

5.1. Введение в линейную регрессию

В статистическом моделировании регрессионный анализ — это набор статистических процедур для изучения зависимостей между случайными переменными. Он включает в себя множество методов моделирования и анализа взаимосвязей между зависимой переменной и одной или несколькими независимыми переменными, называемых также предикторами или регрессорами.

Регрессионный анализ помогает понять, как «типичное» значение зависимой переменной изменяется при изменении одной из независимых переменных, в то время как другие независимые переменные остаются фиксированными.

В настоящее время разработано много методов регрессионного анализа. Наиболее популярными из них являются простая и множественная линейная регрессия.

Достоинства линейной регрессии:

- *скорость и простота получения модели;*
- *интерпретируемость модели.* Линейная модель является прозрачной и понятной для аналитика. По полученным коэффициентам регрессии можно судить о том, как тот или иной фактор влияет на результат, сделать на этой основе дополнительные полезные выводы;
- *широкая применимость.* Большое количество реальных процессов в экономике и бизнесе можно с достаточной точностью описать линейными моделями;
- *изученность данного подхода.* Для линейной регрессии известны типичные проблемы (например, мультиколлинеарность) и их решения, разработаны и реализованы тесты оценки статической значимости получаемых моделей.

Линейная регрессия является одним из наиболее часто используемых алгоритмов в машинном обучении. Этот алгоритм зачастую дает хороший результат даже на небольших наборах данных.

Широкое применение линейной регрессии обусловлено тем, что большое количество реальных процессов в науке, экономике и бизнесе можно описать линейными моделями. Так, с помощью линейной регрессии можно оценивать объем ожидаемых продаж в зависимости от установленной цены.

Линейная регрессия может использоваться для решения различных задач Data Mining, например, таких, как прогнозирование и численное предсказание.

Линейная регрессия строится путем аппроксимации линейной зависимости между входными и выходными переменными. Если ищется связь между одной входной и одной выходной переменными, то имеет место простая линейная регрессия. Для этого определяется уравнение регрессии $y = ax + b$ и строится соответствующая прямая, известная как линия регрессии. Коэффициенты a и b , называемые также параметрами модели, выбираются таким образом, чтобы

сумма квадратов отклонений точек, соответствующих реальным наблюдениям данных, от линии регрессии, была бы минимальной.

Если ищется зависимость между несколькими входными и одной выходной переменными, то имеет место множественная линейная регрессия. Соответствующее уравнение имеет вид:

$$y = a_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n,$$

где n – число входных переменных.

Очевидно, что в данном случае модель будет описываться не прямой, а гиперплоскостью. Коэффициенты уравнения множественной линейной регрессии подбираются так, чтобы минимизировать сумму квадратов отклонения реальных точек данных от этой гиперплоскости.

Построение линейной регрессии заключается в расчете её коэффициентов методом наименьших квадратов. Мерой рассогласования между фактическими значениями, и значениями, оцененными моделью в методе наименьших квадратов, служит сумма квадратов разностей между ними, то есть:

$$\sum_{i=1}^n (y - \hat{y}_x)^2 \rightarrow \min,$$

где \hat{y}_x — оценка, полученная с помощью модели;
 y — фактическое наблюдаемое значение.

Очевидно, что лучшей будет та модель, которая минимизирует данную сумму (откуда и название метода).

Для отбора переменных в модель линейной регрессии могут использоваться несколько методов:

- *принудительное включение (Enter)* — включение в регрессионную модель всех заданных признаков независимо от того, оказывают ли они значимое влияние или нет;
- *пошаговое включение (Forward)* — метод, который базируется на принципе: начать с отсутствия признаков и постепенно найти самые «лучшие», которые будут добавлены в подмножество;
- *пошаговое исключение (Backward)* — метод основан на следующем: начать со всех доступных признаков и последовательными итерациями исключать самые «худшие»;
- *пошаговое включение/исключение (Stepwise)* — модификация метода Forward с тем отличием, что на каждом шаге после включения новой переменной в модель осуществляется проверка на значимость остальных переменных, которые уже были введены в нее ранее;
- *Ridge* — один из методов понижения размерности. Применяется для борьбы с переизбыточностью данных, когда независимые переменные коррелируют друг с другом (мультиколлинеарность), вследствие чего проявляется неустойчивость оценок коэффициентов линейной регрессии;
- *LASSO* — также как и Ridge, применяется для борьбы с переизбыточностью данных;

• *Elastic-Net* — модель регрессии с двумя регуляризаторами $L1$, $L2$. Частными случаями являются модели LASSO $L1 = 0$ и Ridge регрессии $L2 = 0$. Оба регуляризатора помогают улучшить обобщение и ошибки теста, поскольку не допускают переобучения модели из-за шума в данных:

- ✓ $L1$ — реализует это путём отбора наиболее важных факторов, которые сильнее всего влияют на результат;
- ✓ $L2$ — предотвращает переобучения модели путём запрета на непропорционально большие весовые коэффициенты.

Для критерия отбора факторов можно выбрать один из следующих информационных критериев:

- F-тест;
- коэффициент детерминации;
- скорректированный коэффициент детерминации;
- информационный критерий Акаике;
- информационный критерий Акаике скорректированный;
- информационный критерий Байеса;
- информационный критерий Ханнана-Квина;
- порог значимости при добавлении фактора;
- порог значимости при исключении фактора.

Несмотря на свою универсальность, линейная регрессионная модель не всегда пригодна для качественного предсказания зависимой переменной. Например, при нелинейной зависимости. В этом случае нелинейная модель сводится к линейной путем линеаризации переменных (табл. 3.1).

Таблица 5.1. Линеаризация нелинейных моделей

Регрессионная модель	Исходное уравнение	Преобразованное уравнение
Гипербола	$y = a_0 + \frac{a_1}{x}$	$y = a_0 + a_1 \frac{1}{x}$
Квадратичная	$y = a_0 + a_1x + a_2x^2$	$y = a_0 + a_1x_1 + a_2x_2$
Кубическая	$y = a_0 + a_1x + a_2x^2 + a_3x^3$	$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3$
Показательная	$y = a_0a_1^x$	$\ln y = \ln a_0 + \ln x$
Степенная	$y = a_0x^{a_1}$	$\ln y = \ln a_0 + a_1 \ln x$
Экспоненциальная	$y = a_0e^{a_1x}$	$\ln y = \ln a_0 + a_1x$

Кроме того, если выходная переменная является категориальной или бинарной, приходится использовать различные модификации регрессии. Одной из таких модификаций является логистическая регрессия, предназначенная для оценки вероятности того, что зависимая переменная примет значение 0 или 1.

5.2. Методические указания

5.2.1. Множественная регрессия

В файле *Задача 5.1. Услуги связи.xlsx* имеются данные об объеме услуг связи, оказанных населению, и о тех факторах, которые могут влиять на них (рис. 5.1).

	A	B	C	D	E	F	G
	№ п/п	Регион	Объем услуг связи, оказанных населению, на одного жителя, руб. (y)	Наличие квартирных телефонных аппаратов сети общего пользования на 1000 человек населения (x_1)	Число подключенных абонентских устройств подвижной радиотелефонной связи на 1000 человек населения (x_2)	Число активных абонентов фиксированного широкополосного доступа к сети Интернет на 100 человек населения (x_3)	Число активных абонентов подвижной радиотелефонной связи, использующих широкополосный доступ к сети Интернет на 100 человек населения (x_4)
1							
2	1	Республика Башкортостан	3951,0	135,4	1737,9	20,2	71,5
3	2	Республика Марий Эл	4240,8	138,6	1846,6	17,2	70,3
4	3	Республика Мордовия	4238,1	186,3	1590,6	17,9	58,9
14	13	Саратовская область	4845,9	155,4	1798,3	19,8	70,7
15	14	Ульяновская область	4509,5	170,6	1936,6	19,5	67,3

Рис. 5.1

Требуется установить зависимость объема услуг связи от перечисленных факторов. Уровень вероятности суждения принять 0,95.

Решение

Создадим новый пакет *Линейная регрессия*. Выполним импорт исходных данных. Для этого создадим узел сценария, выполняющий действие импорта (рис. 5.2).

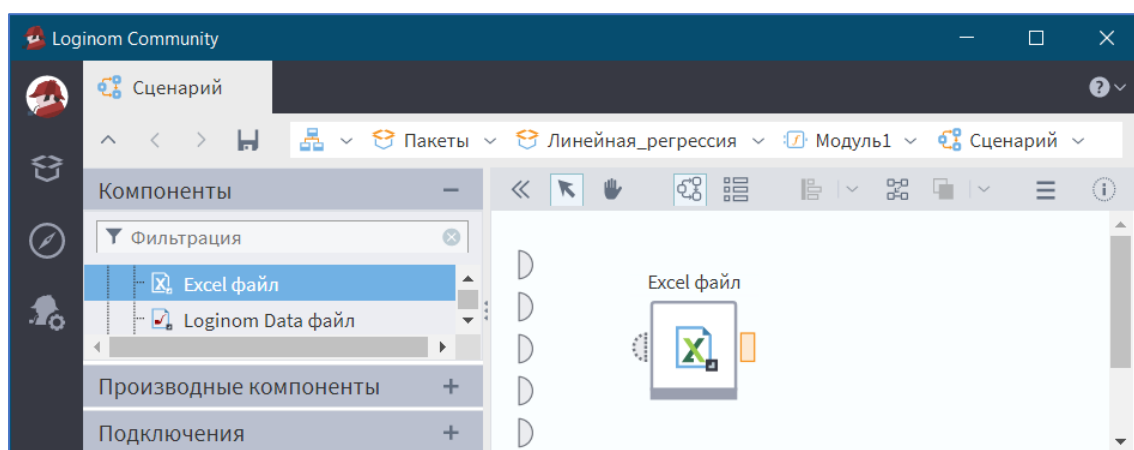


Рис. 5.2

Вызовем *Мастер настройки*. Пройдем шаги мастера, указав в описании узла метку *Задача 5.1. Услуги связи*.

Добавим визуализаторы *Таблица* и *Статистика* к узлу сценария (рис. 5.3).

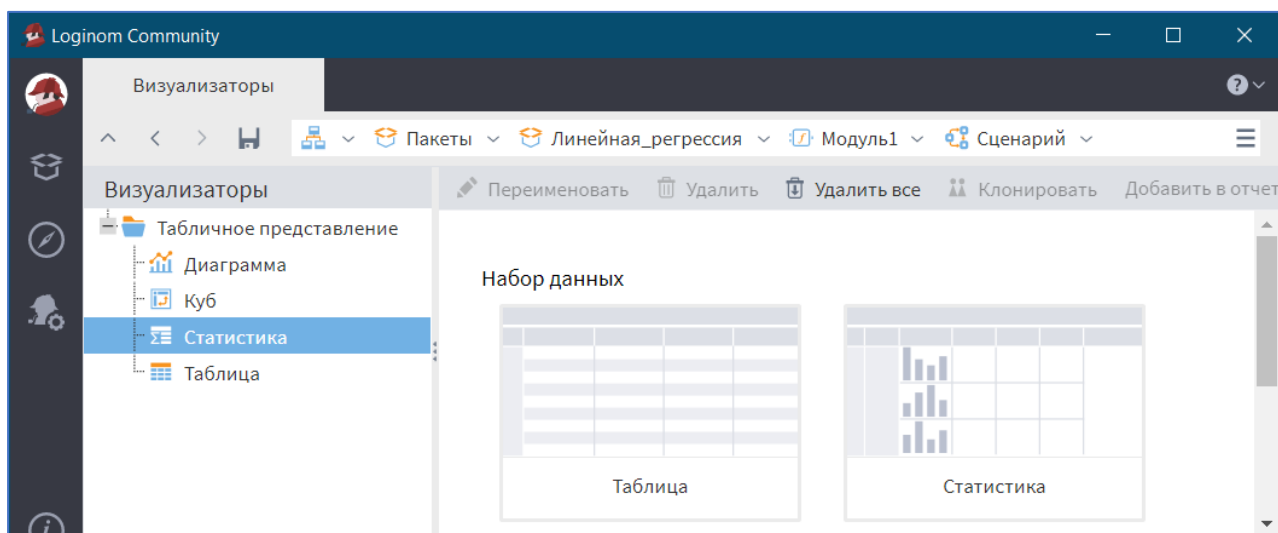


Рис. 5.3

Таблицы с исходными данными и статистическими показателями имеют вид (рис. 5.4–5.5).

