

Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Пермский национальный исследовательский политехнический
университет»
Электротехнический факультет
Кафедра: Информационные технологии и автоматизированные системы

Дисциплина: «Научно-исследовательский семинар»
Лабораторная работа №4
на тему: «Линейная Регрессия»

Выполнил: студент группы АСУ8-23-1м
Шеретов Марк Алексеевич
Проверил: к.т.н., доцент кафедры ИТАС
Суворов Александр Олегович

Пермь 2024

Постановка задачи

Цель работы: изучить применение линейной регрессии в АП Loginom

Задачи проекта:

Используя материал учебного пособия «Анализ данных в АП Loginom»
(автор А.Б. Яковлев), выполнить задания для самостоятельной работы п. 5.3 (стр. 88).

Выполнение работы

Создадим пакет для выполнения работы в LogiDom. На рисунке 1 представлен «пустой» редактор после создания пакета.

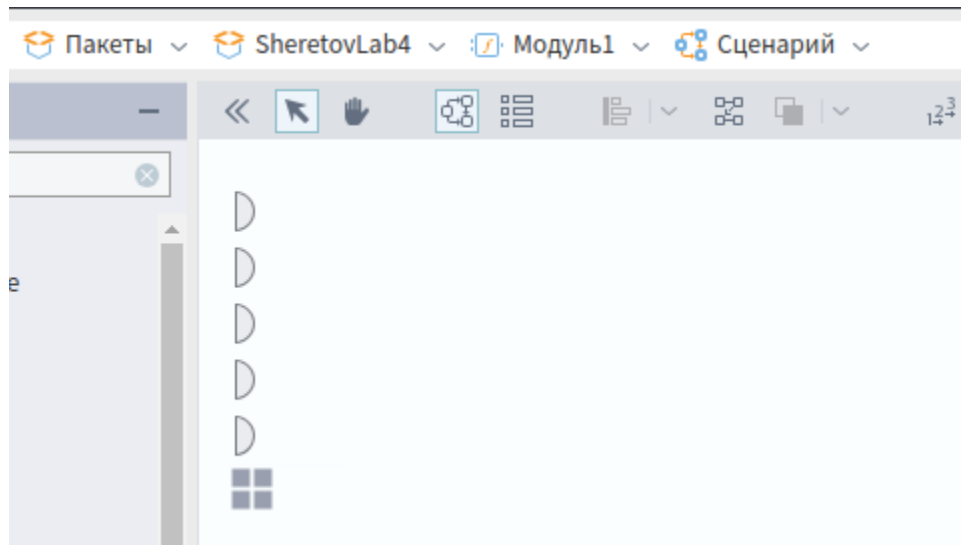


Рисунок 1 — пустой редактор

Задание 1.

Для импорта исходных данных добавим на схему компонент «Excel-файл» и настроим его.

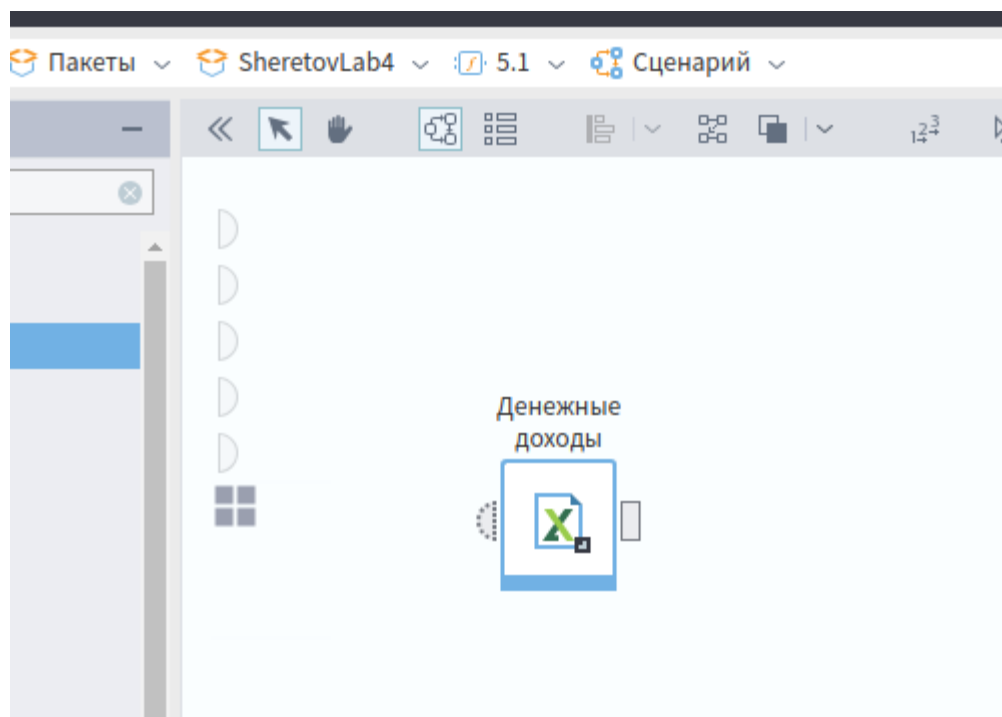


Рисунок 2 — Компонент «Excel-файл» на схеме сценария

Настроим компонент для импорта файла «Денежные доходы» как ранее для остальных excel-файлов. Результат импорта представлен на рисунке 3

#	ab Регион	12 Среднедушевые денежные доходы, руб.	12 Среднемесячная но...	12 Средний размер назна...	9.8 Численность населен...	9.8 Численность занятых, п...
1	Республика Башкортостан	28125	28108	16806	12.5	1.52
2	Республика Марий Эл	18671	23305	16011	22.5	1.46
3	Республика Мордовия	17695	23229	16154	18.8	1.52
4	Республика Татарстан	32609	30224	16963	7.5	1.75
5	Удмуртская Республика	23878	26693	17132	12.3	1.63
6	Чувашская Республика	17872	22908	16254	18.6	1.5
7	Пермский край	28400	30651	17323	14.9	1.52
8	Кировская область	21301	23404	17087	15.9	1.35
9	Нижегородская область	30598	28399	17221	9.6	1.58
10	Оренбургская область	22028	26209	16334	14.8	1.52
11	Пензенская область	21825	25337	16350	14.5	1.44
12	Самарская область	26795	28295	17173	13.8	1.76
13	Саратовская область	19406	23548	16254	17.6	1.53
14	Ульяновская область	22481	24334	16372	14.9	1.43

Рисунок 3 — результат импорта файла «Денежные доходы».

Добавим в сценарий компонент «Линейная регрессия» для проведения регрессионного анализа и подключим к компоненту импорта, как показано на рисунке 4.

