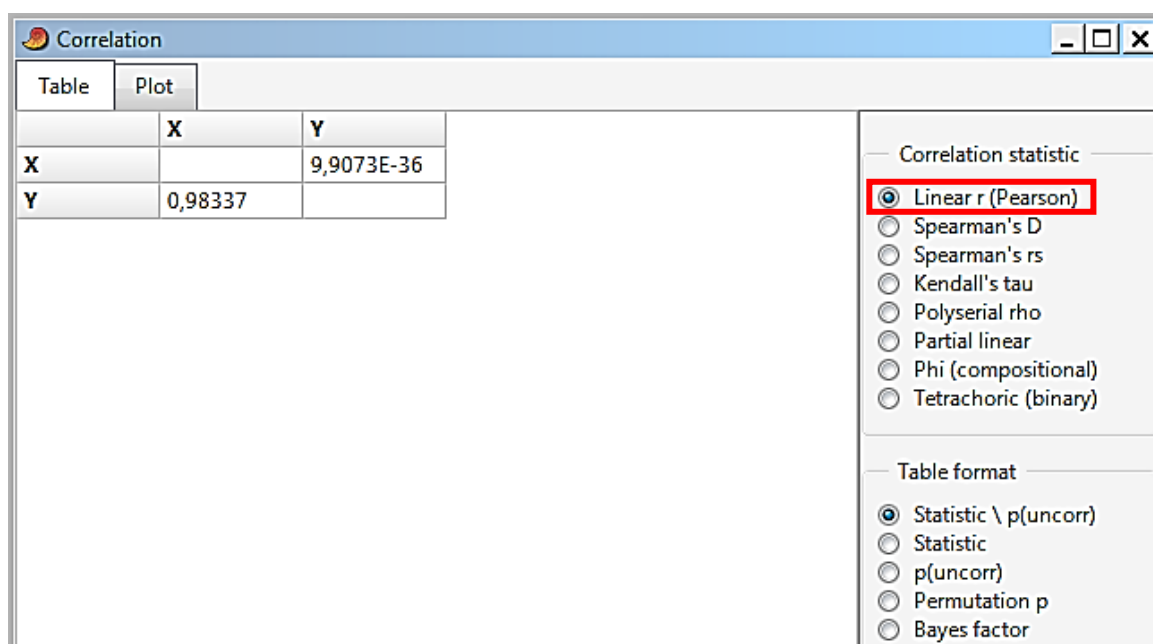


Рисунок 3 – Окно с результатами парного корреляционного анализа с рассчитанным коэффициентом корреляции Пирсона

По умолчанию программа высчитывает параметрический коэффициент корреляции Пирсона, что указано в правой части окна результатов в разделе **Correlation statistic** (*Корреляционная статистика*).

В том случае, если выборка не подчиняется закону нормального распределения, следует рассчитать непараметрический коэффициент корреляции Спирмена, который находится также в разделе **Correlation statistic** (*Корреляционная статистика*) окна результатов чуть ниже коэффициента корреляции Пирсона (рисунок 4).

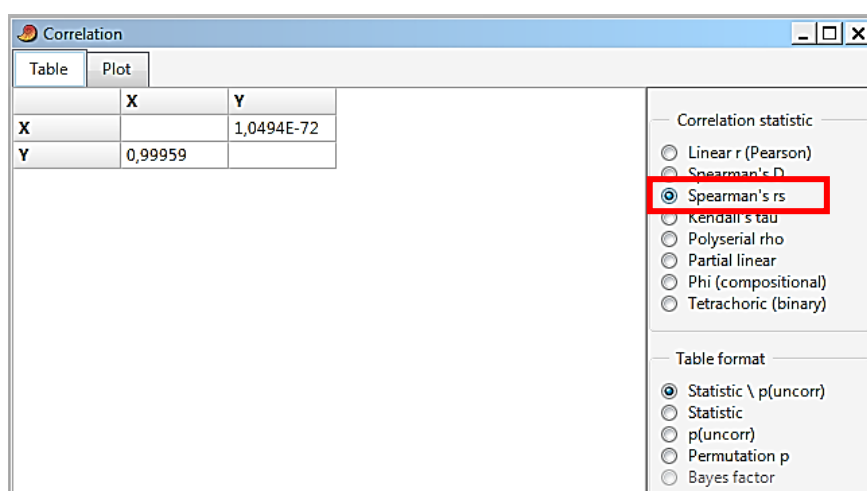


Рисунок 4 – Окно с результатами парного корреляционного анализа с рассчитанным коэффициентом корреляции Спирмена

Для достоверного анализа необходимо предварительно проверить обе выборки на нормальность распределения (в разделе меню **Univariate** → **Normality test**), а затем в зависимости от результата использовать либо коэффициент корреляции Пирсона (r , Pearson), либо Спирмана (r_s , Spearman's). В нашем случае обе выборки распределены нормально (согласно результатам теста Шапиро-Уилка – рисунок 5), поэтому будет использован коэффициент корреляции Пирсона.

Tests for normal distribution		
	X	Y
N	48	48
Shapiro-Wilk W	0,964	0,9644
p(normal)	0,1466	0,1521
Anderson-Darling A	0,6123	0,7081
p(normal)	0,1051	0,06043
p(Monte Carlo)	0,1027	0,0601
Lilliefors L	0,1238	0,1103
p(normal)	0,06147	0,1469
p(Monte Carlo)	0,0652	0,1451
Jarque-Bera JB	1,659	1,133
p(normal)	0,4364	0,5674
p(Monte Carlo)	0,2867	0,4555

Рисунок 5 – Результаты теста Шапиро-Уилка на нормальность распределения

Следующий этап корреляционного анализа – графическое отображение результатов. Для этого есть два способа. Первый способ – это использование графика, который строится автоматически после проведения корреляционного анализа и который находится в окне результатов на закладке **Plot** (*Диаграммы*). Для большей информативности в направлении и силе корреляционной связи в правой части окна с графиком необходимо промаркировать бокс **Ellipses** (*Эллипсы*). Итоговый график отображен на рисунке 6.