#	12 №..	ab Регион	9.0 Объем...	9.0 Налич...	9.0 Число ...	9.0 Число ...	9.0 Число ...
1	1	Республика Ба...	3951	135,4	1737,9	20,2	71,5
2	2	Республика Мар...	4240,8	138,6	1846,6	17,2	70,3
3	3	Республика Мор...	4238,1	186,3	1590,6	17,9	58,9
4	4	Республика Тат...	5253,2	175,3	1850,8	24,5	73,9
14	5	Удмуртская Рес...	4161,9	117,5	1763,4	19,4	66,6

Рис. 5.4

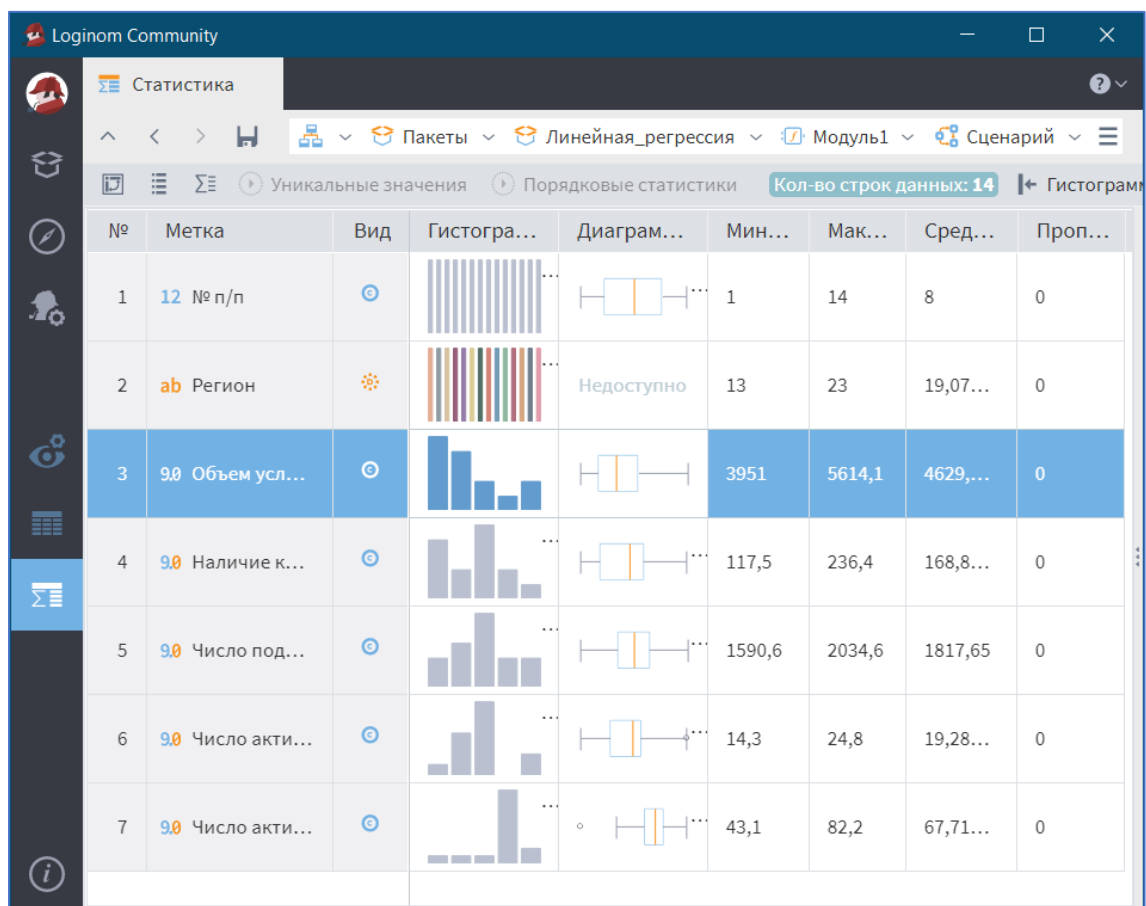


Рис. 5.5

Построим линейную модель на основе импортированных данных. Для этого переместим компонент *Линейная регрессия* в рабочую область сценария. Последовательность обработки данных задается соединением выходного порта узла импорта с входным портом линейной регрессии (рис. 5.6).

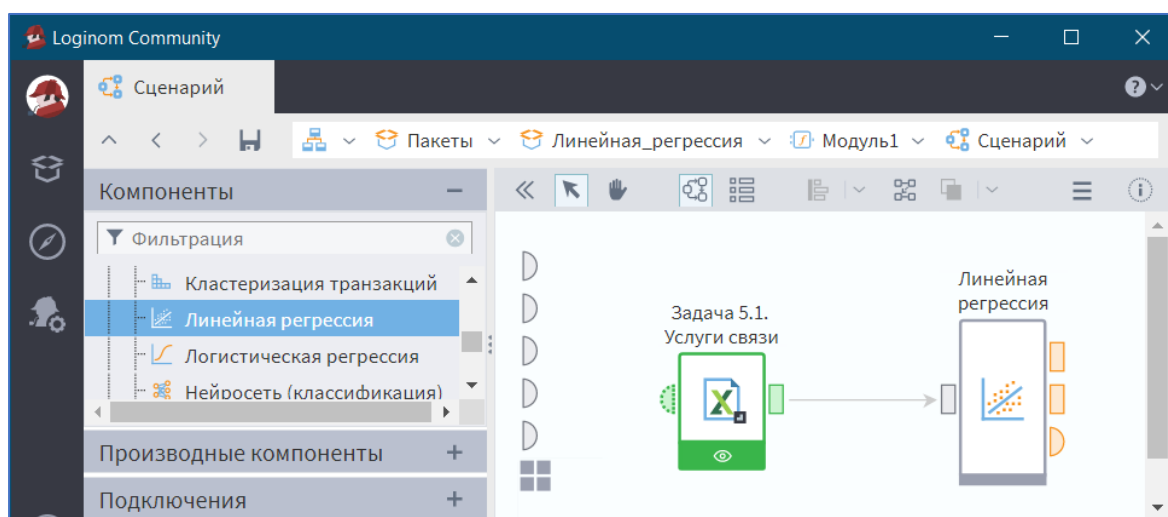


Рис. 5.6

Пройдем шаги *Мастера настройки*. На шаге *Настройка входных столбцов* настроим назначение исходных столбцов данных. Столбец *Объем услуг связи*,

оказанных населению, на одного жителя, руб. зададим как выходной, остальные столбцы — как входные (рис. 5.7).

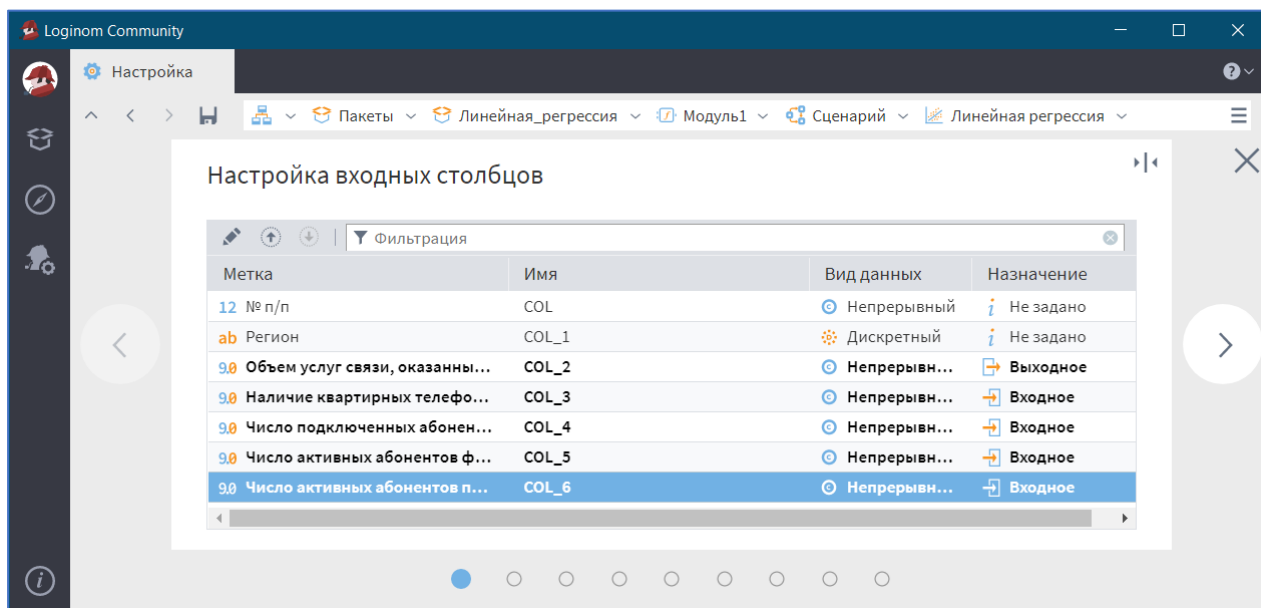


Рис. 5.7

На шагах *Настройка нормализации* и *Разбиение на множества* оставим стандартные параметры по умолчанию (рис. 5.8–5.9).

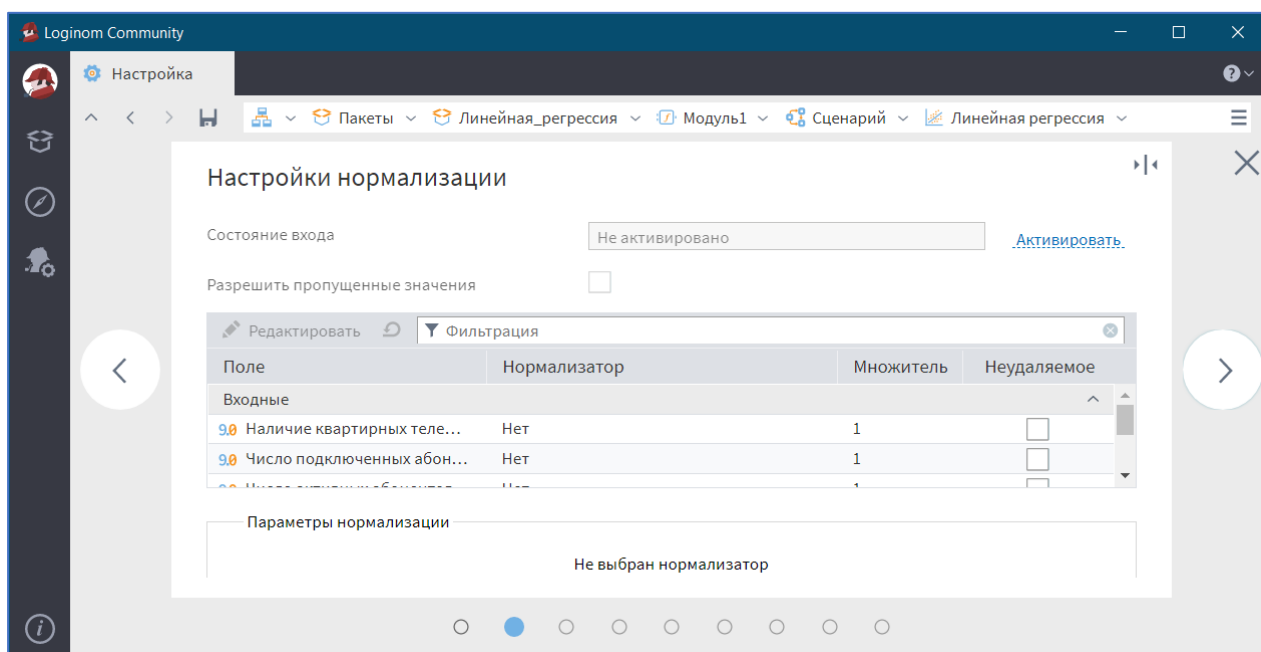


Рис. 5.8

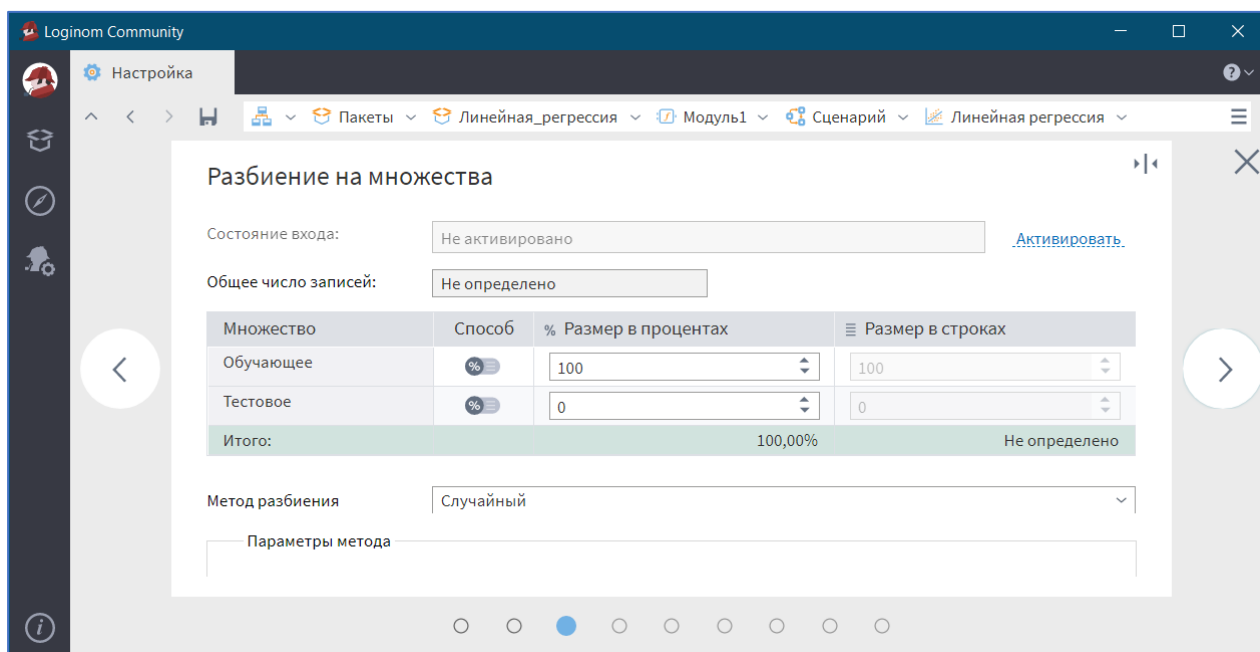


Рис. 5.9

На шаге *Настройка линейной регрессии* снимем флажок с параметра *Автоматическая настройка*, выберем в качестве метода отбора факторов и защиты от переобучения *Пошаговое исключение* и установим флажок на параметре *Использовать детальные настройки* (рис. 5.10).