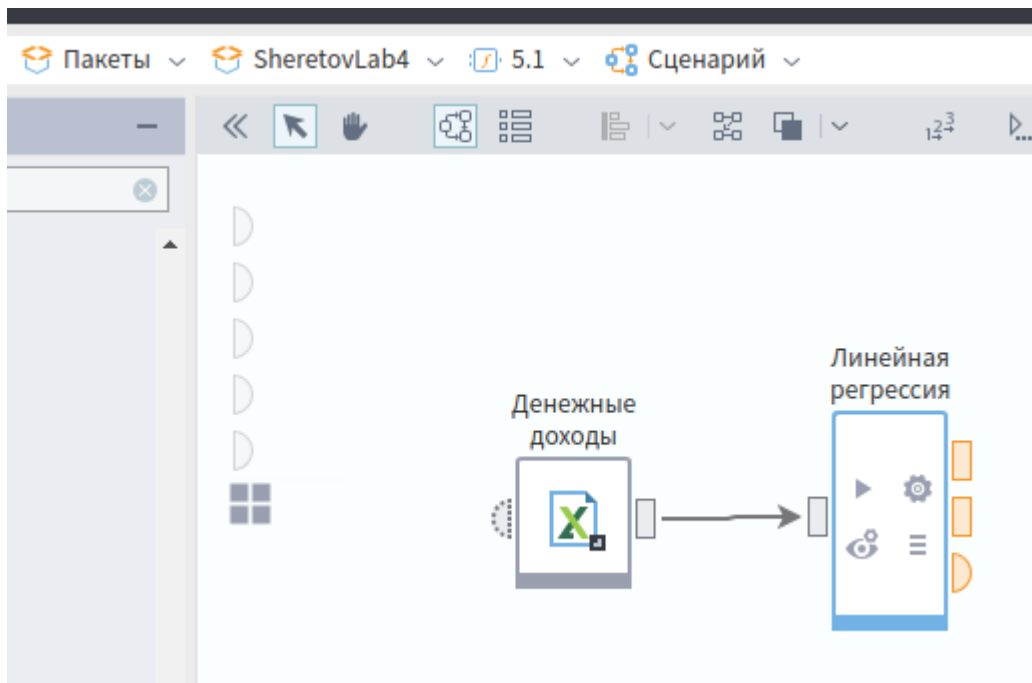


Рисунок 4 — компонент «Линейная регрессия» на схеме сценария

Настроим компонент линейная регрессия:

- произведём настройку входных столбцов как показано на рисунке 5.

Настройка входных столбцов

Метка	И..	Вид данных	Назначение
ab Регион	R...	Дискретный	Не задано
Численность населения с денежными доходами ниже величины прожиточног...	С...	Непрерывн...	Входное
Численность занятых, приходящихся на одного пенсионера, чел.	С...	Непрерывн...	Входное
Среднедушевые денежные доходы, руб.	S...	Непрерывн...	Выходное
Среднемесячная номинальная начисленная заработная плата работников ор...	S...	Непрерывн...	Входное
Средний размер назначенных пенсий, руб.	S...	Непрерывн...	Входное

Рисунок 5 — настройка входных столбцов компонента «Линейная регрессия»

Настройки нормализации оставим без изменения, как показано на рисунке 6.

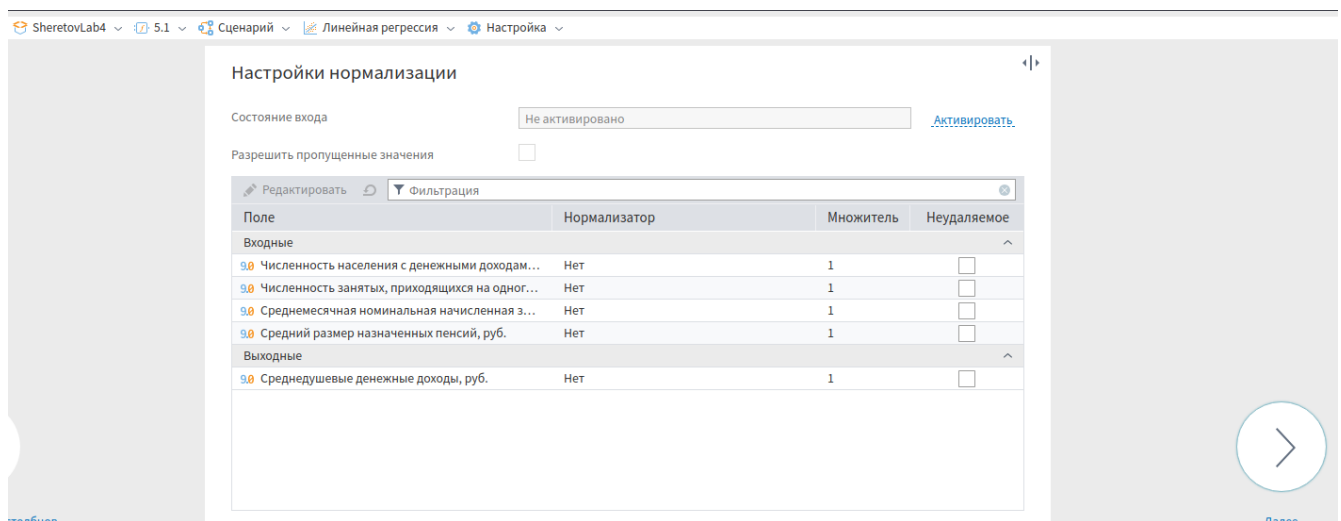


Рисунок 6 — настройки нормализации

- произведём настройку линейной регрессии, как на рисунке 7:
- Отключим «Автоматическую настройку»;
- В качестве метода отбора факторов и защиты от переобучения выберем пошаговое исключение;
- Отметить флажок «Использовать детальные настройки».