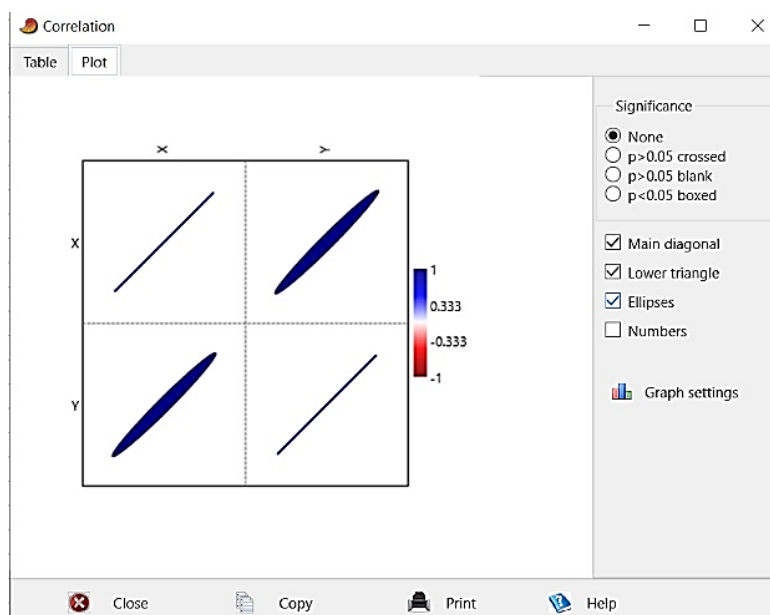


Рисунок 6 – Графическое отображение корреляции окна результатов корреляционного анализа

В настройке графика (**Graph settings**) измените размер рисунка в разделе **Size (pixels)** (*Размер в пикселях*) на 800 по ширине (**Width**) и 600 по высоте (**Height**) и смените шрифт (**Font**) на Arial 16 пт.

Второй способ графического отображения корреляционной связи – использование набора графики PAST из раздела **Plot** (*Графики и диаграммы*). Для этого предварительно выделив диапазон с данными необходимо в этом разделе выбрать пункт меню **XY graph** (*Диаграмма рассеивания по осям X и Y*) – рисунок 7.

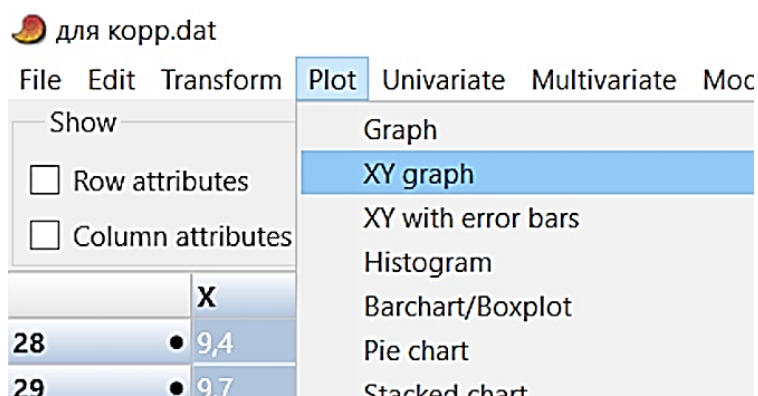


Рисунок 7 – Пункт меню **XY graph** в PAST

Получившаяся диаграмма рассеивания будет отображена на рисунке 8.

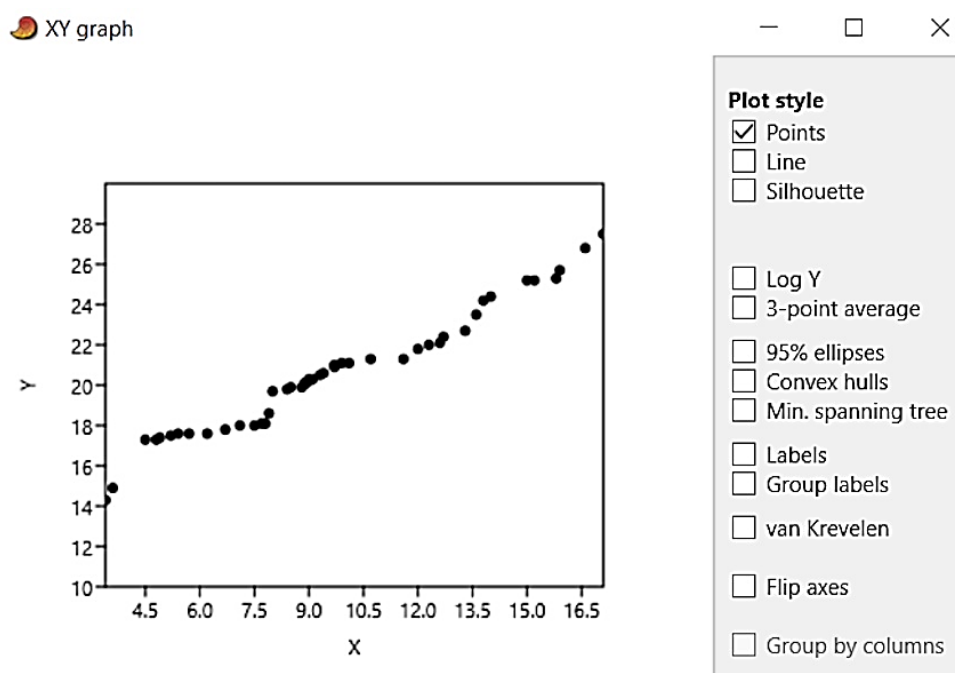


Рисунок 8 – Диаграмма рассеивания в PAST

Для более точного отображения корреляции необходимо на рисунке отобразить диапазон значений, который попадает в область с 95% точности. Для этого справа от рисунка в окне результатов в разделе **Plot style** (*Стиль графика*) необходимо промаркировать пункт **95% ellipses** (рисунок 9).

