Министерство науки и высшего образования Российской Федерации

Федеральное государственное автономное образовательное

учреждения высшего образования

«Национальный исследовательский технологический университет «МИСиС»

Институт компьютерных наук (ИКН)

Кафедра автоматизированных систем управления (АСУ)

КУРСОВАЯ РАБОТА

по дисциплине:

«Прикладной статистический анализ»

на тему:

«Разработка модели прогнозирования развития опасных болезней»

Направление подготовки:

09.03.01 «Информатика и вычислительная техника»

Семестр 5

|  |  |
| --- | --- |
| Выполнил: | Проверил: |
| Харлашкина А.В.\_\_  (Ф.И.О. студента) | Маркарян А.О.  (Ф.И.О преподавателя) |
| БИВТ-21-3\_\_\_\_\_  (№ группы) | \_\_\_\_\_\_\_\_\_\_\_\_\_\_  (оценка) |
| 22.12.2023\_\_\_\_\_  (дата сдачи) | \_\_\_\_\_\_\_\_\_\_\_\_\_\_  (дата проверки) |
| Подпись: \_\_\_\_\_\_\_\_\_\_\_\_ | Подпись: \_\_\_\_\_\_\_\_\_\_\_\_\_ |

|  |  |  |
| --- | --- | --- |
|  |  |  |

Москва – 2023

СОДЕРЖАНИЕ

[ВВЕДЕНИЕ 3](#_Toc154109168)

[1 Анализ характеристик объекта исследования 4](#_Toc154109169)

[1.1 Описание объекта исследования 4](#_Toc154109170)

[1.2 Анализ объекта исследования с помощью статистических показателей 6](#_Toc154109171)

[1.3 Выявление причинно-следственных связей 13](#_Toc154109172)

[1.4 Постановка задачи моделирования 21](#_Toc154109173)

[2 Моделирование статистических зависимостей 21](#_Toc154109174)

[2.1 Формализация и классификация переменных 21](#_Toc154109175)

[2.2 Проверка гипотезы о нормальном распределении 22](#_Toc154109176)

[2.3 Корреляционный анализ 26](#_Toc154109177)

[2.4 Построение регрессионной модели 29](#_Toc154109178)

[2.4.1 Структурная идентификация модели 29](#_Toc154109179)

[2.4.2 Параметрическая идентификация модели 37](#_Toc154109180)

[3 Исследование модели 39](#_Toc154109181)

[3.1 Анализ статистической значимости уравнения регрессии 39](#_Toc154109182)

[3.2 Анализ статистической значимости коэффициентов уравнения регрессии 40](#_Toc154109183)

[3.3 Исследование мультиколлинеарности факторов 41](#_Toc154109184)

[3.4 Применения шагового регрессионного анализа для улучшения модели 43](#_Toc154109185)

[4 Программная реализация и численное исследование результатов моделирования 46](#_Toc154109186)

[4.1 Обоснование выбора и описание программного обеспечения 46](#_Toc154109187)

[4.2 Описание основных модулей программы 46](#_Toc154109188)

[4.3 Численное исследование результатов моделирования 48](#_Toc154109189)

[ЗАКЛЮЧЕНИЕ 50](#_Toc154109190)

[Список используемой литературы 51](#_Toc154109191)

[ПРИЛОЖЕНИЕ 1 52](#_Toc154109192)

# ВВЕДЕНИЕ

Распространение ВИЧ представляет собой одну из наиболее сложных и актуальных проблем здравоохранения в мире, в особенности в развивающихся странах. Россия не является исключением, и вопросы, связанные с ВИЧ, остаются одними из наиболее значимых для общественного здоровья.

Ситуация с распространением ВИЧ в России рядом исследователей характеризуется как эпидемия. В последние годы Россия признана страной с самым высоким уровнем ВИЧ-инфицированных в Европе. По оценкам, пик заражаемости ВИЧ в РФ пришелся на 2017 год, а в течение последующих пяти лет Роспотребнадзор ежегодно отчитывался о 70-100 тысяч новых заражённых.

К 2022 году доля ВИЧ-инфицированных России составляла 1,4% взрослого населения, и в стране проживало более миллиона человек, инфицированных ВИЧ. Поэтому разработка модели прогнозирования динамики инфицирования ВИЧ в России приобретает особую актуальность и представляет собой проблему, требующую специального внимания и комплексного анализа.

**Целью** данного исследования является разработка модели прогнозирования динамики инфицирования ВИЧ в России.

Для достижения этой цели необходимо решить следующие **задачи**:

* Провести анализ динамики инфицирования ВИЧ-инфекции в течение предшествующих лет;
* Проверить гипотезы о нормальном распределении распространения заболеваемости в периоды времени с 2012 года по 2020 год;
* Определить социально-экономические и демографические факторы, оказывающие влияние на уровень распространения ВИЧ;
* Оценить влияние выявленных факторов на динамику распространения заболевания;
* Построить регрессионную модель, отражающую влияние выявленных факторов на развитие эпидемии ВИЧ в России.