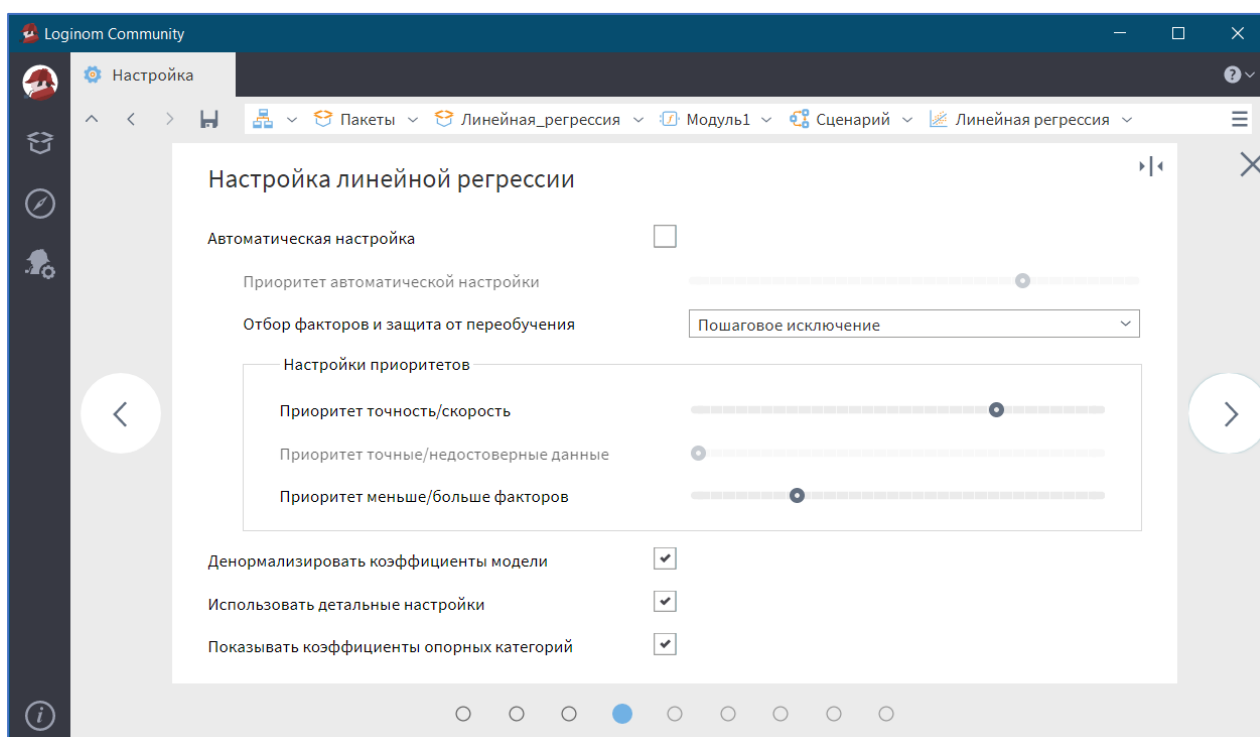


Рис. 5.10

На шаге *Детальная настройка линейной регрессии* установим флажок на параметре *Рассчитать доверительный интервал*, в качестве критерия отбора

факторов выберем *F-тест* и установим *Порог значимости при исключении фактора* — 0,05 (рис. 5.11).

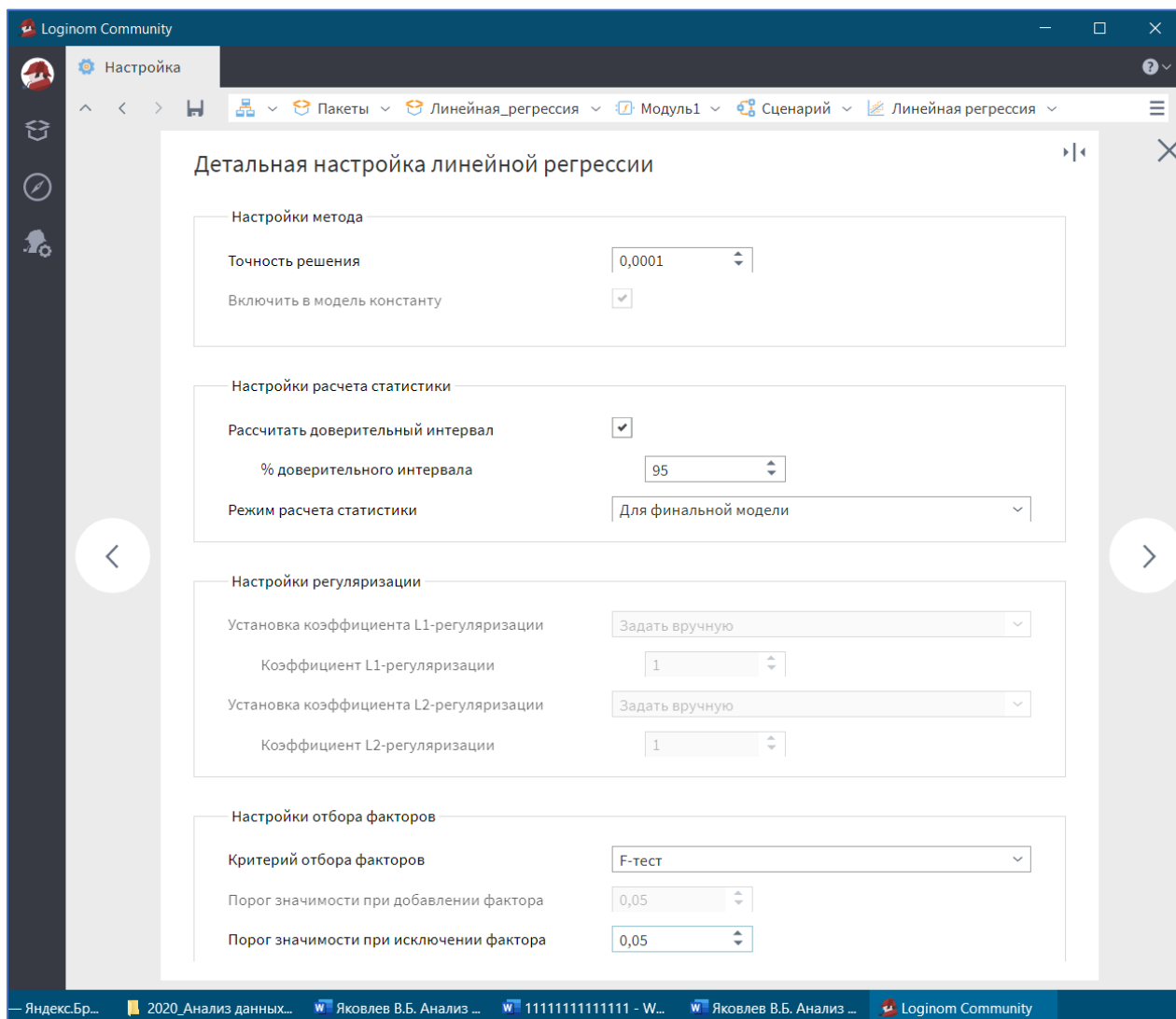


Рис. 5.11

На шаге *Описание узла* оставим стандартные параметры по умолчанию (рис. 5.12).

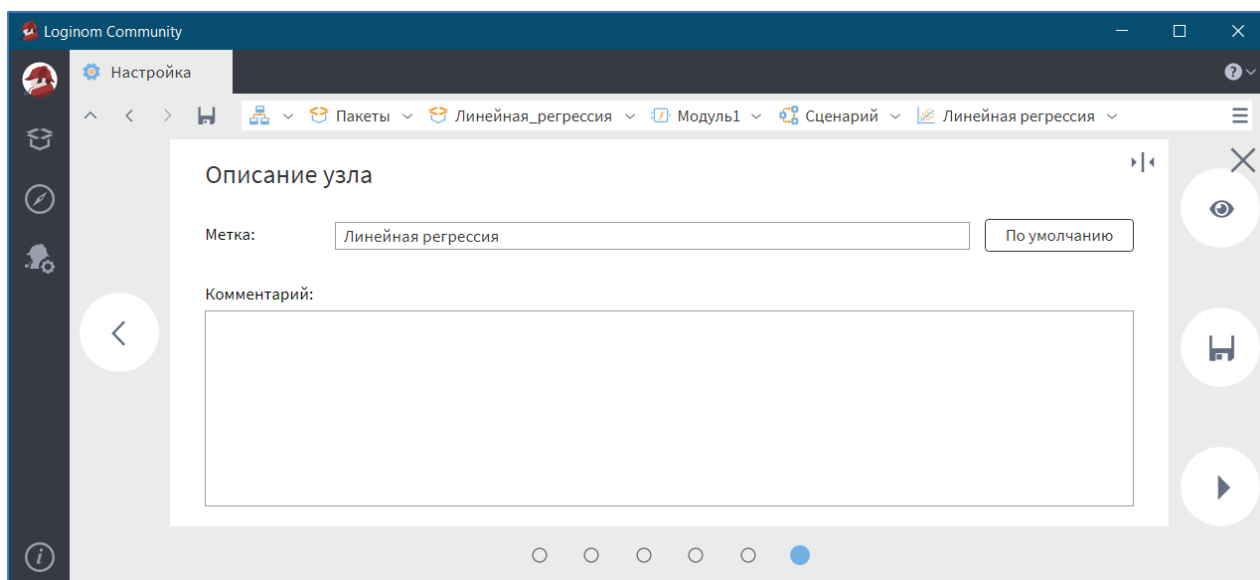


Рис. 5.12

Переобучим узел *Линейная регрессия* (рис. 5.13) и перейдем к настройкам визуализаторов (рис. 5.14).