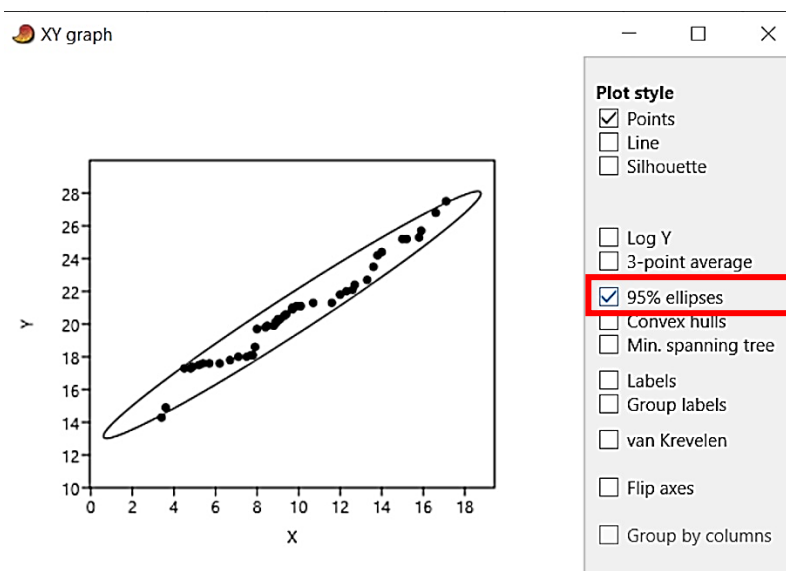


Рисунок 9 – Дополненная диаграмма рассеивания по результатам корреляционного анализа

В получившемся рисунке в настройке графика (**Graph settings**) измените размер рисунка в разделе **Size (pixels)** (*Размер в пикселях*) на 800 по ширине (**Width**) и 600 по высоте (**Height**) и смените шрифт (**Font**) на Arial 16 пт.

Выполните для закрепления задания 1–5 из заданий для самоконтроля в конце лабораторной работы.

2 Линейная регрессия в PAST

Не смотря на то, что расчёт корреляции характеризует направление и силу связи между двумя переменными, а коэффициент корреляции позволяет количественно охарактеризовать степень связи, в то же время с его помощью невозможно предсказать, чему в среднем будет равно значение одного признака при изменении другого на определённое значение.

Регрессионный анализ служит как раз для определения вида этой связи и даёт возможность для прогнозирования значения одной (зависимой) переменной отталкиваясь от значения другой (независимой) переменной.

Часто связь между двумя биологическими особенностями, выраженными какими-либо признаками, имеет линейный характер, которую можно выразить в виде линейного уравнения

$$y = ax + b, \text{ где:}$$

- y и x – анализируемые признаки (y – значение функции, зависимая переменная; x – аргумент, независимая переменная);

- a – коэффициент регрессии (отражает угол наклона линии регрессии – чем больше a отличается от 0, тем сильнее связь между анализируемыми признаками);

- b – свободный член уравнения (при $a = 0$ получаем $y = b$, т.е., другими словами, b – это та точка, в которой линия регрессии пересекается с осью ординат; эта точка называется «**Intercept**»).

Данное уравнение можно использовать для описания связи между 2 признаками только при выполнении обязательных условий:

- 1) Зависимость между признаками носит линейный характер;
- 2) Оба признака распределены нормально.

Такая модель позволяет оценить среднее значение переменной y при известном значении x . Разность между истинным значением и модельным называется *ошибкой* («error») или *остатком* («residual»):

$$E = y - m$$

В идеальном варианте остатки имеют нормальное распределение с нулевым средним и неизвестным, но постоянным разбросом σ^2 , который не зависит от значений x и y . В этом случае говорят о *гомогенности остатков*. Если же разброс зависит еще от каких-либо параметров, то остатки считаются *гетерогенными*. Кроме того, если обнаружилось, что средние значения остатков зависят от x , то это значит, что между y и x имеется нелинейная зависимость.

Для проверки гипотезы о том, что модель значительно отличается от нуля, используется так называемая F-статистика (статистика Фишера). Как обычно, если p -значение меньше уровня значимости (обычно 0,05), то модель считается значимой.

Задание 2. Для примера проведения регрессионного анализа в программе PAST используем данные 2 признаков, рассмотренные при изучении корреляционного анализа (см. таблица 1 данной темы).

Решение:

Создадим электронную таблицу с данными (рисунок 1). Далее для проведения регрессионного анализа нами будет использоваться линейная модель (так как мы проводим линейный регрессионный анализ). Для этого необходимо предварительно выделив диапазон данных перейти по пунктам меню **Model** (*Модель*) → **Linear** (*Линейная*) → **Bivariate regression** → (*Регрессия для двух переменных*) – рисунок 10.