**Объектом** исследования является количество новых случаев заражения ВИЧ в России за текущий год, а **предметом** – модель прогнозирования развития новых случаев заражения ВИЧ в России.

Результаты данного исследования послужат основой для разработки эффективных стратегий предотвращения и управления распространением ВИЧ-инфекции в будущем в России.

1. Анализ характеристик объекта исследования
2. Описание объекта исследования

ВИЧ (вирус иммунодефицита человека) – инфекционное хроническое заболевание, передающееся контактным путем, медленно прогрессирующее и характеризующееся поражением иммунной системы с развитием синдрома приобретенного иммунодефицита (СПИДа).

Согласно классификации Всемирной организации здравоохранения (ВОЗ), существуют 3 стадии распространения эпидемии ВИЧ:

1. На первом этапе заражено менее 1 % населения в целом и менее 5 % — в уязвимых группах.
2. Вторая стадия – концентрирования: когда заражено более 5 % в одной уязвимой группе.
3. Третья – генерализованная стадия: ВИЧ-инфекция выходит за рамки одной группы, эпидемия охватывает широкие слои населения.

На данный момент в России ВИЧ приписывают характер эпидемии. На начало 2022 года инфицированы были 1,137 млн. россиян из 144 млн. общей численности населения.

Распространению заболеваемости ВИЧ в России способствуют следующие факторы:

* + Недостаточное информирование и образование населения о методах предотвращения ВИЧ-инфекции.
  + Низкий уровень доступности качественных услуг по предупреждению, диагностике и лечению ВИЧ.
  + Социальные проблемы, такие как низкий уровень осведомленности, выселение и стигматизация лиц, живущих с ВИЧ, что может затруднять доступ к необходимой помощи и поддержке.
  + Несоблюдение мер профилактики среди групп с высоким риском.
  + Неблагоприятные экономические, политические и социокультурные условия, которые могут затруднять создание продуктивных программ по предотвращению и лечению ВИЧ.
  + ВИЧ-инфекция может протекать бессимптомно в течение 8-10 лет, что делает ее распространение незаметной.

Как показывает статистика, за последние десятилетия заболеваемость ВИЧ в России увеличилась в три раза и достигла 60 случаев на 100 тысяч жителей. Однако, точное число носителей в стране определить сложно, так как медицинские службы не успевают за распространением вируса, а уязвимые группы населения зачастую неохотно проходят обследования.

Положение ВИЧ-инфицированных также усугубилось после вторжения России на Украину и введения ответных западных санкций. В связи с этим были сокращены закупки дорогих зарубежных препаратов для лечения ВИЧ, и в некоторых регионах страны на данный момент наблюдается нехватка лекарств.

Таким образом, по данным Росстата общее количество новых случаев заражения ВИЧ-инфекцией в России за период с 2012 по 2022 год составило 932361 человек. Подробное количество случаев заражения за каждый год представлено в таблице №1.

Таблица №1 – Динамика новых случаев заражения ВИЧ-инфекцией в России

|  |  |
| --- | --- |
| **Год** | **Зарегистрировано новых случаев заражения ВИЧ, чел.** |
| 2012 | 69280 |
| 2013 | 77896 |
| 2014 | 85252 |
| 2015 | 88040 |
| 2016 | 89155 |
| 2017 | 104402 |
| 2018 | 101345 |
| 2019 | 94668 |
| 2020 | 88154 |
| 2021 | 71019 |
| 2022 | 63150 |

На диаграмме (Рисунок 1) распространения ВИЧ в России за 11 лет (2012-2022) заметно, что пик заражаемости пришелся на 2017 год – 104402 новых случаев заражения.



Рисунок 1 – Динамика новых случаев заражения ВИЧ-инфекцией в России

После 2017 наблюдается спад заражаемости ВИЧ среди жителей РФ. В некоторой степени это связано с более низкими объёмами тестирования во время и сразу после пандемии COVID-19. А также с ужесточением некоторых законов, направленных на уязвимые группы населения.

Несмотря на то, что часть ВИЧ-инфицированных в стране до сих пор не выявлена, с каждым годом все большее количество людей проходят тестирования образцов крови на ВИЧ, и осведомленность населения о самом заболевании и способах защиты от заражения значительно увеличивается.

1. Анализ объекта исследования с помощью статистических показателей

Осуществим анализ ключевых статистических показателей.

Для начала проведем подсчет и анализ статистических характеристик (средняя, дисперсия, среднее квадратическое отклонение, мода, медиана, размах, коэффициент вариации) для данных о новых случаях заражения ВИЧ за период с 2012 по 2022 год.