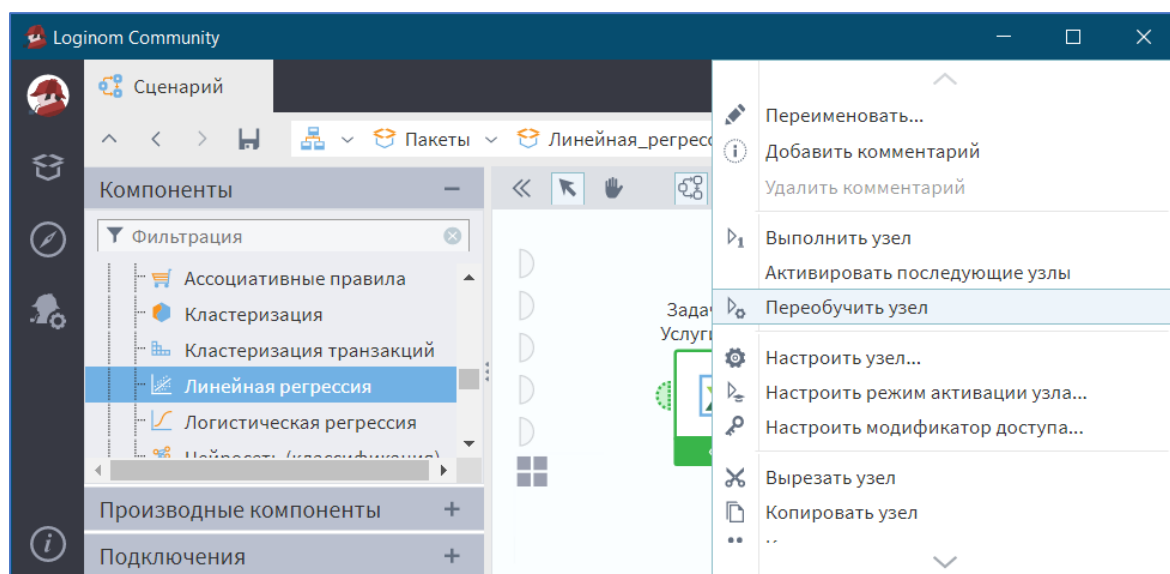


Рис. 5.13

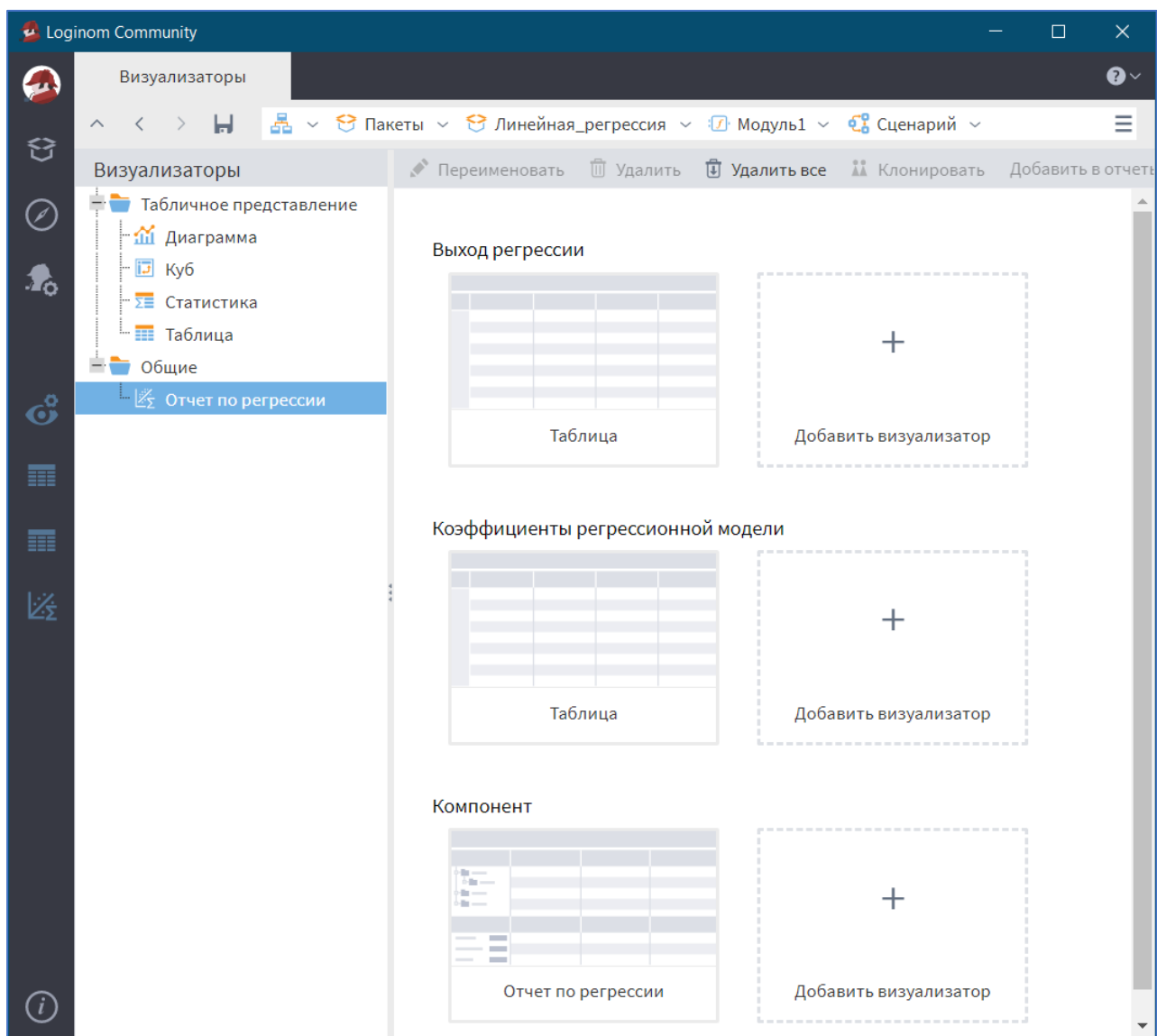


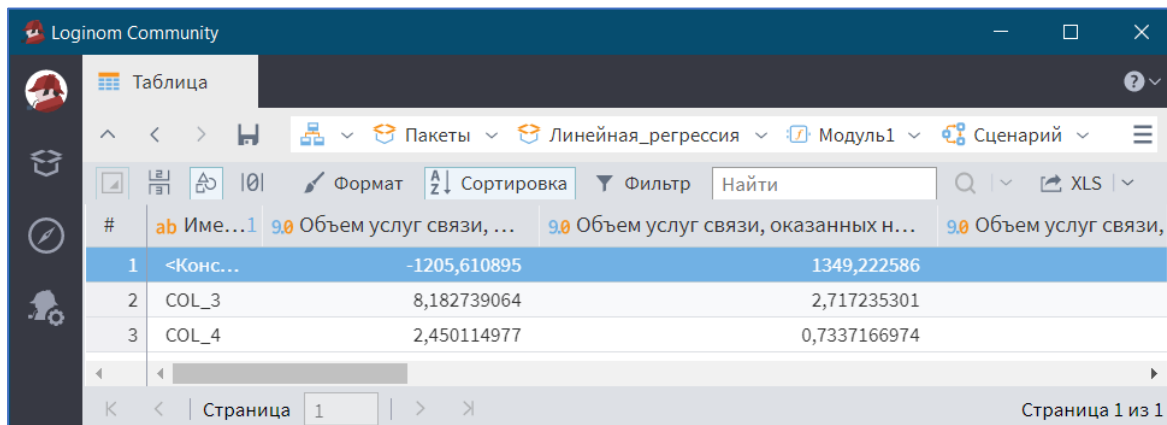
Рис. 5.14

В визуализаторе *Выход регрессии* представлены расчетные значения результативного признака (рис. 5.15).

#	12 № ...	ab Регион	9.8 Объем услуг связи, оказанных населению, на одного ...
1	1	Республика Башкортостан	4160,4
2	2	Республика Марий Эл	4452,9
3	3	Республика Мордовия	4216,0
4	4	Республика Татарстан	4763,5
5	5	Удмуртская Республика	4076,4

Рис. 5.15

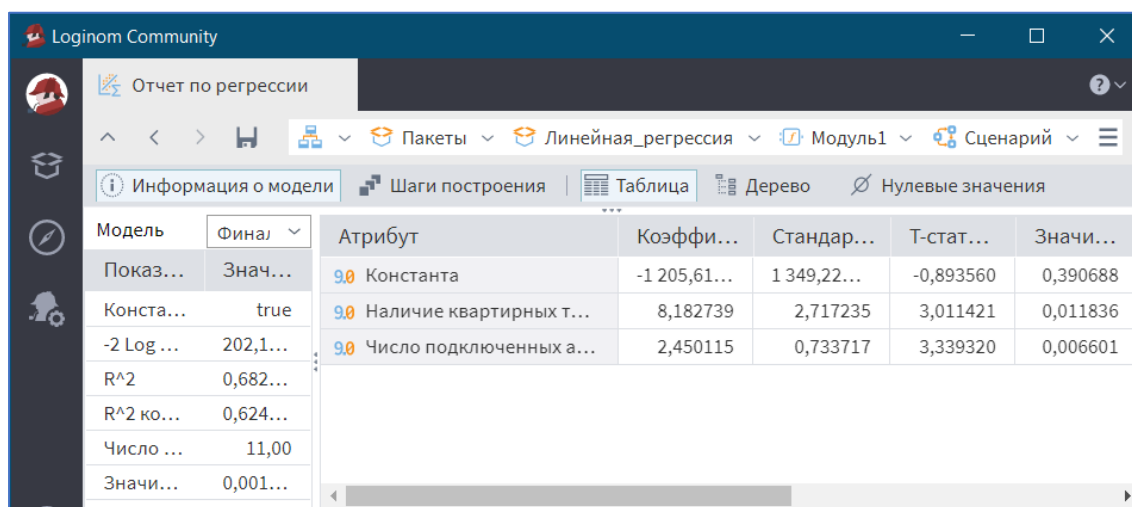
В визуализаторе *Коэффициенты регрессионной модели* приведены регрессионные коэффициенты, их интервальная оценка и уровень статистической значимости (рис. 5.16).



#	Имя	Коэффициент	Интервальная оценка
1	<Конс...	-1205,610895	1349,222586
2	COL_3	8,182739064	2,717235301
3	COL_4	2,450114977	0,7337166974

Рис. 5.16

Визуализатор *Отчет по регрессии* отображает параметры и результаты статистических тестов для анализа регрессионных моделей. Откроем данный визуализатор (рис. 5.17).



Модель	Финал	Атрибут	Коэффи...	Стандар...	Т-стат...	Значи...
Показ...	Знач...	9.0 Константа	-1 205,61...	1 349,22...	-0,893560	0,390688
Конста...	true	9.0 Наличие квартирных т...	8,182739	2,717235	3,011421	0,011836
-2 Log ...	202,1...	9.0 Число подключенных а...	2,450115	0,733717	3,339320	0,006601
R^2	0,682...					
R^2 ко...	0,624...					
Число ...	11,00					
Значи...	0,001...					

Рис. 5.17

Общее качество регрессионной модели определяется с помощью коэффициента детерминации R^2 . В отчете по регрессии *Информация о модели* его величина составляет 0,682. Поскольку максимальное значение $R^2 = 1$, то можно утверждать, что качество регрессионной модели весьма высокое (предлагаемая модель объясняет около 68,2 % дисперсии результативной переменной). Здесь же представлены данные дисперсионного анализа, из которых следует статистическая значимость модели в целом при уровне значимости 0,05.

В отчете по регрессии *Таблица* приведены регрессионные коэффициенты и уровень их статистической значимости.

Как следует из отчета, построенная регрессионная модель имеет вид:

$$\hat{y}_x = -1205,61 + 8,1827x_1 + 2,4501x_2.$$

Все параметры уравнения значимы при заданном уровне значимости 0,05.

Таким образом на объем услуг связи, оказанных населению, на одного жителя оказывают существенное влияние наличие квартирных телефонных аппаратов сети общего пользования и число подключенных абонентских устройств подвижной радиотелефонной связи на 1000 человек населения. Так, увеличение количества квартирных телефонных аппаратов сети общего пользования на 1000 человек населения на 1 шт. позволяет повысить объем услуг связи, оказанных населению, на одного жителя на 8,18 руб., а числа подключенных абонентских устройств подвижной радиотелефонной связи на 1000 человек населения на 1 шт. – на 2,45 руб.

Полученное уравнение регрессии, кроме оценки влияния отдельных факторов на объем услуг связи, оказанных населению, на одного жителя позволяет прогнозировать их в зависимости от величины данных факторов. При этом факторы, влияющие на объем услуг связи, оказанных населению, должны находиться в пределах их изменения в исходной выборочной совокупности.

5.2.2. Парная нелинейная регрессия

В файле *Задача 5.2. Заработная плата.xlsx* имеются данные о заработной плате рабочих и их возрасте (рис. 5.18).

	A	B	C
1	Табельный номер рабочего	Заработная плата в месяц, руб.	Возраст, лет
2	70	47500	25
3	17	44840	24
4	20	47690	43
16	26	47880	26
17	40	55309	28

Рис. 5.18

Требуется определить зависимость заработной платы от возраста рабочих.

Решение

Выполним импорт исходных данных. Для этого создадим узел сценария, выполняющий действие импорта (рис. 5.19).

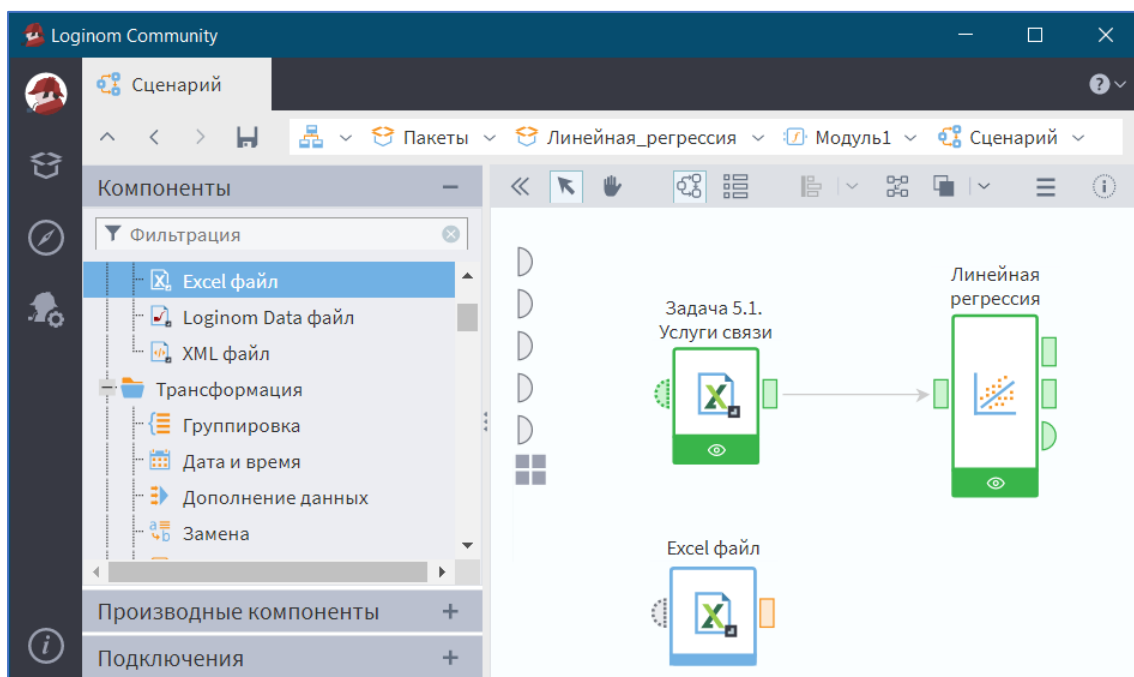


Рис. 5.19

Вызовем *Мастер настройки*. Пройдем шаги мастера. На шаге *Настройка полей* изменим тип данных по второму и третьему столбцу на *Вещественный* (рис. 5.20).