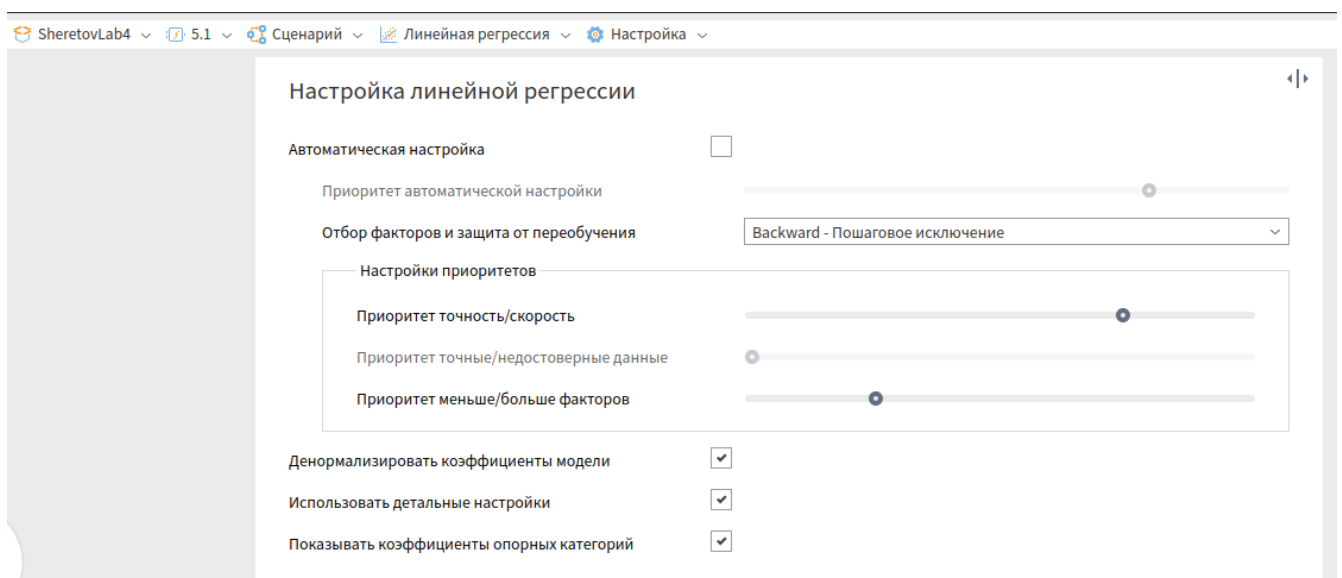
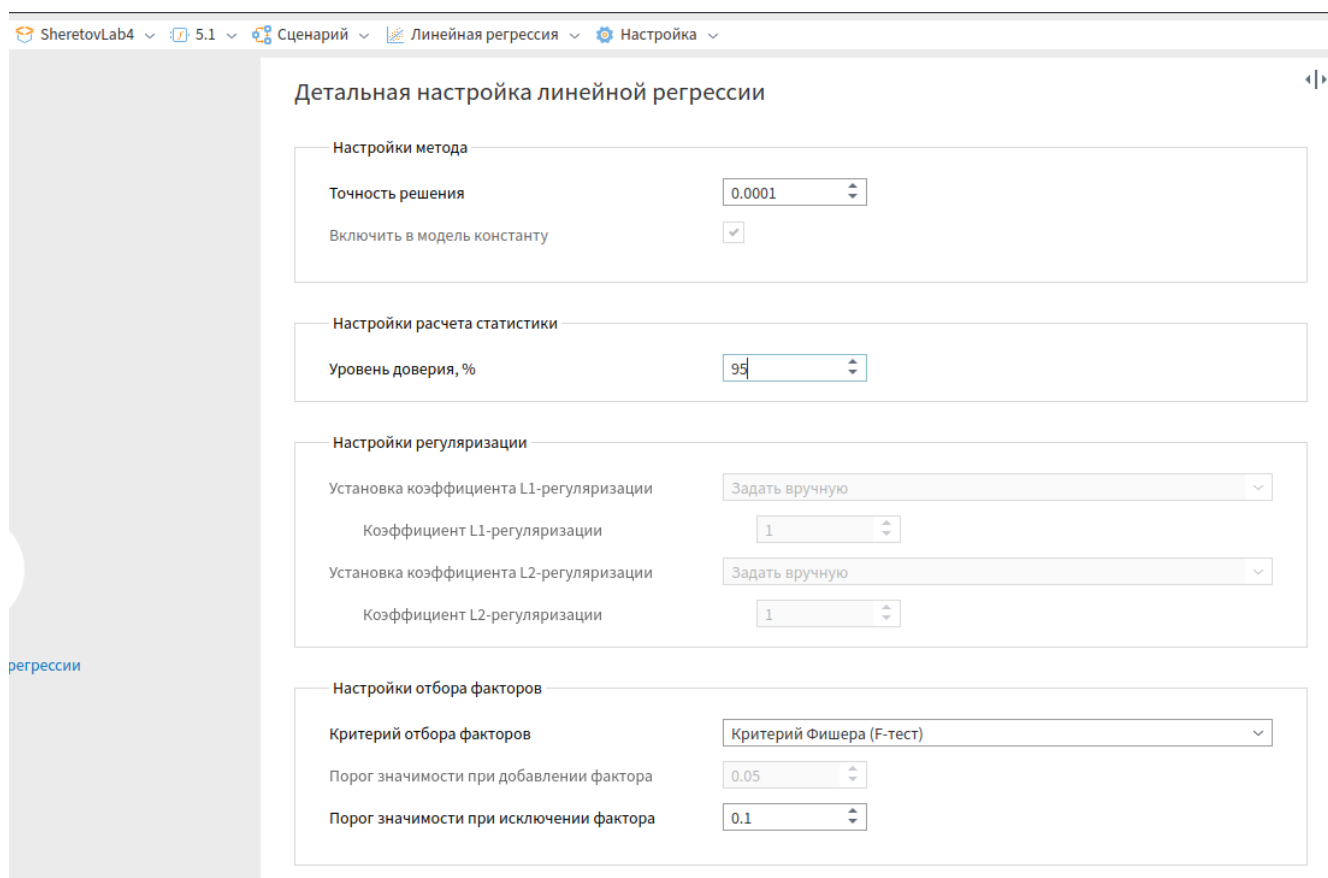


Рисунок 7 — настройка линейной регрессии

На вкладке «Детальная настройка линейной регрессии» произведём следующие настройки (как показано на рисунке 8):

- В качестве критерия отбора факторов выберем «Критерий Фишера (F-тест)»
- Уровень доверия установим равный 95%
- Порог значимости при исключении фактора выберем 0.1



Детальная настройка линейной регрессии

Настройки метода

Точность решения: 0.0001

Включить в модель константу: ☒

Настройки расчета статистики

Уровень доверия, %: 95

Настройки регуляризации

Установка коэффициента L1-регуляризации: Задать вручную

Коэффициент L1-регуляризации: 1

Установка коэффициента L2-регуляризации: Задать вручную

Коэффициент L2-регуляризации: 1

Настройки отбора факторов

Критерий отбора факторов: Критерий Фишера (F-тест)

Порог значимости при добавлении фактора: 0.05

Порог значимости при исключении фактора: 0.1

Рисунок 8 — детальная настройка линейной регрессии

Сохраним настройки узла и в контекстном меню узла на схеме сценария вызовем команду «переобучить узел», как показано на рисунке 9.