Статистические характеристики для данных из таблицы №1 вычисляются по следующим формулам:

1. Средняя – среднее арифметическое наблюдаемых значений случайной величины (N – сумма частот):
2. Дисперсия – мера рассеивания значений относительно средней:
3. Среднее квадратическое отклонение:
4. Модой Мо называется варианта с наибольшей частотой, то есть значение случайной величины, наблюдаемое чаще всего.
5. Медиана Me – варианта, которая делит вариационный ряд на две равные части. Если число вариант нечетное (k = 2m + 1), то Me = xm+1. Если число вариант четное (k = 2m), то
6. Размах – разность между наибольшей и наименьшей вариантами: R = xmax − xmin
7. Коэффициент вариации – выраженное в процентах отношение среднего квадратического отклонения к средней:

Таким образом, получим:

* : характеризует среднее значение количества новых случаев заболевания ВИЧ за год с в период с 2012 по 2022 год.
* 158493332.26: показывает меру разброса данных относительно среднего значения.
* 12589.41: выражает, насколько в среднем каждая точка данных отклоняется от среднего значения.
* 14.85%: отражает уровень изменчивости данных относительно их среднего значения.
* 88040.0: значение, разделяющее набор данных на две равные половины.
* Максимальное значение: 104402 случаев. Показывает наибольшее число заражений в год за 11 лет.
* Минимальное значение: 63150 случаев. Показывает наименьшее число заражений в год за 11 лет.
* Размах: 41252 случаев. Показывает разницу между наибольшим и наименьшим значениями заражений в год. Так как наибольшее число заражений за период с 2012 по 2022 приходится на 2017 год, а наименьшее на 2022, и при этом с 2017 года наблюдается спад, то размах показывает на сколько сократилось ежегодное количество заражений ВИЧ к 2022 году.
* Мода: 2017, значение: 104402. Показывает, что на 2017 год пришелся пик выявленных заражений ВИЧ-инфекцией.

Теперь проведем анализ показателей интенсивности. Они характеризуют степень распространенности событий или явлений в определенной среде (на душу населения, на определенной территории и др.). Для данного исследования найдем уровень заболеваемости ВИЧ.

Уровень заболеваемости ВИЧ – количественный показатель, который показывает количество зарегистрированных новых случаев заражения ВИЧ-инфекцией на 100 тыс. населения ежегодно. Этот показатель позволяет осуществить сравнение уровня распространения заболевания за разные периоды времени (в данном исследовании с 2012 по 2022 годы).

Уровень заболеваемости ВИЧ (Ki) рассчитывается по формуле:

Где Н – численность населения, С – количество новых случаев заражения ВИЧ.

Таблица №2 – Динамика изменения уровня заболеваемости ВИЧ в России

|  |  |
| --- | --- |
| **Год** | **Уровень заболеваемости ВИЧ на 100 тыс. населения** |
| 2012 | 48.43 |
| 2013 | 54.34 |
| 2014 | 59.34 |
| 2015 | 60.19 |
| 2016 | 60.84 |
| 2017 | 71.12 |
| 2018 | 68.1 |
| 2019 | 64.5 |
| 2020 | 60.07 |
| 2021 | 48.25 |
| 2022 | 42.97 |

По данным таблицы №2 уровень заболеваемости ВИЧ на 2022 год составлял 42,97 случаев на 100 тыс. человек населения. Это самый маленький показатель за предыдущие 11 лет (с 2012 по 2022).

Пик приходится на 2017 год – 71,12 случаев заражения ВИЧ-инфекцией на 100 тыс. населения. Это рекордный показатель заболеваемости в стране. Столь высокий показатель, к сожалению, не говорит о реальном всплеске заболеваемости, скорее речь идет о расширении охвата населения диагностическим тестированием.

На основе данных по возрасту людей, заболевших ВИЧ (Таблица №3), за 2022 год, взятых из статистических сборников Росстата, проведем анализ и определим группу (возраст) наиболее подверженных заражению ВИЧ людей в настоящее время.

Для этого вычислим показатели структуры (удельный вес) и показатели сравнения.

Удельный вес a вычисляется по формуле:

Где Σ – общее число новых случаев заболевания ВИЧ за 2022 год, а x – число зарегистрированных заболеваний ВИЧ для определенной возрастной группы.

Показатель сравнения x вычисляется по формуле:

Где a – эталонный показатель. В качестве эталонного показателя выберем минимальное число зарегистрированных случаев. По данным из таблицы №3, таким показателем является число заболевших в возрастной группе от 15 до 20 лет. b – число зарегистрированных заболеваний ВИЧ для определенной возрастной группы.

Таблица №3 – Распределение заболеваемости ВИЧ по возрасту за 2022 год

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Возраст заболевших, лет** | **Зарегистрировано заболевших, чел.** | **Распределение инфицированных ВИЧ в РФ по возрасту, %** | **Сравнение, %** | **Разница, %** |
| Дети до 15 лет | 615 | 0.97 | 121.78 | 21.78 |
| 15-20 | 505 | 0.8 | 100 | 0 |
| 20-29 | 7073 | 11.2 | 1400.59 | 1300.59 |
| 30-39 | 23934 | 37.9 | 4739.41 | 4639.41 |
| 40-49 | 20334 | 32.2 | 4026.53 | 3926.53 |
| 50-59 | 7451 | 11.8 | 1475.45 | 1375.45 |
| 60 и выше | 3238 | 5.13 | 641.19 | 541.19 |

Таким образом, получили, что в 2022 году почти 85% из всех новых случаев заражения ВИЧ были выявлены у экономически активных людей в возрасте 30-59 лет.