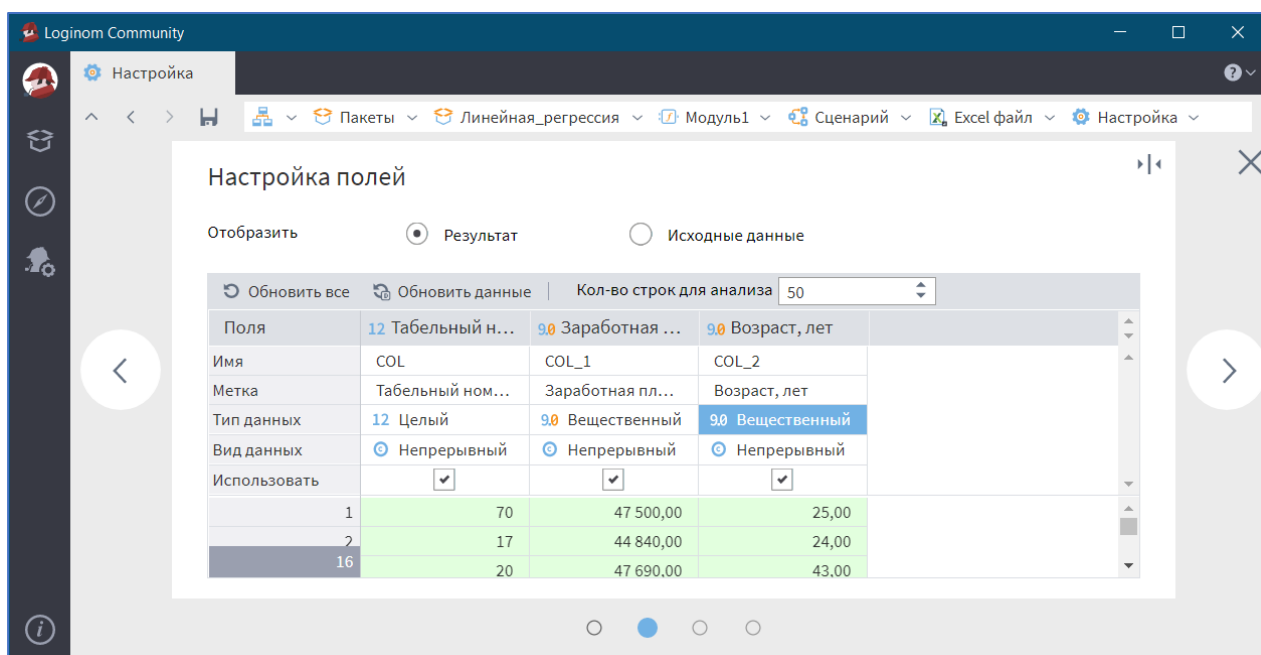


Рис. 5.20

На шаге *Описание узла* укажем метку *Задача 5.2. Зарботная плата* (рис. 5.21).

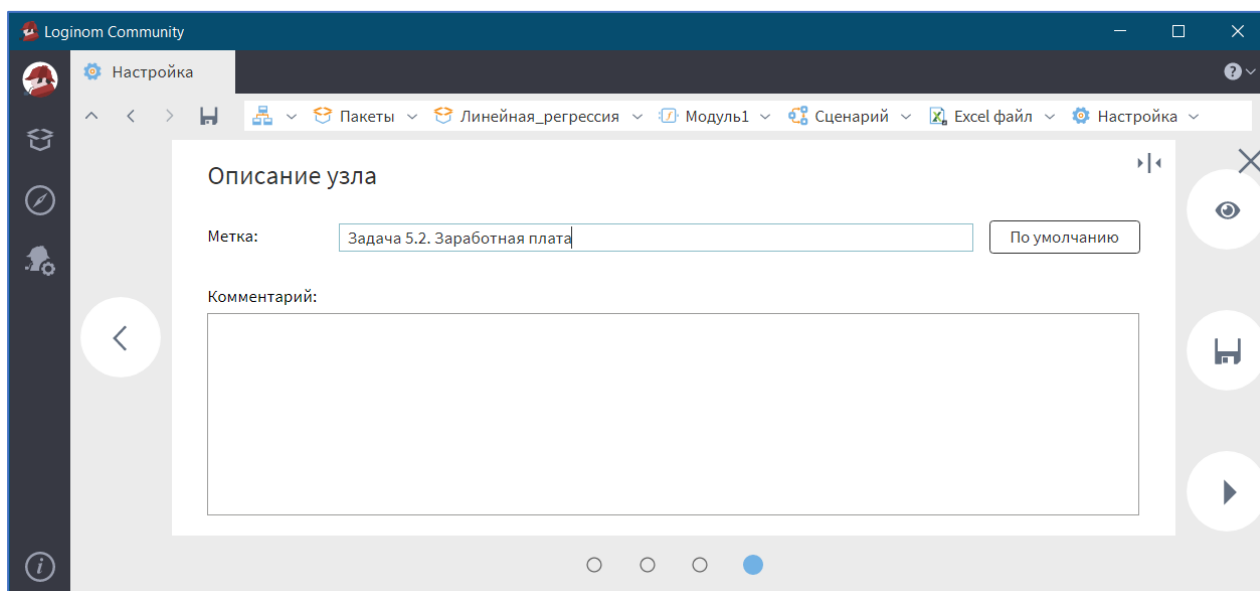


Рис. 5.21

Добавим визуализатор *Таблица* к узлу сценария (рис. 5.22).

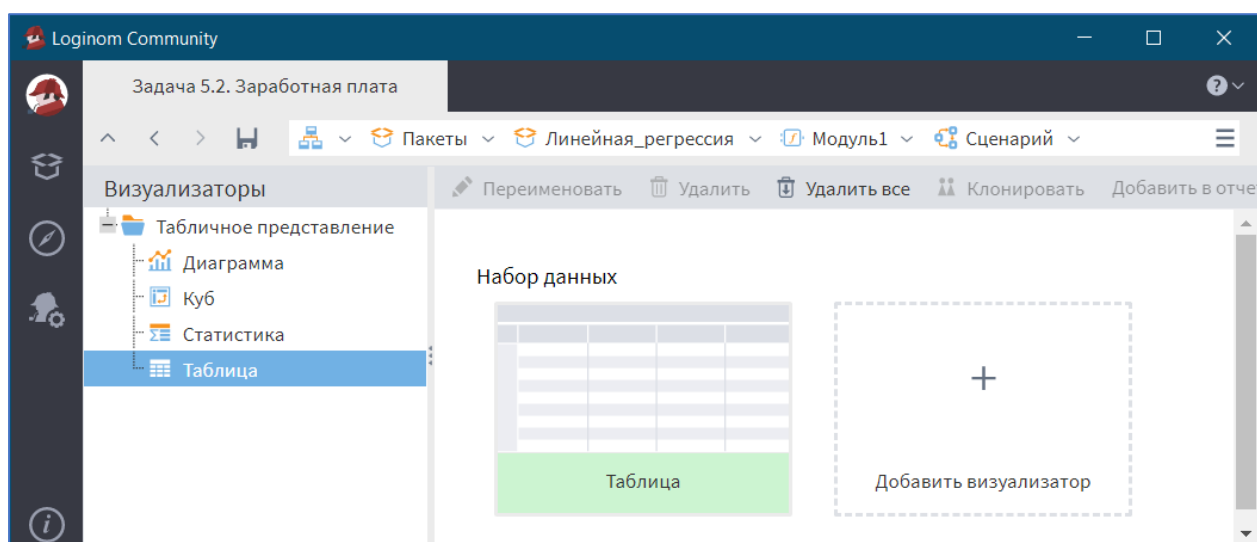


Рис. 5.22

Таблица с исходными данными имеет вид (рис. 5.23).

The screenshot shows the 'Loginom Community' application window. The 'Таблица' (Table) tab is active, displaying a table with the following data:

#	12 Табельный номер рабочего	9.0 Зарплата в месяц, р...	9.0 Возраст, лет
1	70	47500	25
2	17	44840	24
3	20	47690	43
4	35	49590	41
16	44	54720	37

The interface includes a top menu bar with 'Пакеты', 'Линейная регрессия', 'Модуль1', and 'Сценарий'. A toolbar below the menu contains icons for 'Формат', 'Сортировка', 'Фильтр', and a search bar. The bottom status bar indicates 'Страница 1 из 1'.

Рис. 5.23

Для установления направления связи между признаками проведем сортировку данных по столбцу *Возраст, лет*. Для этого переместим компонент *Сортировка* в рабочую область сценария. Последовательность обработки данных задается соединением выходного порта узла импорта с входным портом сортировки (рис. 5.24).

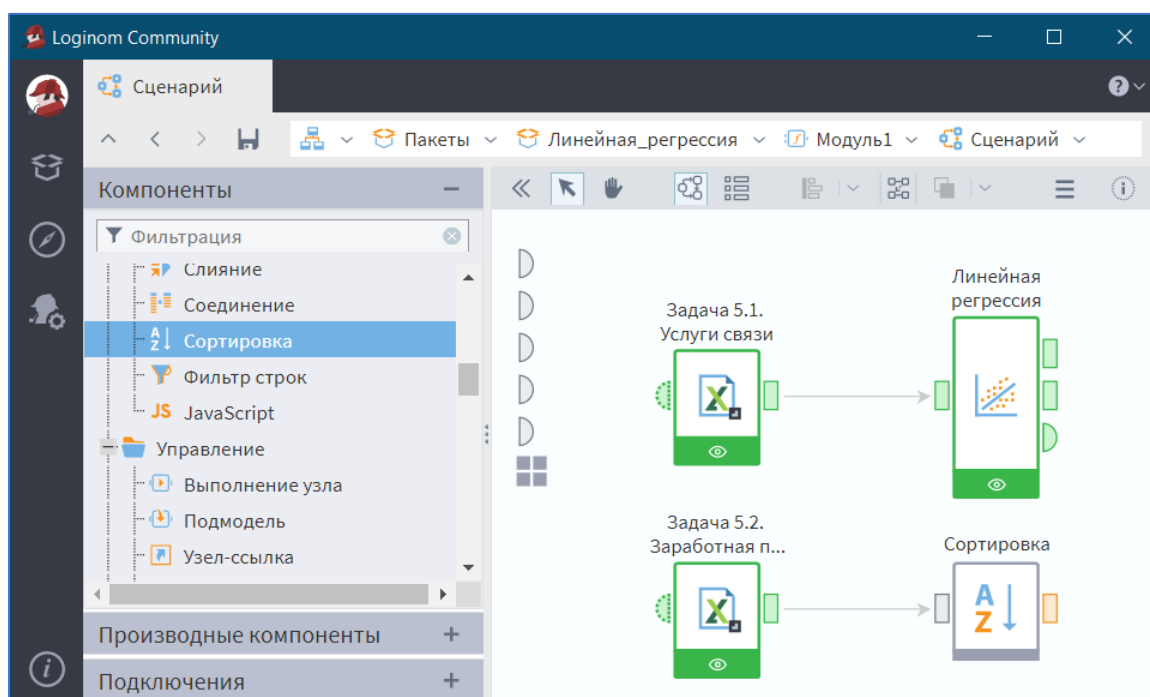


Рис. 5.24

Пройдем шаги *Мастера настройки*. На шаге *Сортировка* установим параметры в соответствии с рис. 5.25.

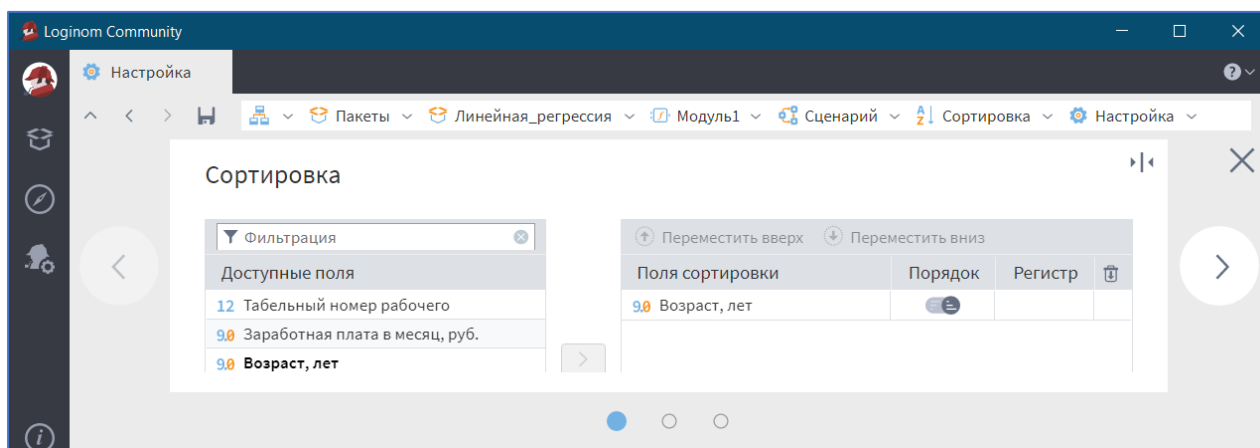


Рис. 5.25

Добавим визуализатор *Диаграмма* к узлу сценария (рис. 5.26).

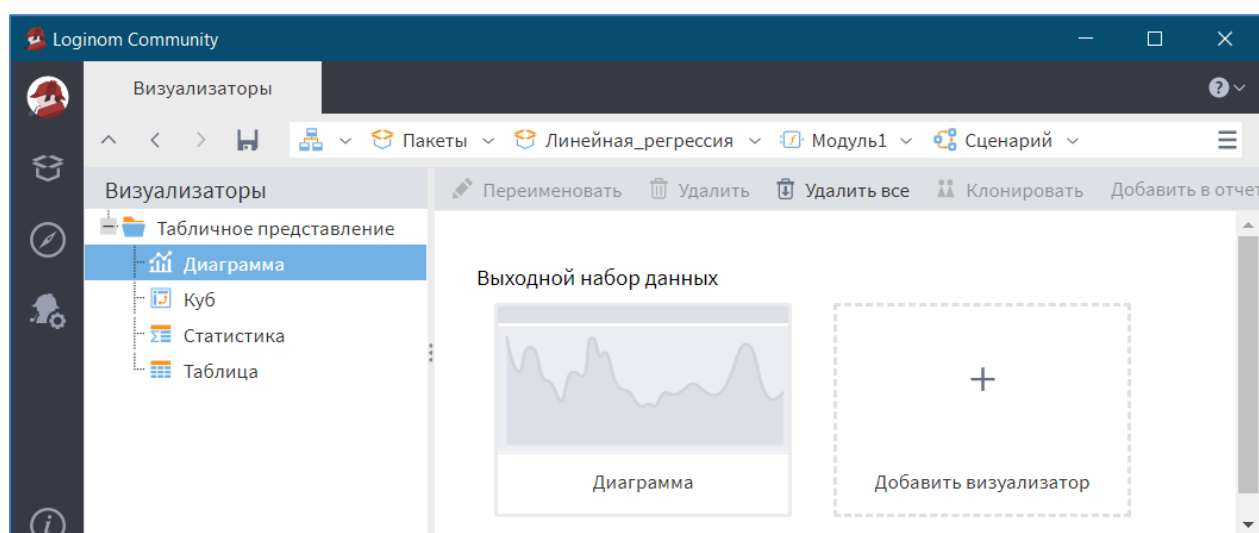


Рис. 5.26

Диаграмма имеет вид (рис. 5.27).