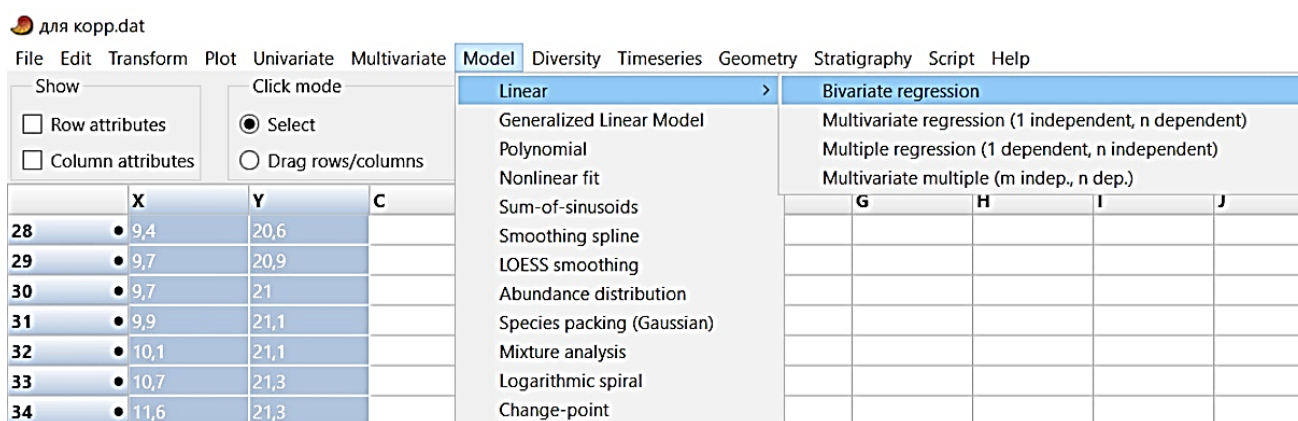


Рисунок 10 – Линейная регрессия для двух переменных в PAST

После выбора появится окно результатов с несколькими вкладками (рисунок 11).

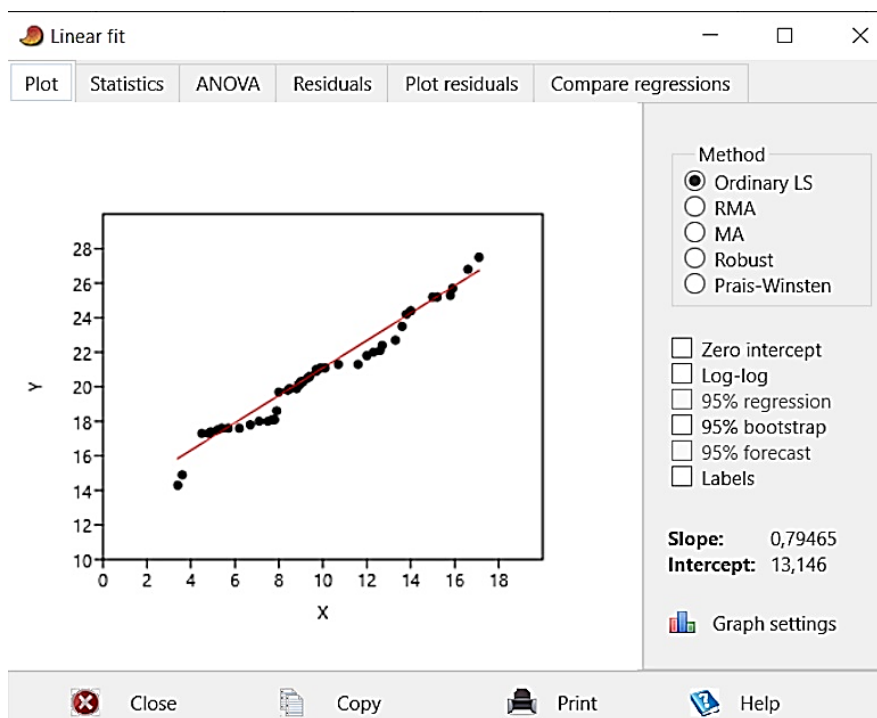


Рисунок 11 – Окно результатов линейной регрессии для двух переменных в PAST

Для нашей дальнейшей работы важны 5 закладок из 6-и имеющихся:

- **Plot** (*График*) – графическое отображение регрессионной зависимости;
- **Statistics** (*Статистики*) – статистические параметры полученной регрессии и корреляции;
- **ANOVA** (*Дисперсионный анализ*) – данные дисперсионного анализа (суммы квадратов и значение критерия Фишера);
- **Residuals** (*Остатки*) – данные по остаткам (отклонениям от линии регрессии);
- **Plot residuals** (*График остатков*).

Кратко рассмотрим каждую из перечисленных закладок.

Закладка «**Plot** (График)»

Появляется первой в окне результатов анализа. Для первоначального анализа этого достаточно, но для более полной характеристики линии тренда желательно добавить на график линии, ограничивающие 95% диапазон данных, выбрав справа в окне закладки опцию **95% regression**. Также следует увеличить размеры рисунка до 800 на 600 пикселей и изменить шрифт на Arial 16 пт. Получившийся график отображен на рисунке 12.