Наглядное и подробное распределение возраста людей, инфицированных ВИЧ в 2022 представлено на диаграмме (Рисунок 2).

Рисунок 2 – Структура заболеваемости ВИЧ за 2022 год

Наиболее наглядно динамику заражаемости ВИЧ за период с 2012 по 2022 год в России можно увидеть, высчитав ряд важных показателей динамики, к которым относятся: абсолютный прирост, темп роста, темп прироста.

Темп роста вычисляется по следующим формулам:

|  |  |
| --- | --- |
| (базисный) |  |
| (цепной) |  |

Формула для вычисления темпа прироста, вытекает из формулы темпа прироста:

Абсолютный прирост (базисный):

Абсолютный прирост (цепной):

Рассчитанные значения показателей динамики представлены в таблице №4. По данным цепного абсолютного прироста можно увидеть, что изначально численность инфицированных ВИЧ в России в год росла, а потом стала падать. Исходя из базисного абсолютного прироста, только в 2022 году наблюдается спад заражаемости по отношению к 2012 году – началу исследуемого периода (на 6130 случаев меньше).

Таблица №4 – Динамический анализ численности новых случаев заражения ВИЧ-инфекцией за 2012–2022 гг.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Год** | **Зарегистрировано новых случаев заражения ВИЧ, чел.** | **Абсолютный прирост** | | **Темп роста, %** | | **Темп прироста, %** | |
|  |
| **базисный** | **цепной** | **базисный** | **цепной** | **базисный** | **цепной** |  |
| 2012 | 69280 | – | – | – | – | – | – |  |
| 2013 | 77896 | 8616 | 8616 | 112.44% | 112.44% | 12.44% | 12.44% |  |
| 2014 | 85252 | 15972 | 7356 | 123.05% | 109.44% | 23.05% | 9.44% |  |
| 2015 | 88040 | 18760 | 2788 | 127.08% | 103.27% | 27.08% | 3.27% |  |
| 2016 | 89155 | 19875 | 1115 | 128.69% | 101.27% | 28.69% | 1.27% |  |
| 2017 | 104402 | 35122 | 15247 | 150.69% | 117.10% | 50.69% | 17.10% |  |
| 2018 | 101345 | 32065 | -3057 | 146.28% | 97.07% | 46.28% | -2.92% |  |
| 2019 | 94668 | 25388 | -6677 | 136.65% | 93.41% | 36.65% | -6.59% |  |
| 2020 | 88154 | 18874 | -6514 | 127.24% | 93.12% | 27.24% | -6.88% |  |
| 2021 | 71019 | 1739 | -17135 | 102.51% | 80.56% | 2.51% | -19.44% |  |
| 2022 | 63150 | -6130 | -7869 | 91.15% | 88.92% | -8.85% | -11.08% |  |

Для полноты анализа следует рассчитать средние показатели ряда динамики (таблица №1), которые являются обобщающей характеристикой его развития. К ним относятся: средний уровень ряда динамики, средний абсолютный прирост, средний темп роста, средний темп прироста.

Средний уровень ряда – среднегодовое количество зарегистрированных новых случаев заболевания ВИЧ в России.

Средний уровень ряда определяется по формуле:

Средний абсолютный прирост:

Средний темп роста:

Средний темп прироста:

Таким образом, после применения формул получились следующие результаты:

* Средний уровень ряда: 84760.09 (человек)
* Средний абсолютный прирост: -613.0 (человек)
* Средний темп роста: 99.08 %
* Средний темп прироста: -0.92 %

Получается, что за период с 2012 года по 2022 год в России число новых случаев заражения ВИЧ-инфекцией ежегодно варьировалось в районе 84760 человек. Исходя из результатов, можно сделать вывод, что за 11 лет ежегодно численность заражения ВИЧ в России падала на 613 человек (примерно на 1% в год). Однако, исходя из графика распространения ВИЧ-инфекции в России за 2012-2022 (Рисунок 1) видно, что изначально численность новых случаев заражения росла, а уже потом (после 2017 года) стала падать, тем самым уменьшив общее число новых заражений ВИЧ в год до уровня меньшего 2012 году.

С помощью полученных данных можно осуществить экстраполяцию динамического ряда (Таблица №1). Стоит помнить, что в основе методов экстраполяции лежит предположение, что факторы, которые обусловили развитие новых случаев заражения ВИЧ, остаются неизменными и в течение следующего периода.

Для получения прогноза новых случаев заражения ВИЧ на 2023 год в России воспользуемся тремя разными способами:

1. Прогнозирование по среднему абсолютному приросту

Формула для вычисления прогнозируемого числа заражений ВИЧ на 2023 год по среднему абсолютному приросту:

В результате применения данного метода прогноз новых случаев заболеваний ВИЧ на 2023 в РФ год составил 62537 человек.

1. Прогнозирование по среднему темпу роста

Формула для вычисления прогнозируемого числа заражений ВИЧ на 2023 год по среднему темпу роста:

Получилось, что прогнозирование по среднему темпу роста на 2023 год составило 62567.66 новых случаев заражения ВИЧ-инфекцией.