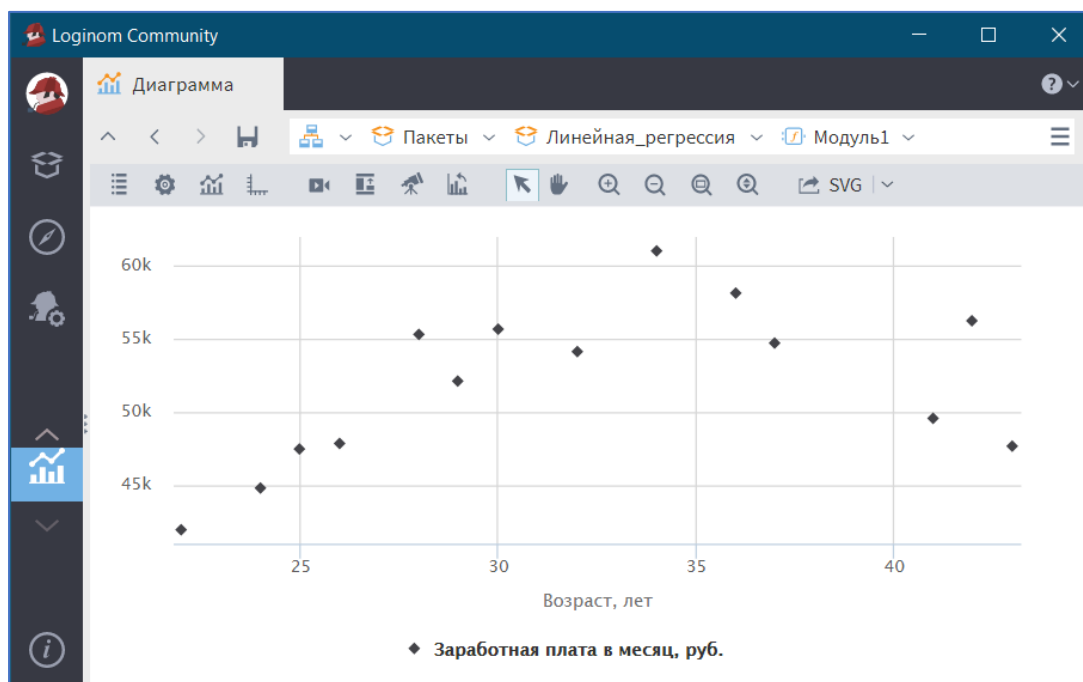


Рис. 5.27

Из данных диаграммы видно, что с увеличением возраста рабочих повышается их заработная плата ввиду одновременного увеличения опыта и повышения квалификации. При этом с определенного возраста ввиду старения организма и снижения производительности труда заработная плата рабочих постепенно снижается. То есть здесь наблюдается параболическая зависимость заработной платы от возраста рабочих. Такую зависимость можно отобразить с помощью модели квадратичной регрессии:

$$y = a_0 + a_1x + a_2x^2.$$

Для решения приведем данную модель к линейному виду путем замены x^2 новой переменной. Уравнение примет вид:

$$y = a_0 + a_1x_1 + a_2x_2.$$

Добавим в импортированные данные новый столбец *Возраст в квадрате*. Для этого переместим компонент *Калькулятор* в рабочую область сценария. Последовательность обработки данных задается соединением выходного порта узла сортировки с входным портом калькулятора (рис. 5.28).

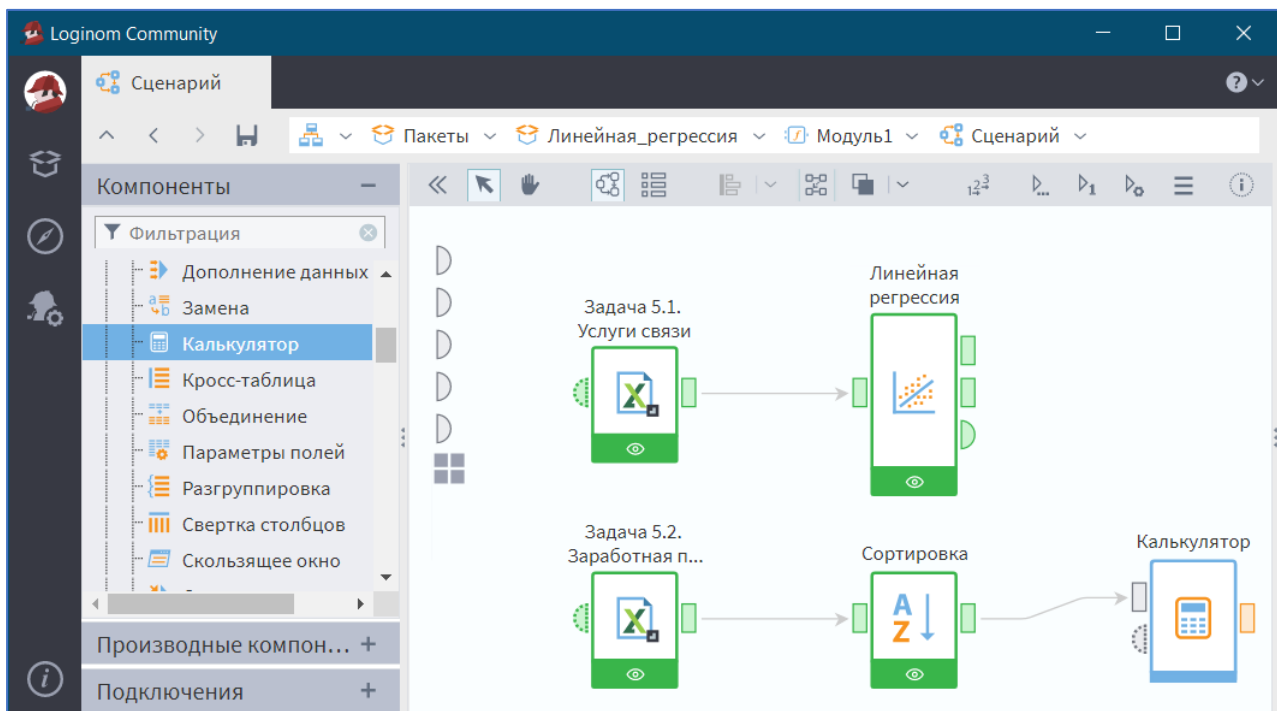


Рис. 5.28

Пройдем шаги *Мастера настройки*. На шаге *Калькулятор* вычислим новый столбец в соответствии с рис. 5.29.

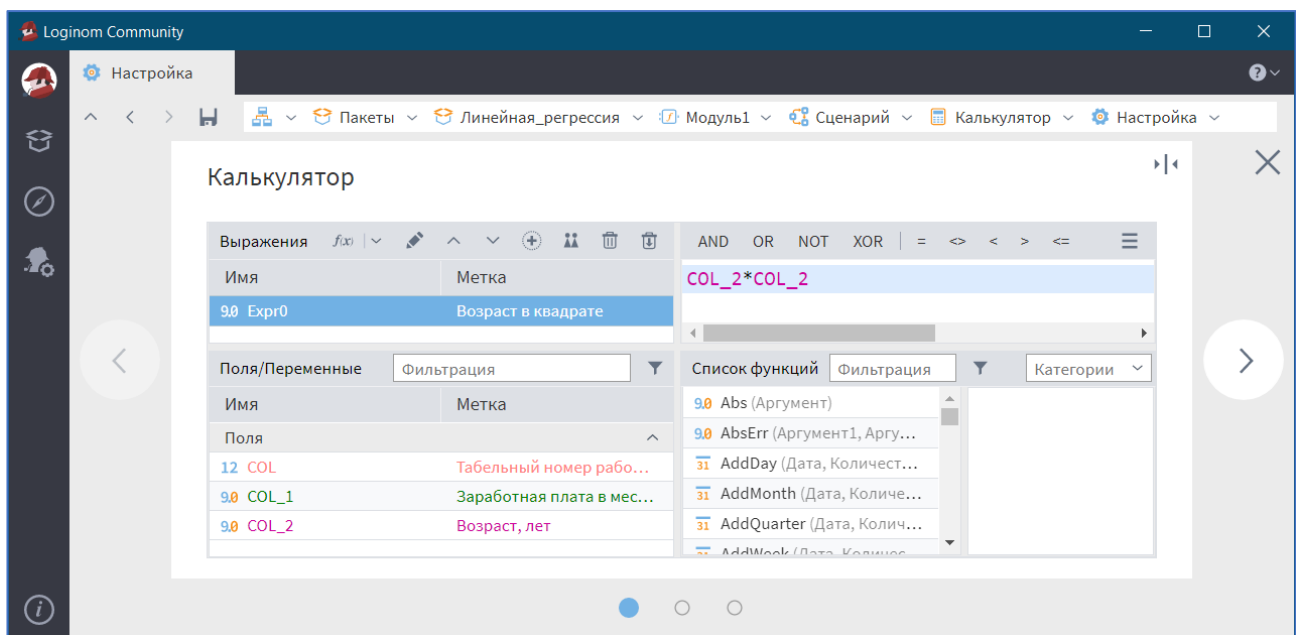


Рис. 5.29

Добавим визуализатор *Таблица* к узлу сценария (рис. 5.30).

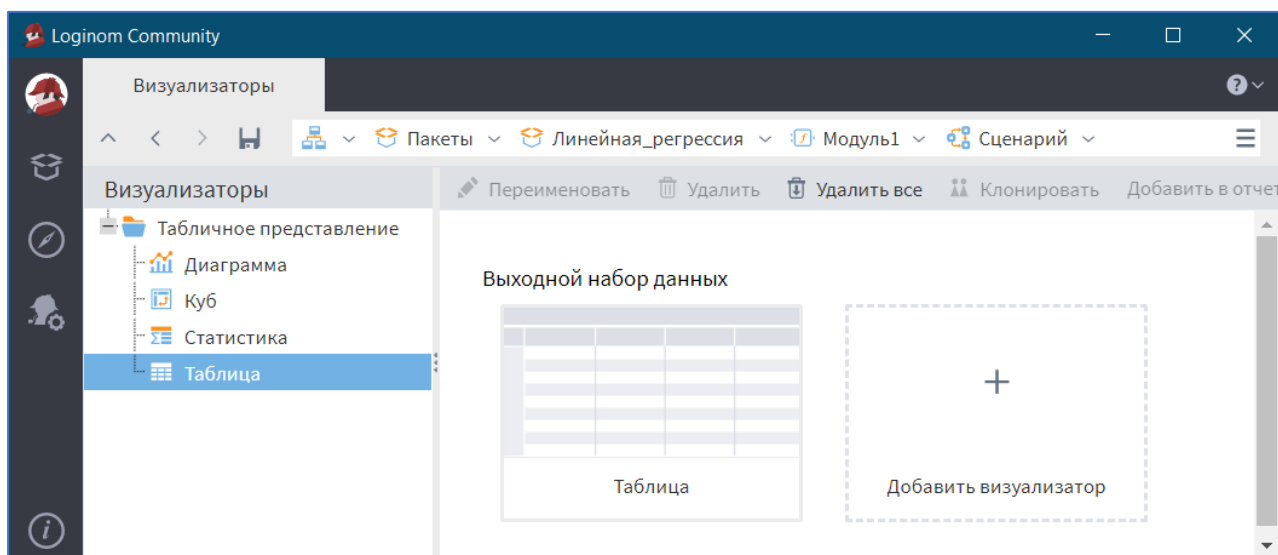


Рис. 5.30

Таблица с преобразованными данными имеет вид (рис. 5.31).

#	12 Табелный но...	9.0 Зарботная плата в меся...	9.0 Возраст, ...	9.0 Возраст в квадр...
1	32	41990	22	484
2	17	44840	24	576
3	70	47500	25	625
4	45	47500	25	625
16	26	47880	26	676

Рис. 5.31

Построим линейную модель на основе преобразованных данных. Для этого переместим компонент *Линейная регрессия* в рабочую область сценария. Последовательность обработки данных задается соединением выходного порта узла калькулятора с входным портом линейной регрессии (рис. 5.32).