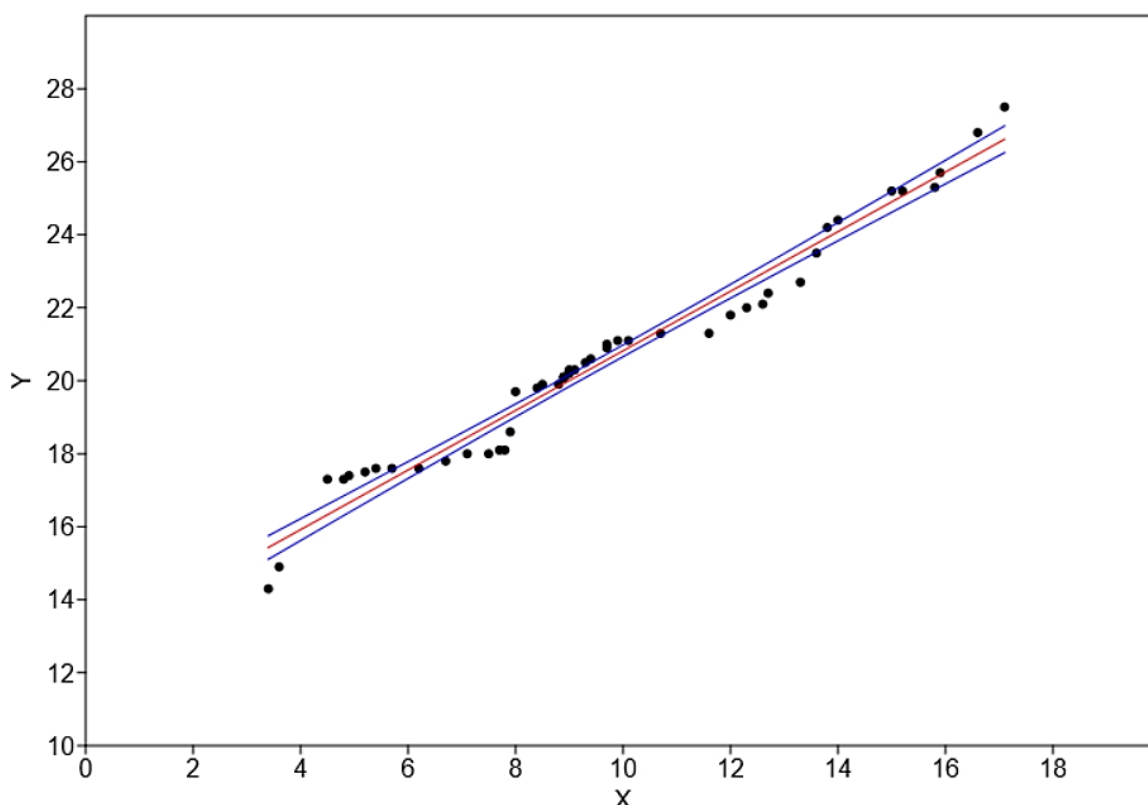


Рисунок 12 – График линейной регрессии в PAST

Закладка «**Statistics** (Статистики)»

В данной закладке отображаются результаты корреляционной связи и регрессии (рисунок 13).

Ordinary Least Squares Regression: X-Y

Slope a:	0,81635	Std. error a:	0,022229
t:	36,724	p (slope):	9,9073E-36
Intercept b:	12,657	Std. error b:	0,22916

95% bootstrapped confidence intervals ($N=1999$):

Slope a:	(0,76606, 0,87379)
Intercept b:	(12,076, 13,213)

Correlation:	
r:	0,98337
r^2:	0,96702
t:	36,724
p (uncorr.):	9,9073E-36
Permutation p:	0,0001

Рисунок 13 – Статистические результаты закладки **Statistics** (*Статистики*)

Здесь нас интересуют следующие показатели:

1) По корреляции (Correlation):

- r – значение коэффициента корреляции Пирсона
- r^2 – коэффициент детерминации R^2 (квадрат коэффициента множественной корреляции). Очень важный показатель регрессионного анализа, изменяется в пределах от 0 до 1 и отражает своеобразное «качество» рассчитанной регрессии, показывая долю (в процентах) общего разброса выборочных точек, которая объясняется построенной регрессией (например, при $R^2 = 0,96702$, как в нашем случае, следует вывод о том, что практически 97% дисперсии зависимой переменной Y объясняется вариацией независимой переменной X);
- p – уровень значимости коэффициента корреляции.

2) По регрессии (Ordinary Least Squares Regression: X-Y):

- **Slope a** – коэффициент при независимой переменной x ;
- **Intercept b** – свободный член уравнения, точка пересечения

Исходя из полученных данных, можно сказать, что уравнение полученной регрессии будет иметь вид: $y=0.81635x+12.657$.

Закладка «ANOVA (Дисперсионный анализ)»

В данной закладке (рисунок 14) приведены данные дисперсионного анализа в виде суммы квадратов регрессии и остатков, а также F-статистика (показатель критерия Фишера и его уровень значимости).