Это значение примерно такое же, как и значение прогнозирования по среднему абсолютному приросту. Следовательно, прогнозирование двумя этими способами показало, что заболеваемость ВИЧ в России на 2023 год составит приблизительно 62500 человек.

Тем самым, прогнозируется маленький спад по сравнению с 2022 годом и несущественный, если смотреть на ежегодную динамику всего периода в целом.

1. Прогнозирование методом аналитического выравнивания

Если посмотреть на график распространения ВИЧ-инфекции в России за 2012-2022 года (Рисунок 1), то можно обнаружить параболическую зависимость между данными. Такая зависимость используется для описания динамических рядов, в которых меняется направление развития – со снижения показателей на их рост и наоборот. В данном случае сначала показатели увеличивались (с 2012 по 2017 года), а потом пошли на спад (2018-2022 года).

Параболическая зависимость:

Параметр c характеризует изменение интенсивности развития заболеваемости ВИЧ в единицу времени: при c > 0 – ускоренное развитие, при c < 0 – замедленное.

Поэтому, с помощью метода наименьших квадратов, необходимо найти параметры параболической функции, аппроксимирующей исходный динамический ряд.

В результате функция, наиболее точно, описывающая динамику заболеваемости ВИЧ в России, получилась следующей:

Если подставим вместо t значение года – 2023 год, то получим прогнозируемое число новых случаев заражения ВИЧ на 2023 год.

Таким образом, прогнозирование методом аналитического выравнивания на 2023 год составило 48655.7 новых случаев заражения ВИЧ в России.

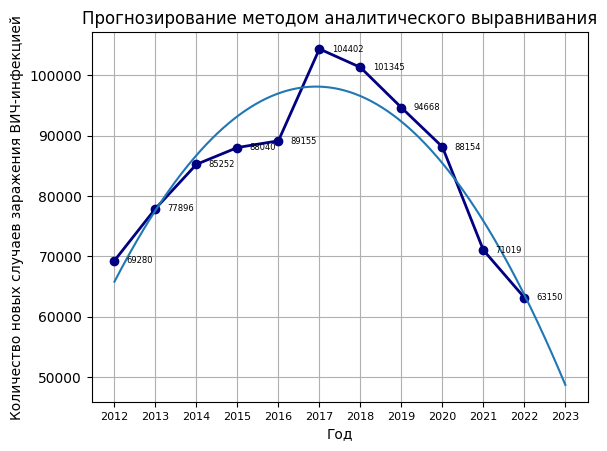


Рисунок 3 – Аппроксимация исходной кривой параболой

Как видно из графика с аппроксимирующей кривой (Рисунок 3) дальнейшее развитие ВИЧ инфекции в России должно идти на спад. Однако, вероятнее всего данное прогнозирование является наименее точным в сравнении со всеми остальными. Такой относительно низкий показатель заболеваемости, к сожалению, трудно представить в настоящее время.

Тем не менее на основе прогнозирования можно предположить, в 2023 году будет наблюдаться незначительный спад числа новых заражений ВИЧ-инфекцией.

1. Выявление причинно-следственных связей

Динамика выявления новых случаев заболеваемости ВИЧ в России за период с 2012 по 2022 год сначала неутешительно росла до 2017 года, а потом стала падать, и к 2022 году наблюдается тенденция к снижению уровня заболеваемости ВИЧ-инфекцией по всей стране.

Следует выделить несколько важных факторов, от которых косвенно или напрямую зависит распространения ВИЧ-инфекции в России. Выявление данных факторов поможет детальнее разобраться в проблеме распространения вируса по стране и составить более точную модель прогнозирования, которая будет учитывать самые разные аспекты.

В первую очередь, стоит отметить, что люди с низкими доходами часто оказываются в более уязвимом положении, когда речь идет о доступе к медицинским услугам и профилактике заболеваний. Они могут иметь ограниченный доступ к средствам для защиты от ВИЧ и не иметь средств на получение качественной медицинской помощи и антиретровирусной терапии, что увеличивает риск заражения ВИЧ и распространения вируса. Также люди с низкими доходами могут быть вынуждены заниматься видами работы, связанными с высоким риском заражения ВИЧ (например, проституцией), чтобы обеспечить себе и своим семьям пропитание.

Таким образом, численность населения с денежными доходами ниже границы бедности оказывает значительное влияние на распространение ВИЧ в России, увеличивая уязвимость групп населения и ухудшая доступ к профилактическим и лечебным мерам. Следовательно, можно обозначить связь: чем выше уровень бедности в стране, тем выше показатели заболеваемости ВИЧ в России.

Как видно из графика динамики численности населения с денежными доходами ниже границы бедности за период с 2012 по 2022 год (Рисунок 4), с 2012 года наблюдался рост, а после 2015 года спад, который продолжается и на текущий момент.