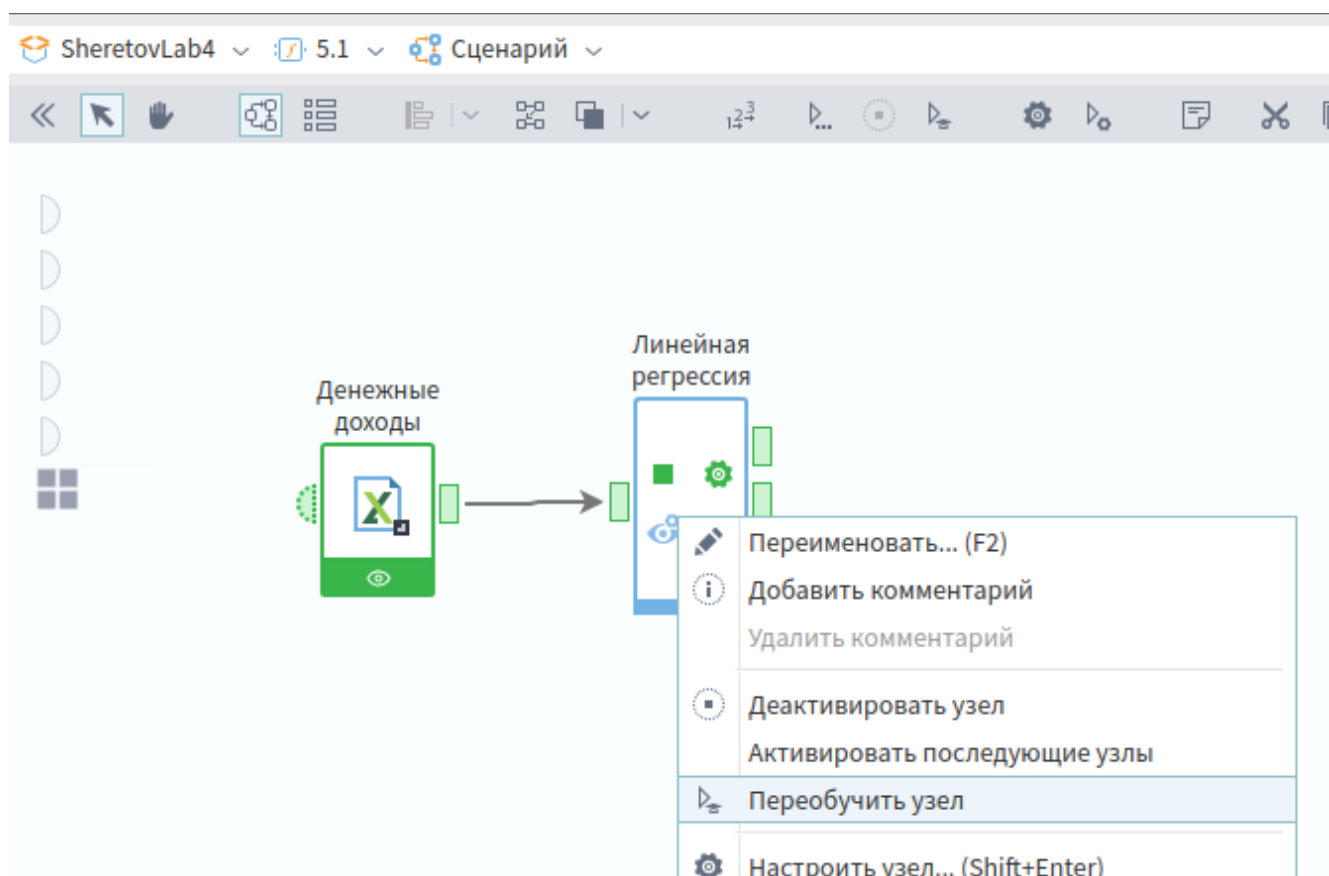
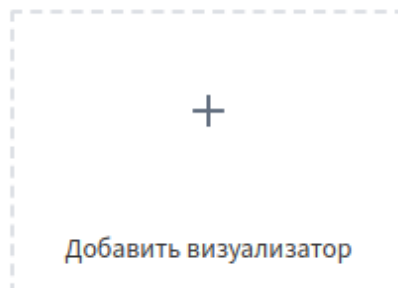


Рисунок 9 — переобучение узла «Линейная регрессия»

Добавим визуализаторы к каждому из выходов компонента «Линейная регрессия» (рисунок 10):

- Таблицу выхода регрессии;
- Таблицу коэффициентов регрессионной модели;
- Отчёт по регрессии.

Выход регрессии



Коэффициенты регрессионной модели



Компонент

Отчет по регрессии

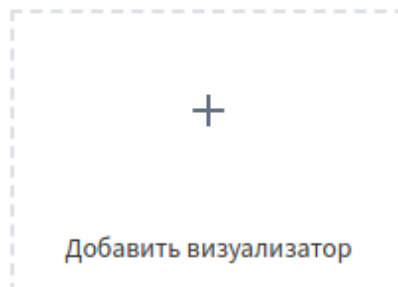


Рисунок 10 — визуализаторы линейной регрессии

Вывод визуализаторов представлен на рисунках 11 — 13 ниже.

#	9.0 Среднедушевые денежные доходы, руб.	Регрессия	ab Регион	9.0 Числен...	9.0 Сре...	9.0 Численн...	9.0 Сре...	9.0 Средни...
1	16842.75234		Республика Марий Эл	22.5	18671	1.46	23305	16011
2	18565.2272		Республика Мордовия	18.8	17695	1.52	23229	16154
3	18297.70796		Чувашская Республика	18.6	17872	1.5	22908	16254
4	19514.94313		Саратовская область	17.6	19406	1.53	23548	16254
5	23912.13192		Оренбургская область	14.8	22028	1.52	26209	16334
6	23066.46033		Пензенская область	14.5	21825	1.44	25337	16350
7	21729.47314		Ульяновская область	14.9	22481	1.43	24334	16372
8	27197.7029		Республика Башкортостан	12.5	28125	1.52	28108	16806
9	32050.27645		Республика Татарстан	7.5	32609	1.75	30224	16963
10	20182.21519		Кировская область	15.9	21301	1.35	23404	17087
11	25685.20122		Удмуртская Республика	12.3	23878	1.63	26693	17132
12	26774.92888		Самарская область	13.8	26795	1.76	28295	17173
13	28946.69999		Нижегородская область	9.6	30598	1.58	28399	17221
14	28918.27935		Пермский край	14.9	28400	1.52	30651	17323

Рисунок 11 — таблица «Выход регрессии»

#	ab Имена входных полей	ab Метки входных полей	ab 9.0 Среднедушевые ...	9.0 Сре...	9.0 Среднедуш...	9.0 Среднеду...	9.0 Среднедушевы...	9.0 Среднедушев...
1	Численност_населения_s_депе...	Численность населения с денежными доходами ниже величины прожиточного минимума, % от ...	-488.9090776	148.67...	-3.268456191	0.00722449653	-816.1391096	-161.6790456
2	Srednemesyachnaya_nominaln...	Среднемесячная номинальная начисленная заработная плата работников организаций, руб.	1.138009532	0.2101...	5.414112487	0.0002120641...	0.6753774573	1.600641606
3	<Константа>		1321.894443	7330.5...	0.1803266841	0.8601753204	-14812.54549	17456.33438

Рисунок 12 — таблица «Коэффициенты регрессионной модели»