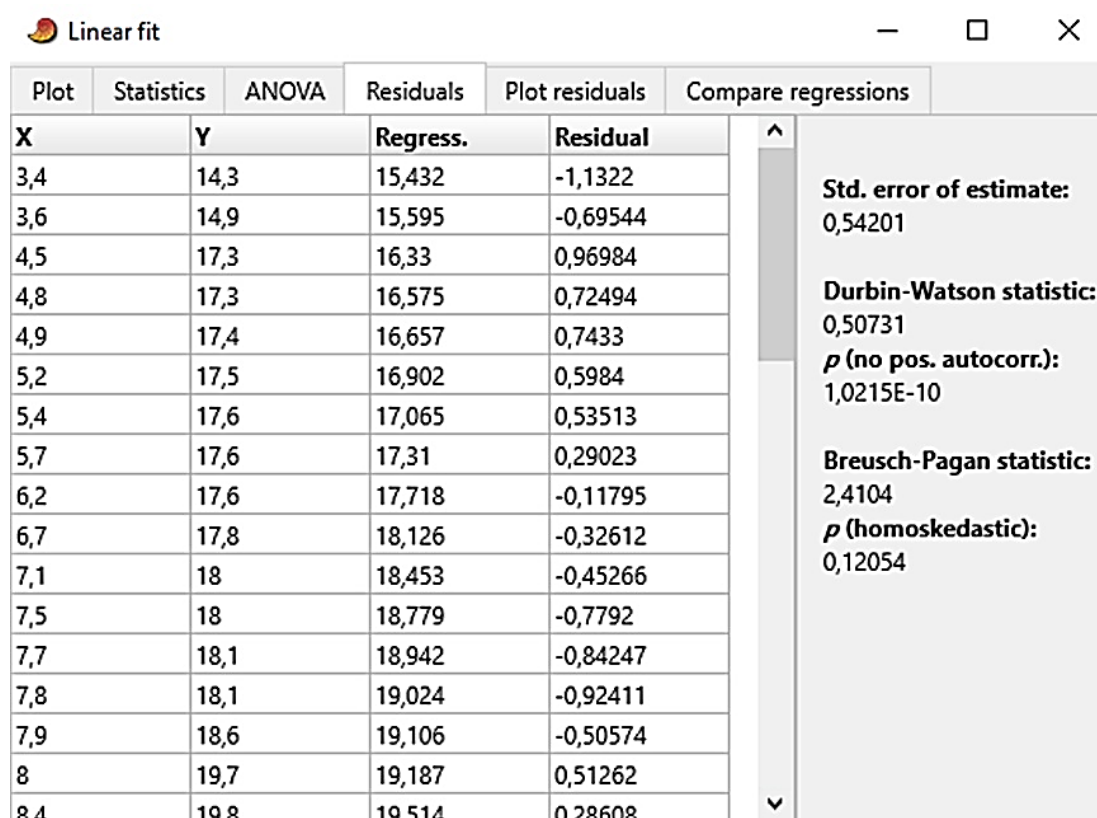
Ordinary Least Squares Regression: X-Y

	SS	df	MS	F	p
Regression	396,19	1	396,19	1348,6	4,9536E-36
Residual	13,513	46	0,29377		
Total SS	409,71				

Рисунок 14 – Статистические результаты закладки ANOVA (Дисперсионный анализ)

Закладка «Residuals (Остатки)»

В данной закладке (рисунок 15) приведены данные по отклонениям каждого из значений переменных от линии тренда регрессии.



Plot	Statistics	ANOVA	Residuals	Plot residuals	Compare regressions
X	Y	Regress.	Residual		
3,4	14,3	15,432	-1,1322		
3,6	14,9	15,595	-0,69544		
4,5	17,3	16,33	0,96984		
4,8	17,3	16,575	0,72494		
4,9	17,4	16,657	0,7433		
5,2	17,5	16,902	0,5984		
5,4	17,6	17,065	0,53513		
5,7	17,6	17,31	0,29023		
6,2	17,6	17,718	-0,11795		
6,7	17,8	18,126	-0,32612		
7,1	18	18,453	-0,45266		
7,5	18	18,779	-0,7792		
7,7	18,1	18,942	-0,84247		
7,8	18,1	19,024	-0,92411		
7,9	18,6	19,106	-0,50574		
8	19,7	19,187	0,51262		
8,4	19,8	19,514	0,28608		

Std. error of estimate:
0,54201
Durbin-Watson statistic:
0,50731
p (no pos. autocorr.):
1,0215E-10
Breusch-Pagan statistic:
2,4104
p (homoskedastic):
0,12054

Рисунок 15 – Статистические результаты закладки Residuals (Остатки)

Здесь также можно посмотреть стандартную ошибку остатков (**Std. Error of estimate**).

Закладка «Plot residuals (График остатков)»

Данная закладка имеет важное значение при проверке остатков в регрессионном анализе и, следовательно, насколько можно доверять полученной модели и сформированному уравнению регрессии. По умолчанию PAST показывает график остатков в виде диаграммы рассеивания (**Scatter plot**), показанный на рисунке 16.

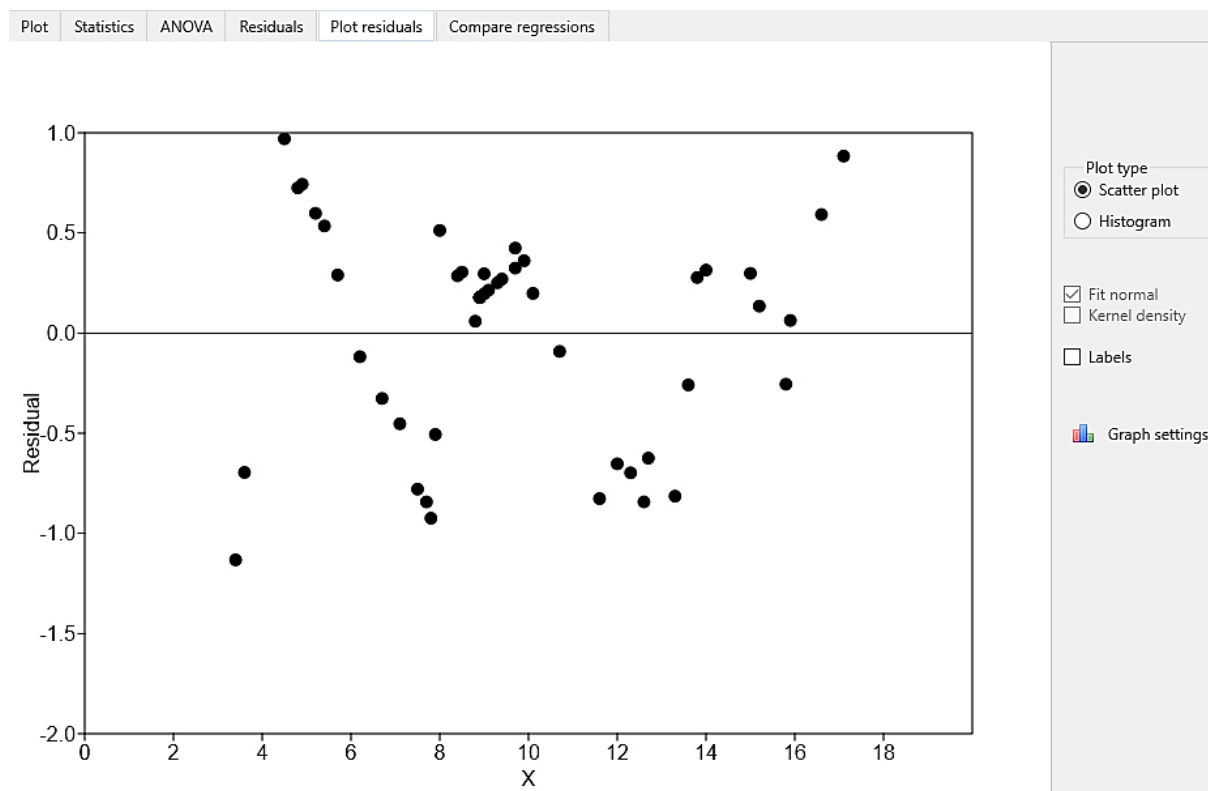


Рисунок 16 – Диаграмма рассеивания остатков закладки **Plot residuals** (*График остатков*)

Проверка о рабочей модели выполняется в том случае, если точки на данном графике будут располагаться хаотично и не проявлять закономерностей. В нашем случае график демонстрирует именно такую зависимость. В противном случае (в расположении точек имеется тенденция – разброс увеличивается слева направо, точки тесно укладываются вдоль прямой и др.), линейный регрессионный анализ неприменим.