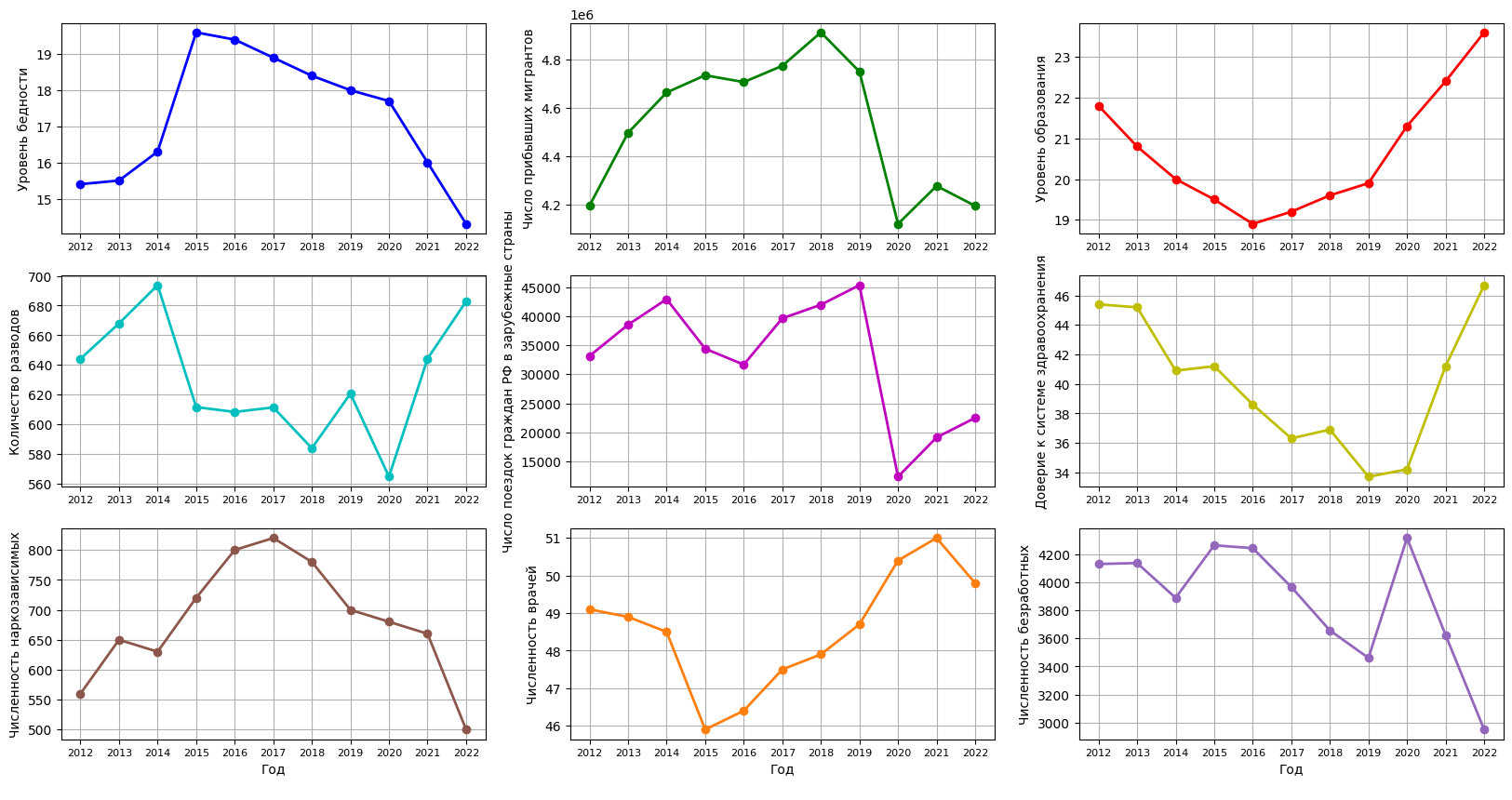


Рисунок 4 – Динамика численности населения с денежными доходами ниже границы бедности

На распространение ВИЧ в России также может влиять количество мигрантов из-за высокого уровня миграционной мобильности и различий в распространении ВИЧ в различных регионах. Люди, приезжающие из других стран в Россию, особенно из стран, где наблюдается высокий уровень распространения ВИЧ, могут быть разносчиками вируса. Они также могут сталкиваться с дискриминацией, доступом к медицинским услугам и барьерами в получении информации о профилактике ВИЧ.

Получается, что чем выше число мигрантов в стране, тем выше число новых случаев заражения ВИЧ.

Как видно из графика динамики числа мигрантов в России за период с 2012 по 2022 год (Рисунок 5), с 2012 года наблюдался рост мигрантов в стране, а после 2018 года – спад. В 2021 году численность мигрантов поднялась, но на момент 2022 года вновь опустилась на уровень 2012. Получается, что данное наблюдение соответствует динамике распространения ВИЧ в России.