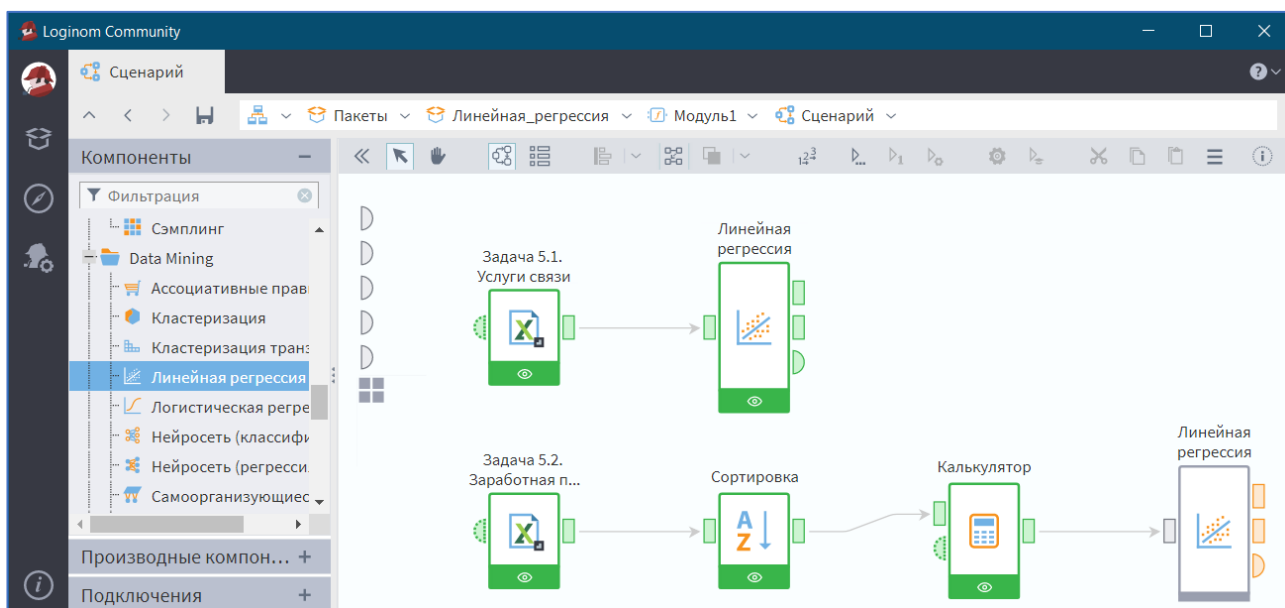


Рис. 5.32

Пройдем шаги *Мастера настройки*. На шаге *Настройка входных столбцов* настроим назначение исходных столбцов данных. Столбец *Заработная плата в месяц, руб.* зададим как выходной, столбцы *Возраст, лет* и *Возраст в квадрате* — как входные (рис. 5.33).

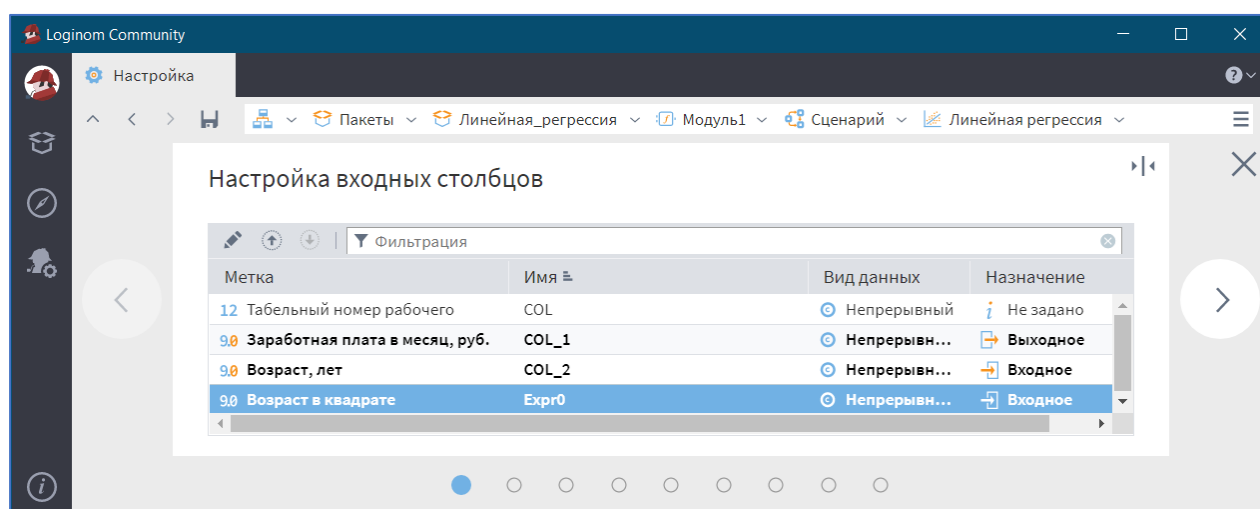


Рис. 5.33

На шагах *Настройка нормализации* и *Разбиение на множества* оставим стандартные параметры по умолчанию (рис. 5.34–5.35).

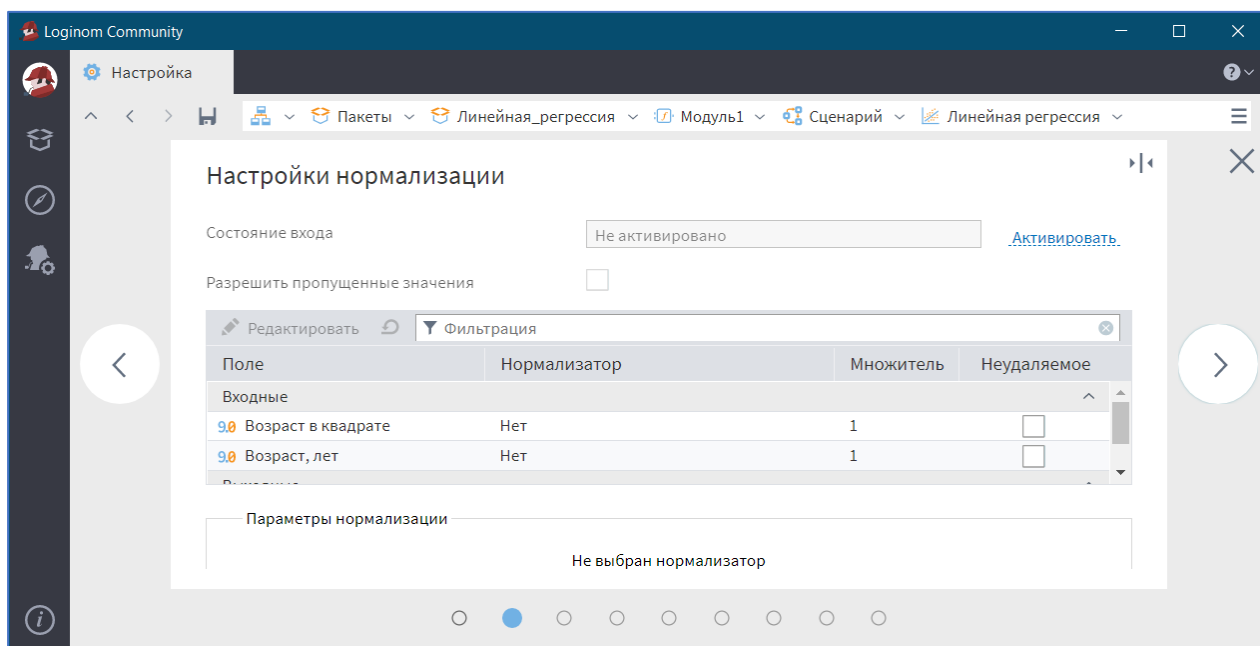


Рис. 5.34

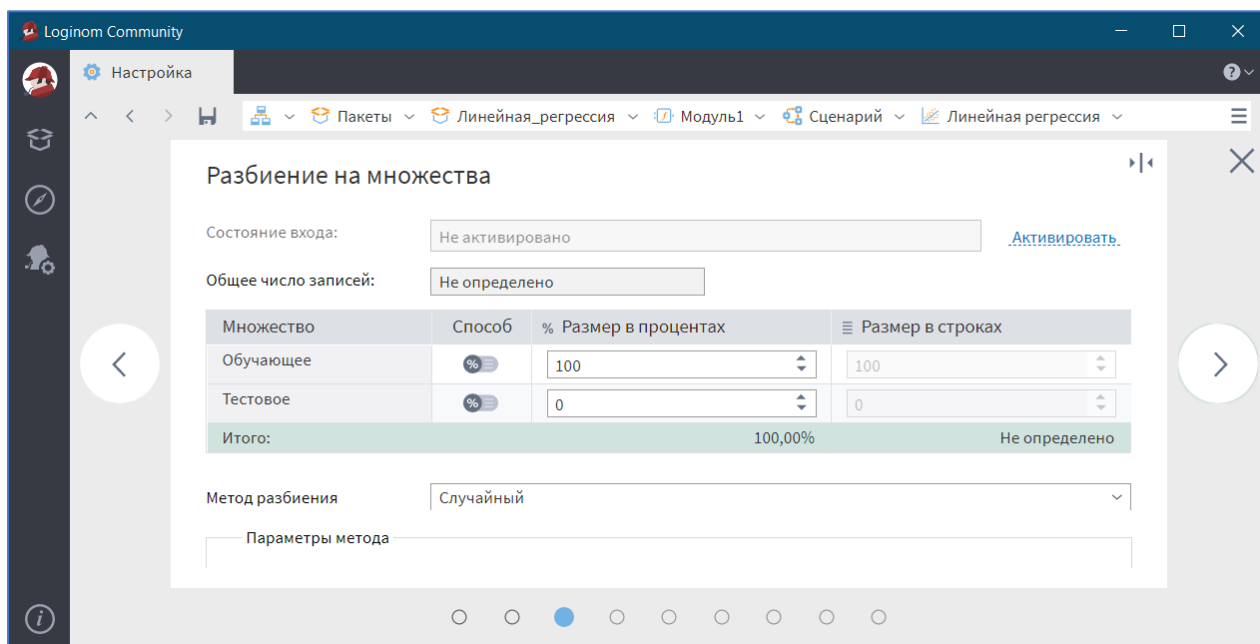


Рис. 5.35

На шаге *Настройка линейной регрессии* снимем флажок с параметра *Автоматическая настройка*, выберем в качестве метода отбора факторов и защиты от переобучения *Принудительное включение* (рис. 5.36).

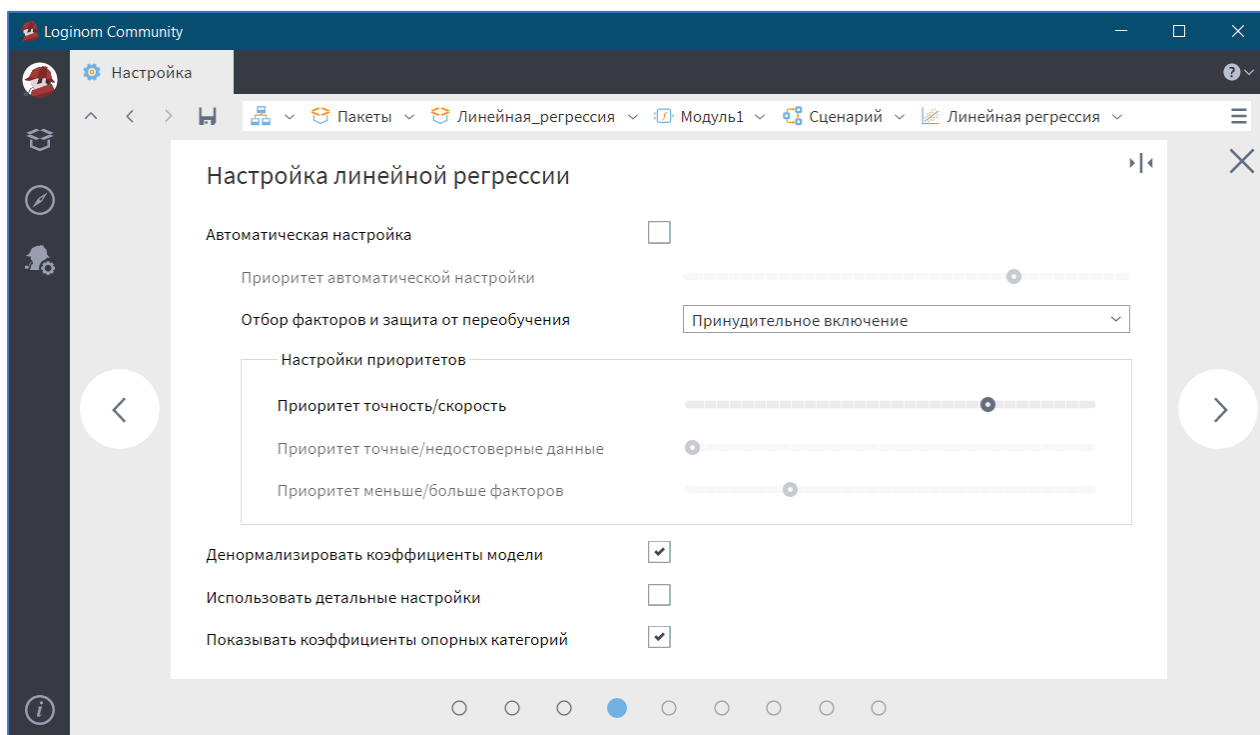


Рис. 5.36

На шаге *Описание узла* оставим стандартные параметры по умолчанию (рис. 5.37).