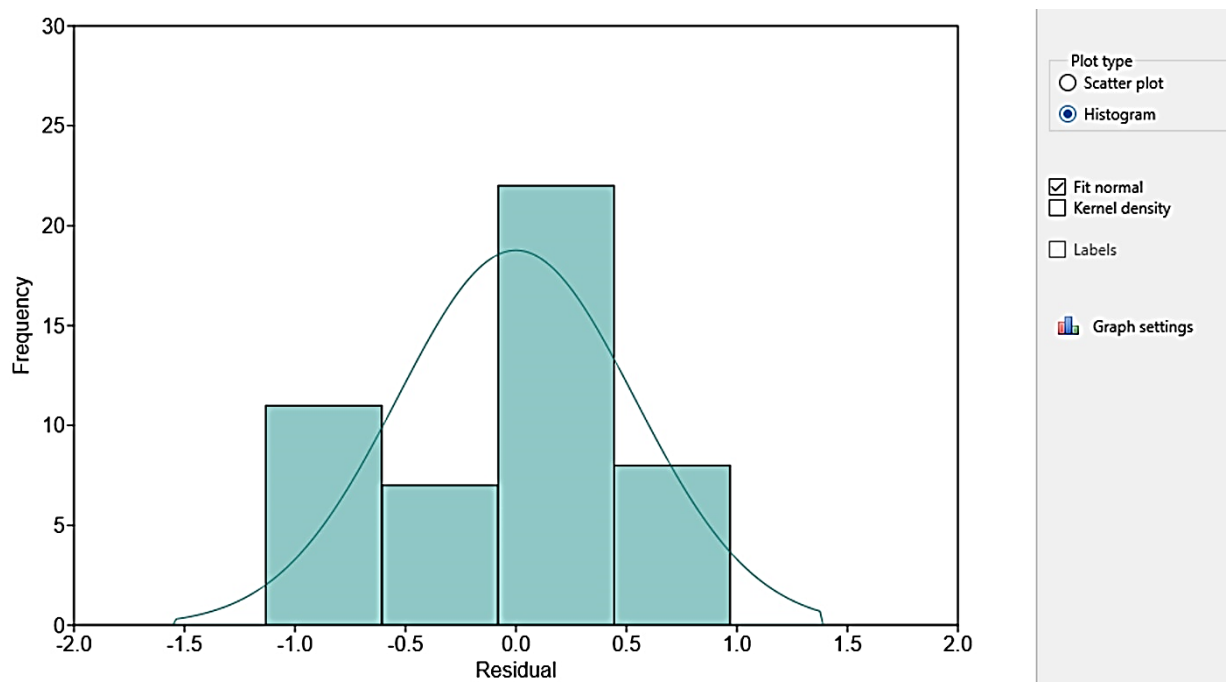


Рисунок 16 – Гистограмма остатков закладки **Plot residuals** (*График остатков*)

Также подобное условие (соответствие линейной модели) нашей полученной регрессии будет сохраняться если остатки будут подчиняться закону нормального распределения. Это также можно проверить в этой закладке просто выбрав справа в окне в разделе **Plot type** (*Тип графика*) график в виде гистограммы (**Histogram**), при этом обратите внимание, чтобы опция **Fit normal** (*Линия нормального распределения*) там же, справа, была промаркирована. Полученную гистограмму сравните с рисунком 17.

Таким образом, в результате проведенного нами регрессионного анализа можно сказать следующее:

1. Выявленная зависимость имеет достоверно линейную модель распределения.

2. Полученная модель имеет высокую степень прямой связи и около 97% всех значений переменных укладываются в пределы модели.

3. Формула зависимости переменной y от переменной x имеет вид:
 $y = 0.81635x + 12.657$.

Выполните для закрепления задания 6–10 из заданий для самоконтроля в конце лабораторной работы

Задания для самоконтроля

1) Надо установить, есть ли корреляция между высотой головы (x) и длиной 3-го членика усика (y) *Drosophila funebris*:

x	15	16	15	15	16	16	17	18	18	17	17	17	15	16	15	15	15	17
y	29	31	32	33	32	33	33	36	36	35	35	35	35	33	31	31	31	35
x	15	13	15	14	17	15	16	15	15	16	15	16	15	16	18	17	14	15
y	33	30	32	31	35	33	33	32	30	33	33	33	30	31	34	34	31	33
x	14	15	15	13	15	16	14	15	15	15	14	15	15	15	16	18	15	14
y	31	31	33	30	30	33	30	33	31	32	30	31	31	32	33	35	32	32

2) Получены следующие данные о продолжительности беременности у кроликов при различных размерах помета (число крольчат в помете (x) и длительность беременности в днях (y)):

x	1	8	3	5	7	8	4	8	3	4	4	8	8	5	7	6	6	5	6	6
y	33	30	31	31	31	32	31	31	32	33	32	31	31	31	31	30	31	32	32	31
x	6	5	7	8	10	6	7	6	7	6	5	10	7	8	8	6	5	6	5	4
y	32	32	31	32	31	31	30	31	31	32	31	30	32	32	31	31	31	32	30	31
x	6	8	6	5	8	7	6	5	9	5	3	4	7	8	9	5	6	2	2	4
y	31	31	32	32	31	30	32	31	31	31	32	32	31	31	31	31	31	32	33	33

Есть ли корреляция между длительностью плодоношения и размерами помета? Определите необходимый коэффициент корреляции и постройте диаграмму рассеивания. Сделайте вывод.