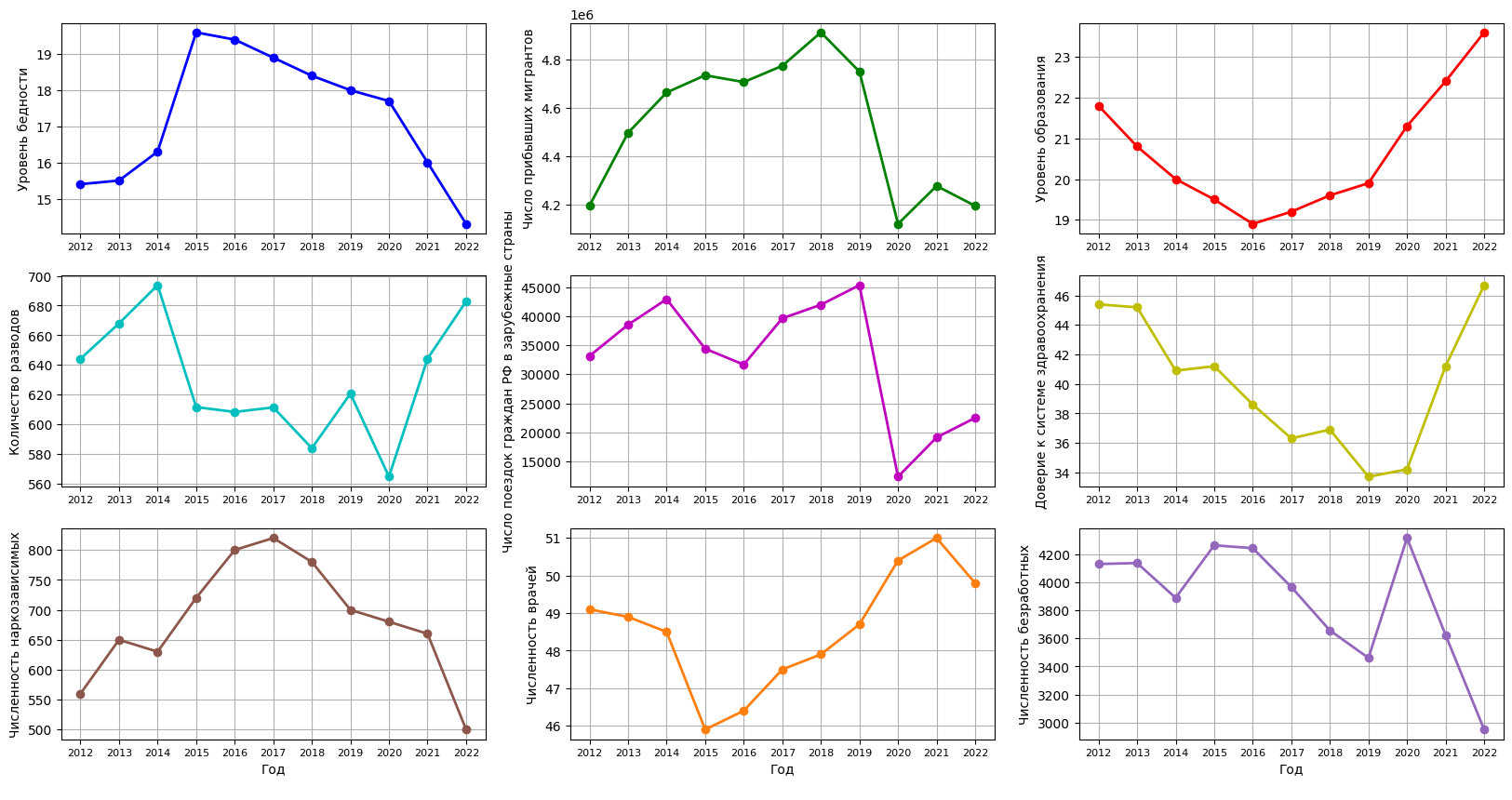


Рисунок 5 – Динамика численности прибывших мигрантов

Уровень образования в России имеет прямое влияние на распространение ВИЧ в стране по нескольким причинам:

* Студенты, проходящие образовательные программы, получают информацию о ВИЧ, методах защиты и профилактике этого вируса. Чем больше студентов охвачено образовательными программами, тем больше людей получат необходимые знания о ВИЧ, что способствует снижению распространения вируса.
* Высшее и среднее профессиональное образование формирует социально ответственное поведение у студентов, что в свою очередь может привести к более осознанному образу жизни.
* Образованные люди более склонны к поиску медицинской помощи и следованию рекомендациям врачей, что способствует более эффективной профилактике и лечению ВИЧ.

Таким образом, увеличение валового коэффициента охвата населения образовательными программами среднего профессионального образования и высшего образования может способствовать снижению распространения ВИЧ в России через повышение уровня информированности, социальной ответственности и доступа к медицинской помощи.

Исходя из графика динамики уровня охвата образовательными программами среднего профессионального образования и высшего образования (Рисунок 6), видно, что с 2012 года уровень образованности в стране падал, в то время как уровень ВИЧ рос. С 2017 года уровень образования начал вновь расти, в то время как уровень заболеваемости ВИЧ падал. Следовательно, распространение ВИЧ в России сильно зависит от уровня охвата образованием населения.