Показатель	Значение	Атрибут	Коэффициент	Стандартная ошибка	T-статистика	P-значение	Нижняя граница ДИ	Верхняя граница ДИ
Константа	Включена	9.0 Константа	1.321.894443	7.330.553712		0.180327	0.860175	-14.812.545491
Логарифм функции правдоподобия	-118.575365	9.0 Численность населения с денежными доходами ниж...	-488.909078	148.674347		-3.288456	0.007224	-816.139110
Коэффициент детерминации	0.939012	9.0 Среднемесячная номинальная начисленная зарбот...	1.138010	0.210193		5.414112	0.000212	0.675377
Коэффициент детерминации (скорр.)	0.927923							
Стандартное отклонение	1.301.550330							
Число степеней свободы ошибки	11.00							
Число степеней свободы модели	2.000000							
F-статистика	84.681321							
P-значение модели	2.083788e-7							
Критерий Акаике	17.367909							
Критерий Акаике (скорр.)	17.539338							
Критерий Байеса	17.504850							
Критерий Ханнана-Куинна	17.355233							

Рисунок 13 — вывод «Отчёта о регрессии»

Полученная модель приняла вид:

$$y = 1.138 x_1 - 488.9 x_2 + 1321.89, \text{ где:}$$

x_1 - Среднемесячная номинальная начисленная заработная плата работников организации, руб.

x_2 — Численность населения с денежными доходами ниже величины прожиточного минимума, % от общей численности населения

y — среднедушевые доходы.

Коэффициент детерминации = 93.9% - достаточно высокая точность модели.

Получившуюся модель можно проинтерпретировать так:

- При росте среднемесячной заработной платы на 1000 рублей, средний доход на душу населения вырастет на 1138 рублей.
- При росте численности населения с доходами ниже прожиточного минимума на 1%, средний доход на душу населения уменьшится на 488.9 рублей.

Задание 2

Для выполнения задания 2 был создан отдельный сценарий «5.2», добавим на схему сценария компонент «Excel-файл» для импорта исходных данных, как показано на рисунке 14.

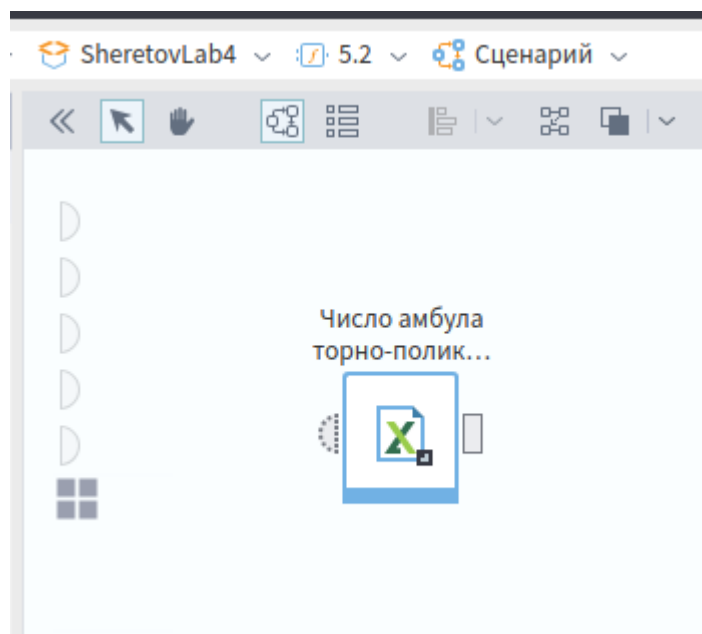
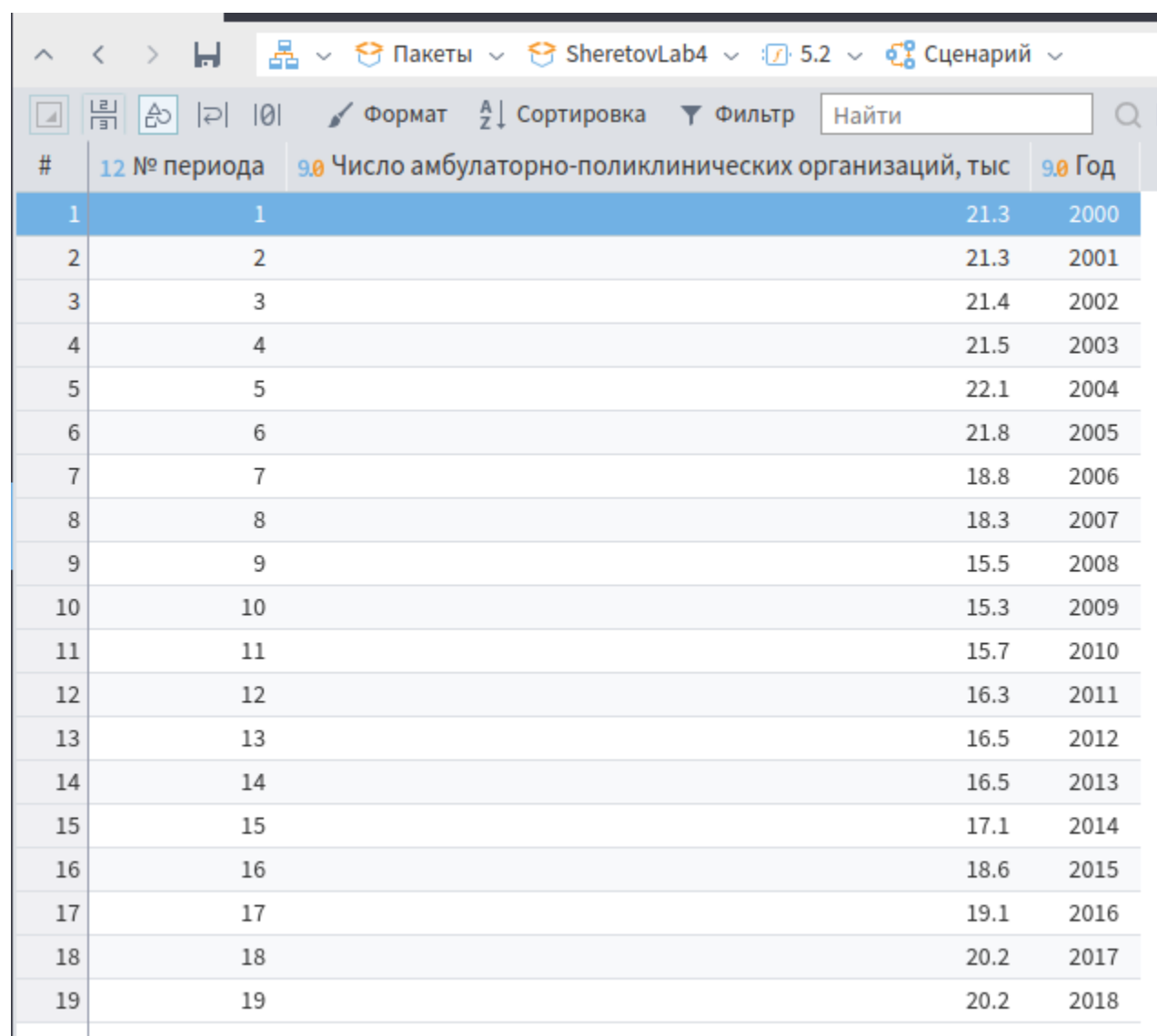


Рисунок 14 — Excel-файл на схеме сценария

Для просмотра исходных данных к компоненту «Excel-файл» был добавлены визуализаторы «Таблица» и «Диаграмма», вывод которых представлен на рисунках 15 и 16.