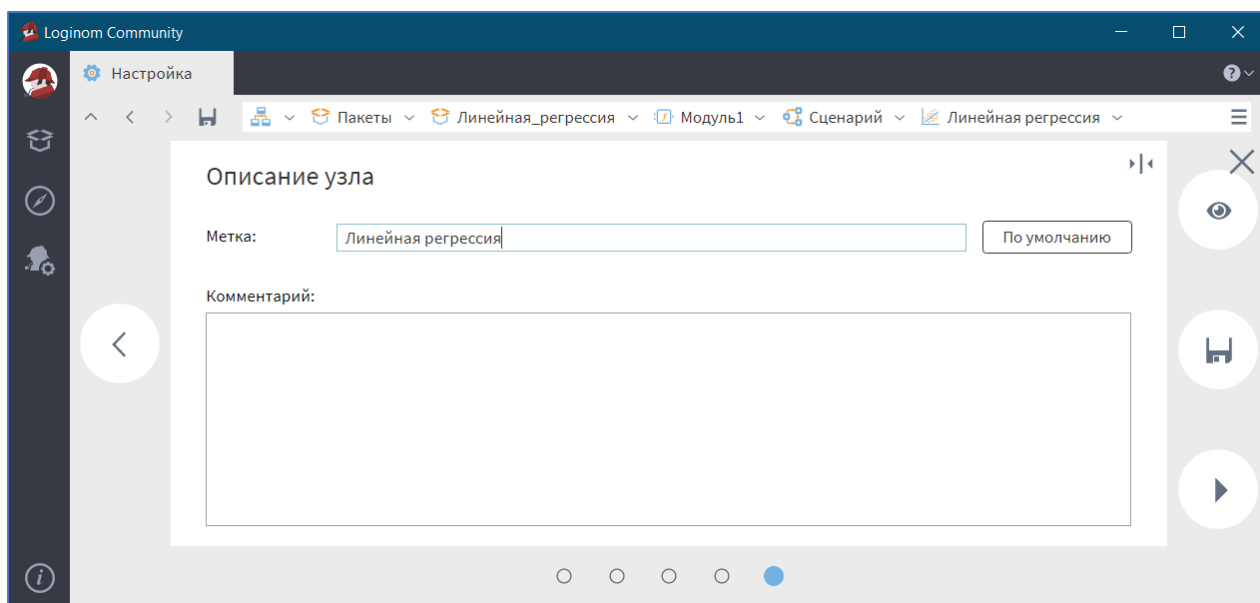


Рис. 5.37

Переобучим узел *Линейная регрессия* и перейдем к настройкам визуализаторов (рис. 5.38).

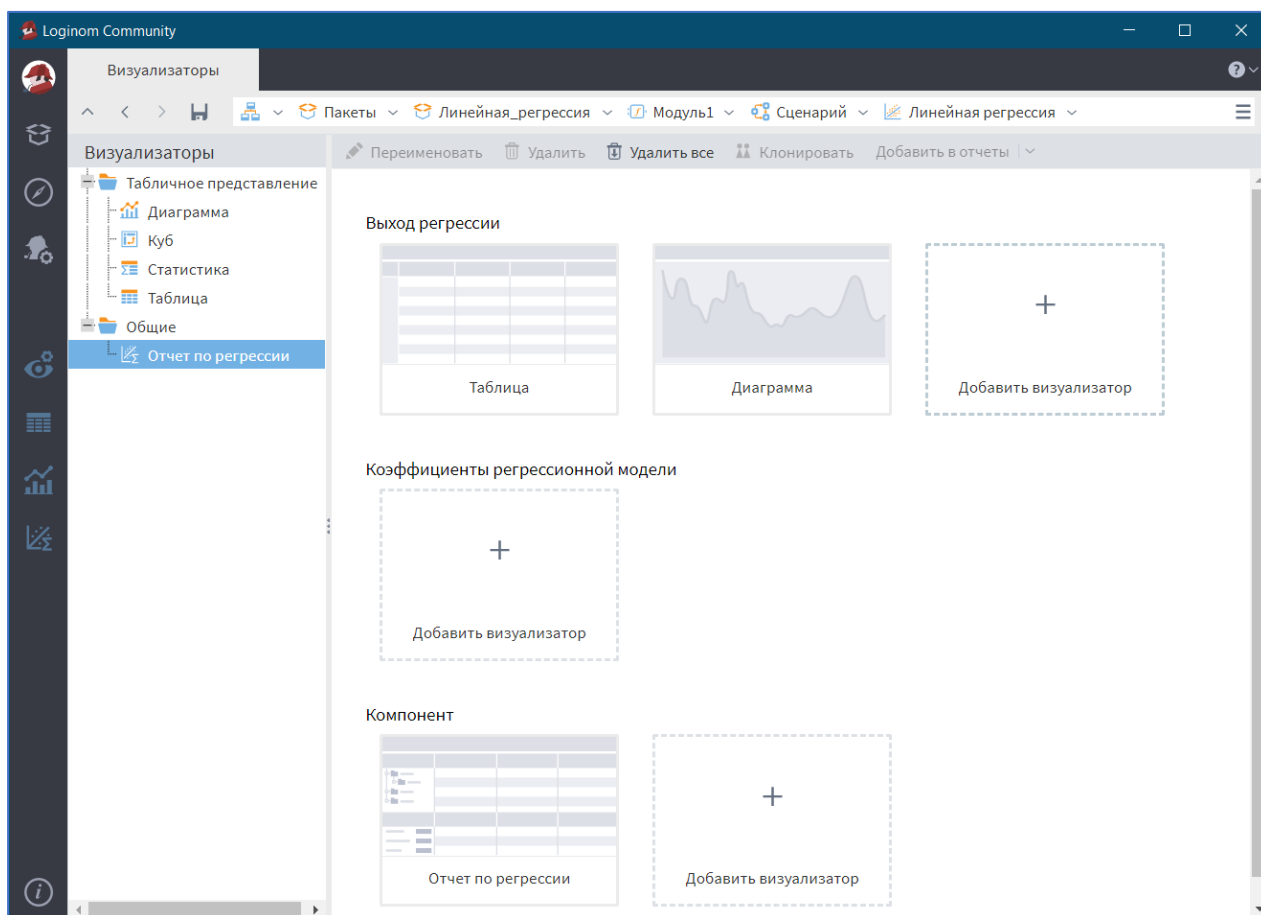


Рис. 5.38

Откроем основной визуализатор *Отчет по регрессии* (рис. 5.39).

Модель	Финг	Атрибут	Коэффи...	Станд...	Т-стат...	Значи...
Пока...	Знач...	9.0 Константа	-73 086,10...	17 995,...	-4,061...	0,001347
Конст...	true	9.0 Возраст в квадрате	-109,608084	17,326...	-6,325...	0,000026
-2 Log...	296,6...	9.0 Возраст, лет	7 570,608...	1 135,6...	6,666547	0,000015
R^2	0,816...					
R^2 к...	0,788...					
Числ...	13,00					
Знач...	0,000...					

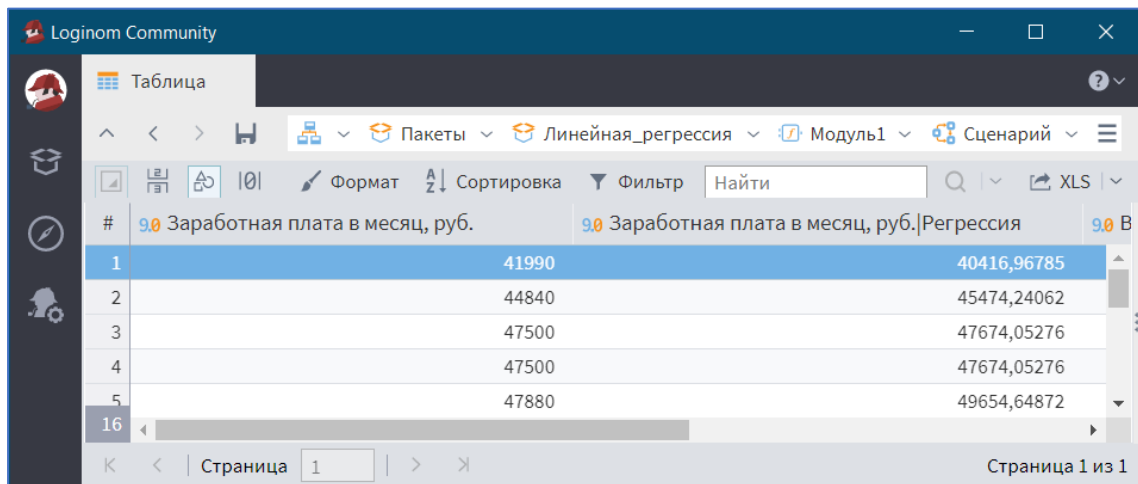
Рис. 5.39

Коэффициент детерминации $R^2 = 0,816$ свидетельствует о высоком качестве регрессионной модели. Полученная модель объясняет около 81,6 % дисперсии результативной переменной. Данные дисперсионного анализа показывают, что модель статистически значима при заданном уровне значимости 0,05.

Построенная квадратичная регрессионная модель имеет вид:

$$y = -73086,1 + 7570,61x - 109,61x^2.$$

Таблица и диаграмма с исходными и рассчитанными по модели данными выводятся в следующем виде (рис. 5.40-5.41).



#	Зарплата в месяц, руб.	Регрессия
1	41990	40416,96785
2	44840	45474,24062
3	47500	47674,05276
4	47500	47674,05276
5	47880	49654,64872
16		

Рис. 5.40

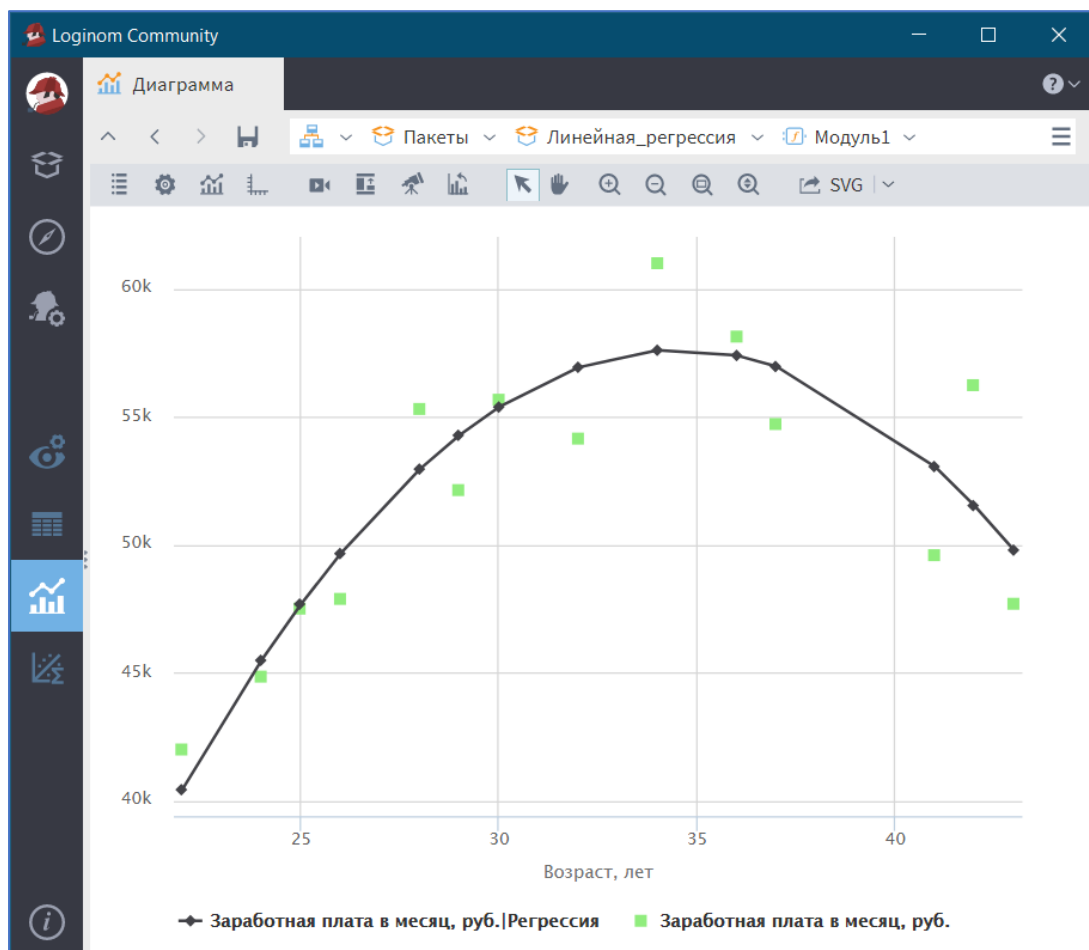


Рис. 5.41

5.3. Задания для самостоятельной работы

Задание 5.1

В файле *Задача 5.3. Денежные доходы.xlsx* имеются данные о среднедушевых денежных доходах населения и о тех факторах, которые могут влиять на них (рис. 5.42).

	A	B	C	D	E	F
1	Регион	Среднедушевые денежные доходы, руб.	Среднемесячная номинальная начисленная заработная плата работников организаций, руб.	Средний размер назначенных пенсий, руб.	Численность населения с денежными доходами ниже величины прожиточного минимума, % от общей численности населения	Численность занятых, приходящихся на одного пенсионера, чел.
2	Республика Башкортостан	28125	28108	16806	12,5	1,52
3	Республика Марий Эл	18671	23305	16011	22,5	1,46
4	Республика Мордовия	17695	23229	16154	18,8	1,52
5	Республика Татарстан	32609	30224	16963	7,5	1,75
14	Саратовская область	19406	23548	16254	17,6	1,53
15	Ульяновская область	22481	24334	16372	14,9	1,43

Рис. 5.42

Требуется установить зависимость среднедушевых доходов от отобранных факторов.

Задание 5.2

В файле *Задача 5.4. Число амбулаторно-поликлинических организаций.xlsx* имеются данные о числе амбулаторно-поликлинических организаций в Российской Федерации за 19 лет (рис. 5.43).

	A	B	C
1	№ периода	Год	Число амбулаторно-поликлинических организаций, тыс
2	1	2000	21,3
3	2	2001	21,3
4	3	2002	21,4
19	18	2017	20,2
20	19	2018	20,2

Рис. 5.43

Требуется построить кубическую модель временного ряда.