3) Были получены следующие данные о весе ягнят-баранчиков (единцов) – у и весе баранов – их отцов – х (в кг).

х	76,6	72,2	67,0	66,5	63,3	65,4	63,9	63,1	63,0	62,5	62,2
у	4,56	4,79	4,49	4,32	4,59	4,32	4,67	4,29	4,57	4,20	4,12
х	61,0	60,2	60,0	59,6	59,5	58,9	58,0	57,8	57,6	57,0	
у	4,13	4,70	3,80	4,23	3,76	4,08	4,61	4,37	4,30	4,0	
х	56,8	55,4	55,0	53,8	53,7	52,0	51,4	51,0	50,9	48,5	
у	3,82	4,12	4,19	4,16	4,09	4,12	4,02	4,31	4,06	4,03	

Есть ли корреляция между весом баранчиков и весом их отцов? Определите необходимый коэффициент корреляции и постройте диаграмму рассеивания. Сделайте вывод.

4) У 40 серебристо-черных лисиц были измерены (в см) длина туловища х и длина хвоста у:

х	70	65	66	65	71	68	64	57	66	65	67	62	67	62	63	57	64	66	69	58
у	40	40	40	40	40	42	39	38	41	43	39	45	43	38	40	40	41	45	43	37
х	63	67	67	67	65	65	67	70	65	71	69	64	64	66	69	72	66	66	67	66
у	45	38	39	37	42	38	38	38	38	40	39	43	43	42	40	41	47	47	40	40

Определите необходимый коэффициент корреляции между длиной туловища и длиной хвоста серебристо-белых лисиц. Постройте диаграмму рассеивания, сделайте вывод.

5) Были получены следующие данные о весе х (в кг) и длине туловища у (в см) 100 серебристо-черных лисиц:

х	4,7	4,6	5,2	5,1	5,3	5,3	4,6	4,8	5,8	5,7	4,5	5,7	5,0	4,8	4,7	5,2	4,6			
у	70	65	69	70	66	68	65	71	69	68	57	73	65	67	71	62	69			
х	5,5	5,5	4,8	4,7	6,0	5,1	5,	4,5	5,0	5,0	4,9	5,5	5,2	5,6	5,2	5,7	5,3			
у	62	63	67	64	64	66	68	69	58	63	67	74	67	67	70	65	71			
х	5,4	5,3	4,6	5,6	5,1	4,9	5,2	5,3	5,0	5,3	5,6	5,0	5,1	5,5	5,6	5,2	5,0			
у	63	64	64	66	63	69	62	72	66	66	67	67	66	63	67	62	71			
х	5,5	5,6	5,0	6,7	4,7	5,3	5,0	5,1	5,0	5,1	4,8	5,0	6,0	5,5	4,6	4,5	4,5			
у	67	66	66	69	64	69	70	62	68	68	72	68	67	66	69	65	65			
х	5,4	5,0	4,9	5,0	5,7	5,9	5,6	5,1	5,1	4,6	4,9	6,2	5,6	5,2	5,1	4,5				
у	65	65	64	66	66	67	62	63	64	69	69	68	65	69	67	68				
х	4,8	5,5	6,0	5,3	4,8	5,3	5,1	5,4	4,7	5,0	5,9	5,0	5,2	5,6	5,2	5,1				
у	61	64	62	66	59	65	62	68	61	67	69	69	66	66	67	70				

Определите, есть ли корреляция между весом и длиной туловища у лисиц? Определите необходимый коэффициент корреляции и постройте диаграмму рассеивания. Сделайте обоснованный вывод.

6) У 20 взрослых мужчин были измерены высота (длина тела) x (в см) и вес y (в кг):

x	165	176	175	168	167	172	175	180	179	173
y	56	75	70	61	61	63	72	80	76	68
x	166	178	169	169	170	176	180	169	177	176
y	58	76	60	64	63	71	78	63	75	71

Определите коэффициент регрессии, постройте график и уравнение линейной регрессии, проверьте остатки.

7) Предполагается, что между количеством настриженной шерсти y и живым весом овец x имеется зависимость. Для 10 овец были получены следующие данные (в кг):

x	50	55	60	50	65	60	50	55	50	65
y	4,0	4,2	4,1	4,2	4,5	4,3	4,1	4,4	4,0	4,2

Определите коэффициент регрессии, постройте график и уравнение линейной регрессии, проверьте остатки.

8) Для 10 петушков леггорнов 15-дневного возраста были получены следующие данные о весе их тела x (в г) и весе гребня y (в мг):

x	83	72	69	90	90	95	95	91	75	70
y	56	42	18	84	56	107	90	68	31	48

Определите коэффициент регрессии, постройте график и уравнение линейной регрессии, проверьте остатки.

9) Путем еженедельного взятия проб с поля было изучено изменение высоты растений сои y (в см) с возрастом x (в неделях):

x	1	2	3	4	5	6	7
y	5	13	16	23	33	38	40

Определите коэффициент регрессии, постройте график и уравнение линейной регрессии, проверьте остатки.

10) Для установления связи между содержанием фосфора в почве x и содержанием фосфора в злаковых растениях y было проведено 9 анализов со следующими результатами:

x	1	4	5	9	13	11	23	23	28
y	64	71	54	81	93	76	77	95	109

Определите коэффициент регрессии, постройте график и уравнение линейной регрессии, проверьте остатки.