#	№ периода	Число амбулаторно-поликлинических организаций, тыс	Год
1	1	21.3	2000
2	2	21.3	2001
3	3	21.4	2002
4	4	21.5	2003
5	5	22.1	2004
6	6	21.8	2005
7	7	18.8	2006
8	8	18.3	2007
9	9	15.5	2008
10	10	15.3	2009
11	11	15.7	2010
12	12	16.3	2011
13	13	16.5	2012
14	14	16.5	2013
15	15	17.1	2014
16	16	18.6	2015
17	17	19.1	2016
18	18	20.2	2017
19	19	20.2	2018

Рисунок 15 — таблица исходных данных

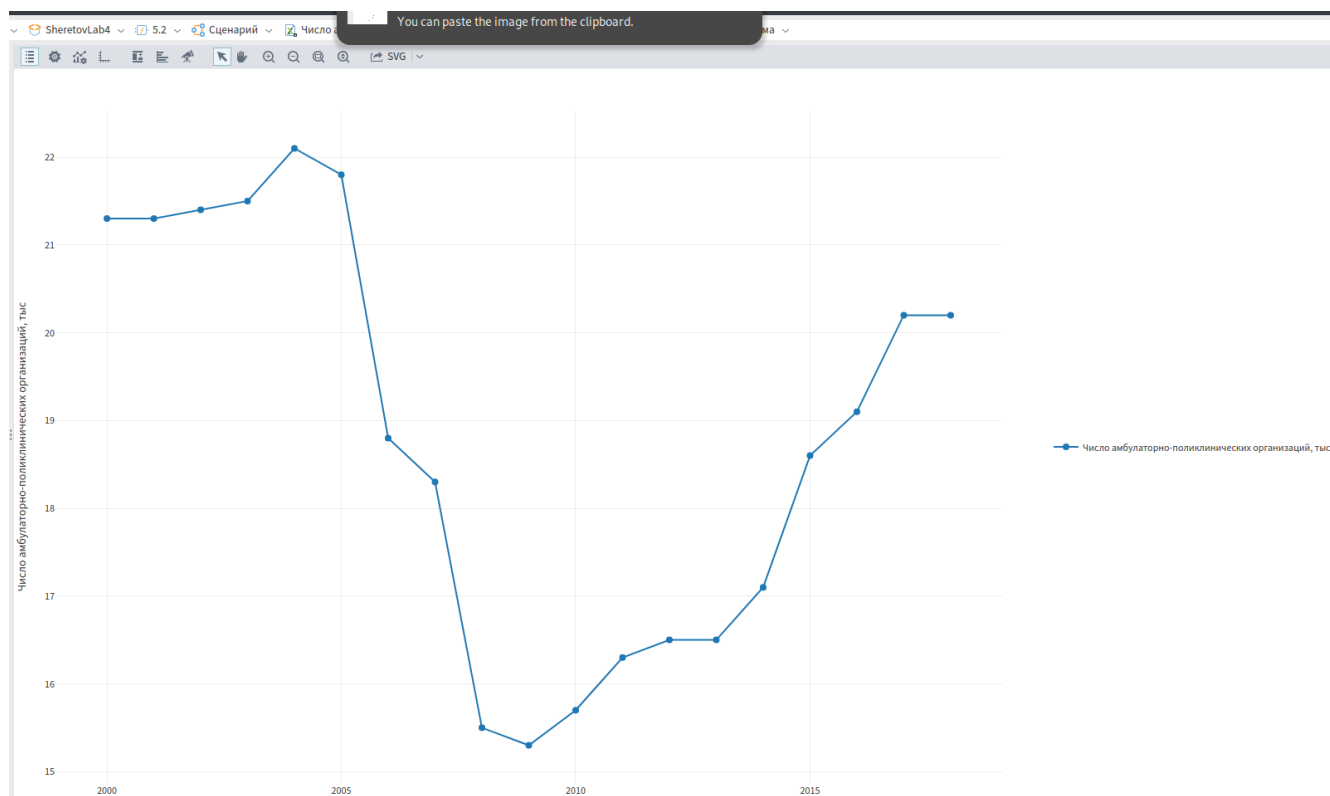


Рисунок 16 — диаграмма исходных данных

Попробуем представить зависимость в виде модели кубической регрессии:

Модель:

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3$$

И приведём модель к линейному виду, используя дополнительные переменные:

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3, \text{ где:}$$

y — число амбулаторных клинических организаций,

x_1 — год,

x_2 — год в квадрате,

x_3 — год в кубе.

Добавим компонент «Калькулятор» и с его помощью рассчитаем x_2 и x_3 .

Настройки калькулятора показаны на рисунке 17 и 18 для x_2 и x_3 соответственно.

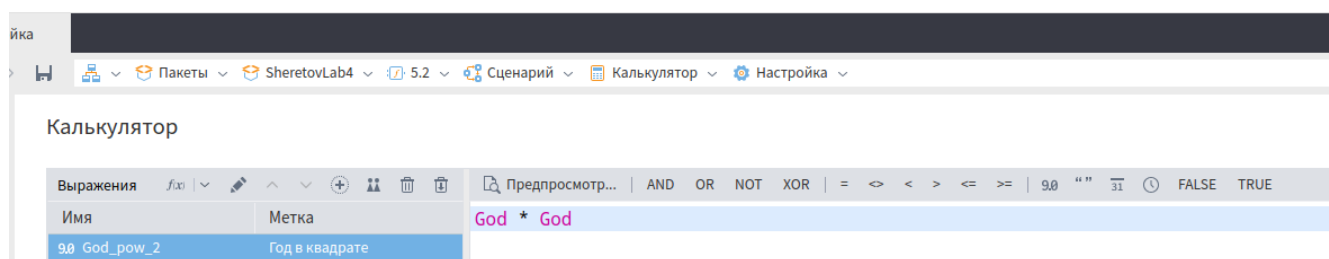


Рисунок 17 — расчёт x_2

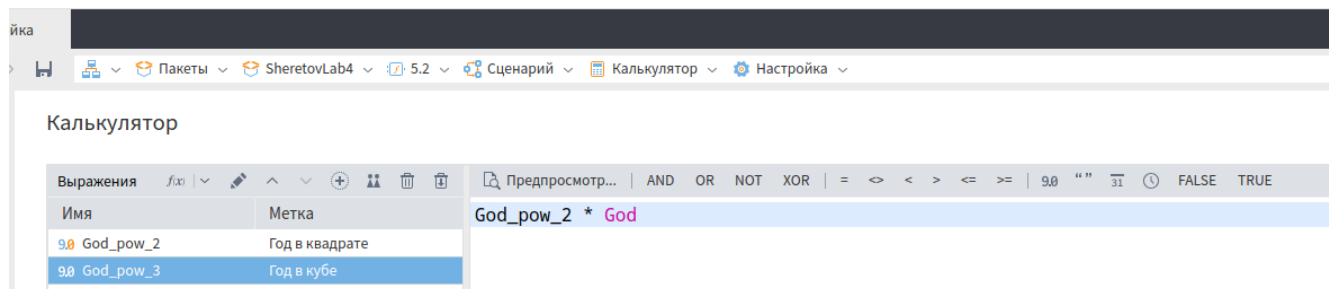


Рисунок 18 — расчёт x_3

Добавим компонент «Линейная регрессия в наш сценарий» и передадим ему на вход результат работы компонента «Калькулятор», как показано на рисунке 19.

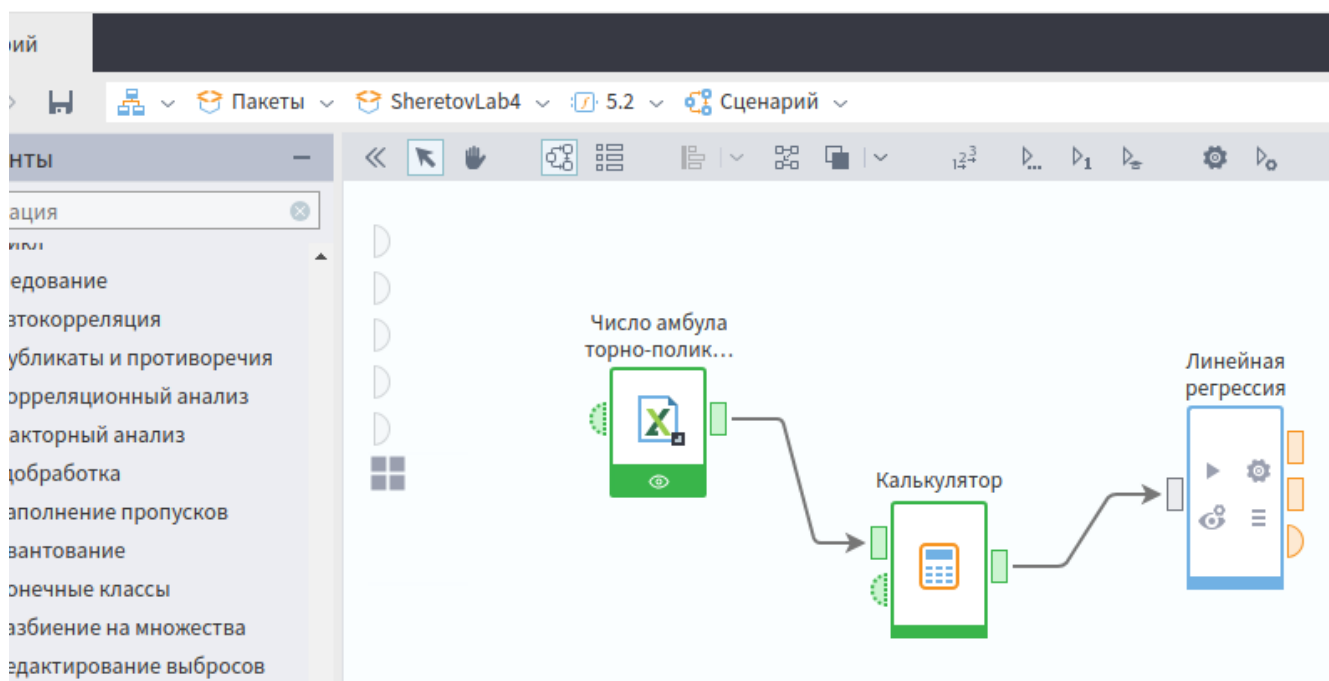


Рисунок 19 — сценарий после добавления компонента «Линейная регрессия»

Настроим входные столбца компонента «Линейная регрессия». В качестве выходного столбца выберем столбец «Число амбулаторных клинических организаций», а различные вариации столбца «Год» выберем входным параметром, как показано на рисунке 20.

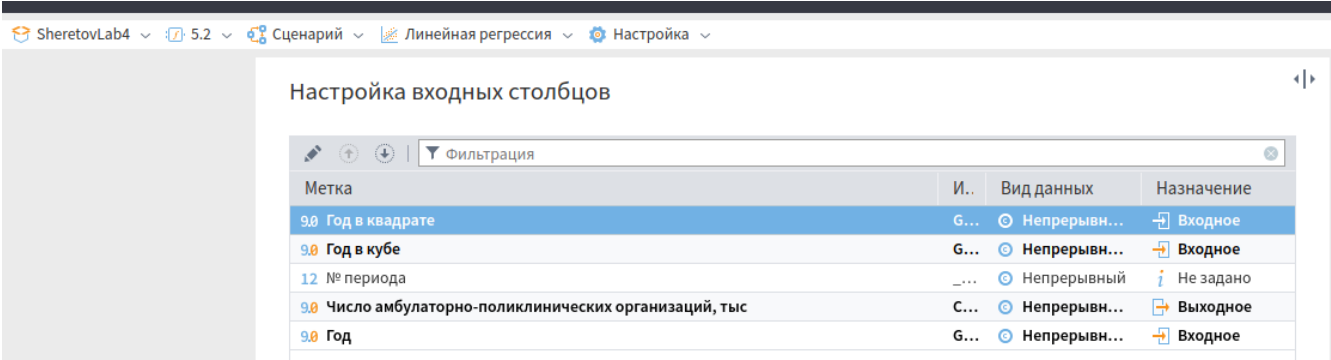


Рисунок 20 — настройка входных столбцов

Все остальные настройки оставим без изменений и переобучим модель.

Добавим визуализаторы для каждого выходного порта компонента «Линейная регрессия» - две таблицы и отчёт о регрессии. Вывод визуализаторов представлен на рисунках 21 - 23.

Число амбулаторно-поликлинических организаций, тыс	Регрессия	Год в квадрате	Год в кубе	№ периода	Год
23.26407588	4000000	8000000000	1	21.3	2000
22.25965564	4004001	8012006001	2	21.3	2001
21.31716637	4008004	8024024008	3	21.4	2002
20.44363267	4012009	8036054027	4	21.5	2003
19.64607913	4016016	8048096064	5	22.1	2004
18.93153033	4020025	8060150125	6	21.8	2005
18.30701086	4024036	8072216216	7	18.8	2006
17.7795453	4028049	8084294343	8	18.3	2007
17.35615826	4032064	8096384512	9	15.5	2008
17.0438743	4036081	8108486729	10	15.3	2009
16.84971802	4040100	8120601000	11	15.7	2010
16.780714	4044121	8132727331	12	16.3	2011
16.84388683	4048144	8144865728	13	16.5	2012
17.04626111	4052169	8157016197	14	16.5	2013
17.39486142	4056196	8169178744	15	17.1	2014
17.89671234	4060225	8181353375	16	18.6	2015
18.55883845	4064256	8193540096	17	19.1	2016
19.38826436	4068289	8205738913	18	20.2	2017
20.39201464	4072324	8217949832	19	20.2	2018

Рисунок 21 — таблица «Выход регрессии»

#	Имена входных полей	Метки вх...	Число амбулаторно-поликлин...	Число ам...	Число амбула...	Число амбулатор...	Число амбу...	Число амбул...
1	<Константа>		-9254214.077	2.986415704	-3098769.56	5.74786752E-90	-9254207.712	-9254220.443
2	God_pow_2	Год в квадр...	-6.997133966	443.0326165	-0.01579372196	0.987607116	937.3045352	-951.2988031
3	God_pow_3	Год в кубе	0.001170764527	847.4402571	1.381530459E-6	0.9999989159	1806.277322	-1806.27498
4	God	Год	13938.32849	0	0	0	13938.32849	13938.32849

Рисунок 22 — таблица «Коэффициенты регрессионной модели»

Показатель	Значение	Атрибут	Коэффициент	Стандартная ошибка	T-статистика	P-значение	Нижняя граница ДИ	Верхняя граница ДИ
Константа	Включена	Константа	-9,254,214.077...	2.986416	-3,098,769.560111	5.747868e-90	-9,254,220.442829	-9,254,207.712040
Логарифм функции правдоподобия	-31.515458	Год в квадрате	-6.997134	443.032617	-0.015794	0.987607	-951.298803	937.304535
Коэффициент детерминации	0.702619	Год в кубе	0.001171	847.440257	0.000001	0.999999	-1,806.274980	1,806.277322
Коэффициент детерминации (скорр.)	0.643143	Год	13,938.328494	0.000000	∞	0.000000	13,938.328494	13,938.328494
Стандартное отклонение	1.430414							
Число степеней свободы ошибки	15.00							
Число степеней свободы модели	3.000000							
F-статистика	11.813451							
P-значение модели	0.000312							
Критерий Акаике	3.738469							
Критерий Акаике (скорр.)	3.888845							
Критерий Байеса	3.937299							
Критерий Ханнана-Куинна	3.772119							

Рисунок 23 — отчёт о регрессии

В результате работы «Линейной регрессии» на автоматических настройках, получилась следующая модель:

$$y = -9254214 + 13938x + 0.001x^2 - 6.997x^3,$$

где y — Число амбулаторных клинических организаций, x — год.

Коэффициент детерминации для этой модели: 70.26%, что можно считать довольно хорошим результатом, показывающим довольно хорошую точность модели. Большое значение параметра «F-статистики», в купе с «P-значением модели» превосходящим 0, можно интерпретировать, как успех регрессионной модели.

Построим диаграмму по результатам регрессии для сравнения с исходными данными (рисунок 24).

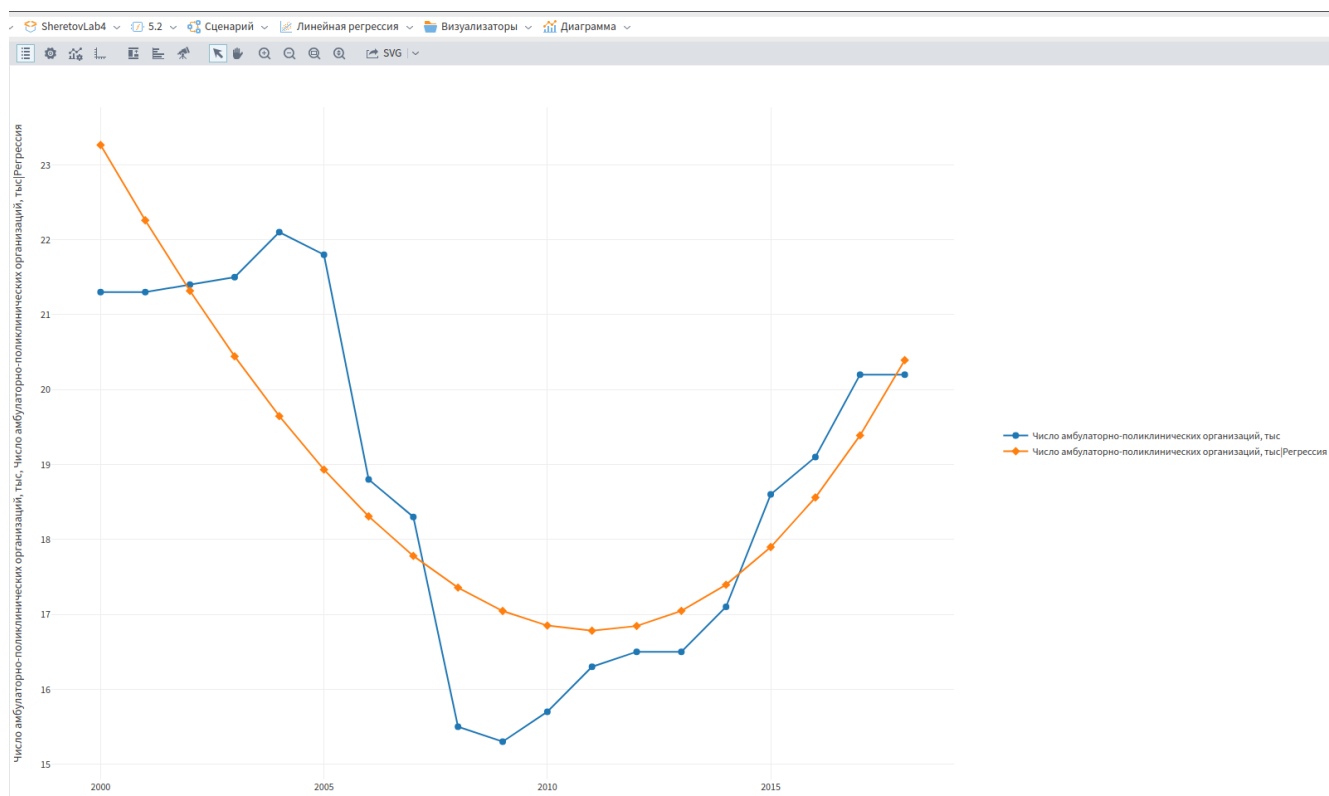


Рисунок 24 — графики исходных данных и регрессионной модели