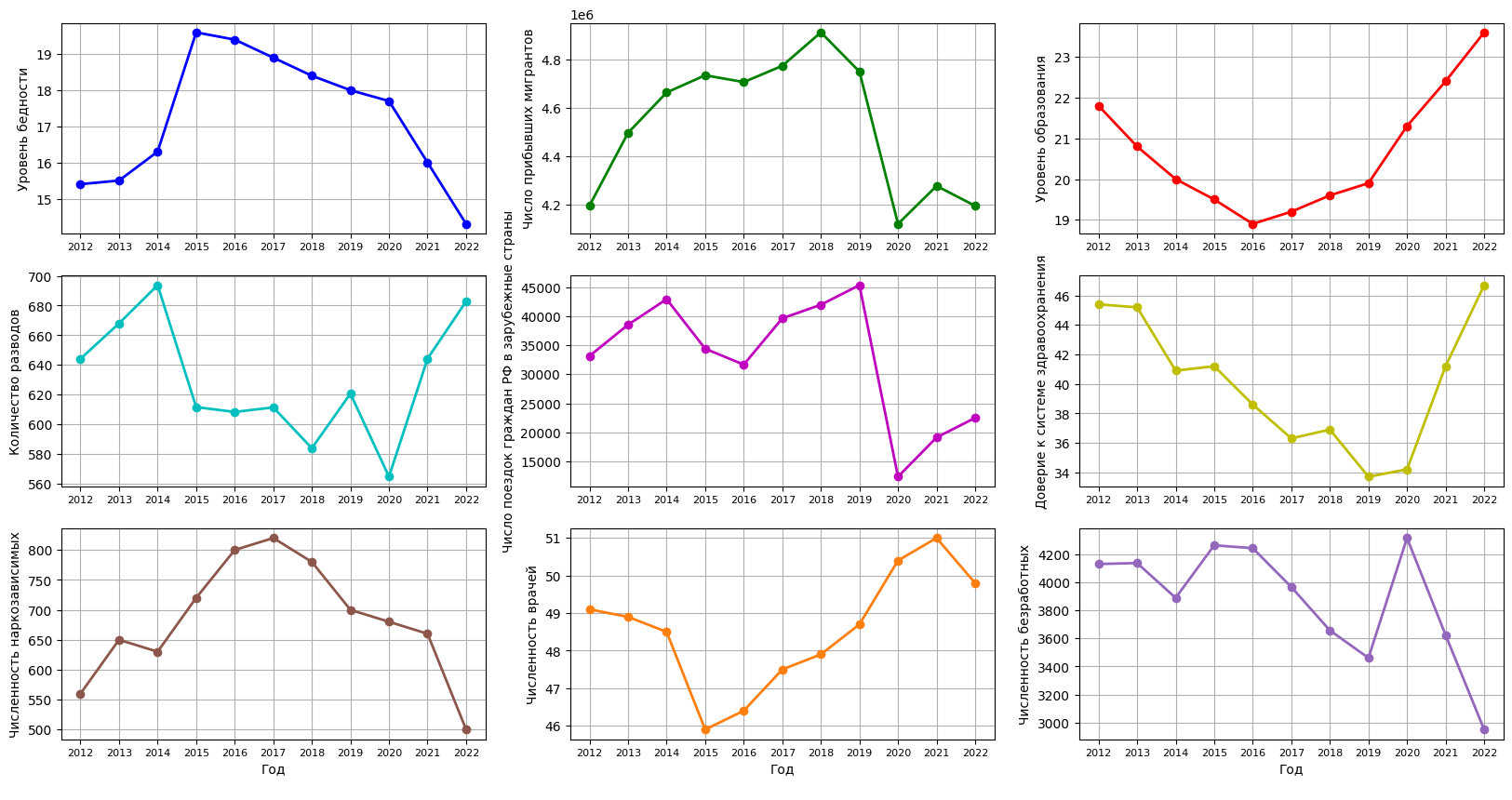


Рисунок 6 – Динамика уровня образования в России

Следует предположить, что ВИЧ также может зависеть от количества разводов по стране. Однако количество разводов может влиять, как негативно на распространение ВИЧ, так и позитивно.

С одной стороны, развод может вызвать эмоциональную травму, депрессию и стресс у людей, что в свою очередь может привести к изменению их поведения, включая более рискованные сексуальные практики или применение наркотических веществ, что увеличивает вероятность заражения ВИЧ.

С другой стороны, с учетом социокультурных особенностей России, можно отметить, что количество разводов в стране может оказать положительное воздействие на уровень распространения ВИЧ. В российском обществе брак имеет большое значение для многих людей, однако часто люди находят себя в несчастливых браках и начинают искать новые отношения или увлечения, оставаясь при этом в формально зарегистрированных браках. Такие ситуации могут способствовать распространению инфекции ВИЧ через непостоянные половые контакты или изменяющиеся партнерские отношения. Следовательно, увеличение числа разводов может привести к снижению числа таких рисковых сексуальных поведений и увеличению доверия в отношениях между людьми, что в свою очередь поможет в ограничении распространения ВИЧ-инфекции в обществе.

На основании графика динамики числа разводов (Рисунок 7) в России, видно, что зависимость распространения ВИЧ скорее достаточно косвенно зависит от числа разводов в стране. Однако, можно увидеть, что уменьшение числа разводов приводит к уменьшению распространения ВИЧ-инфекции в России.

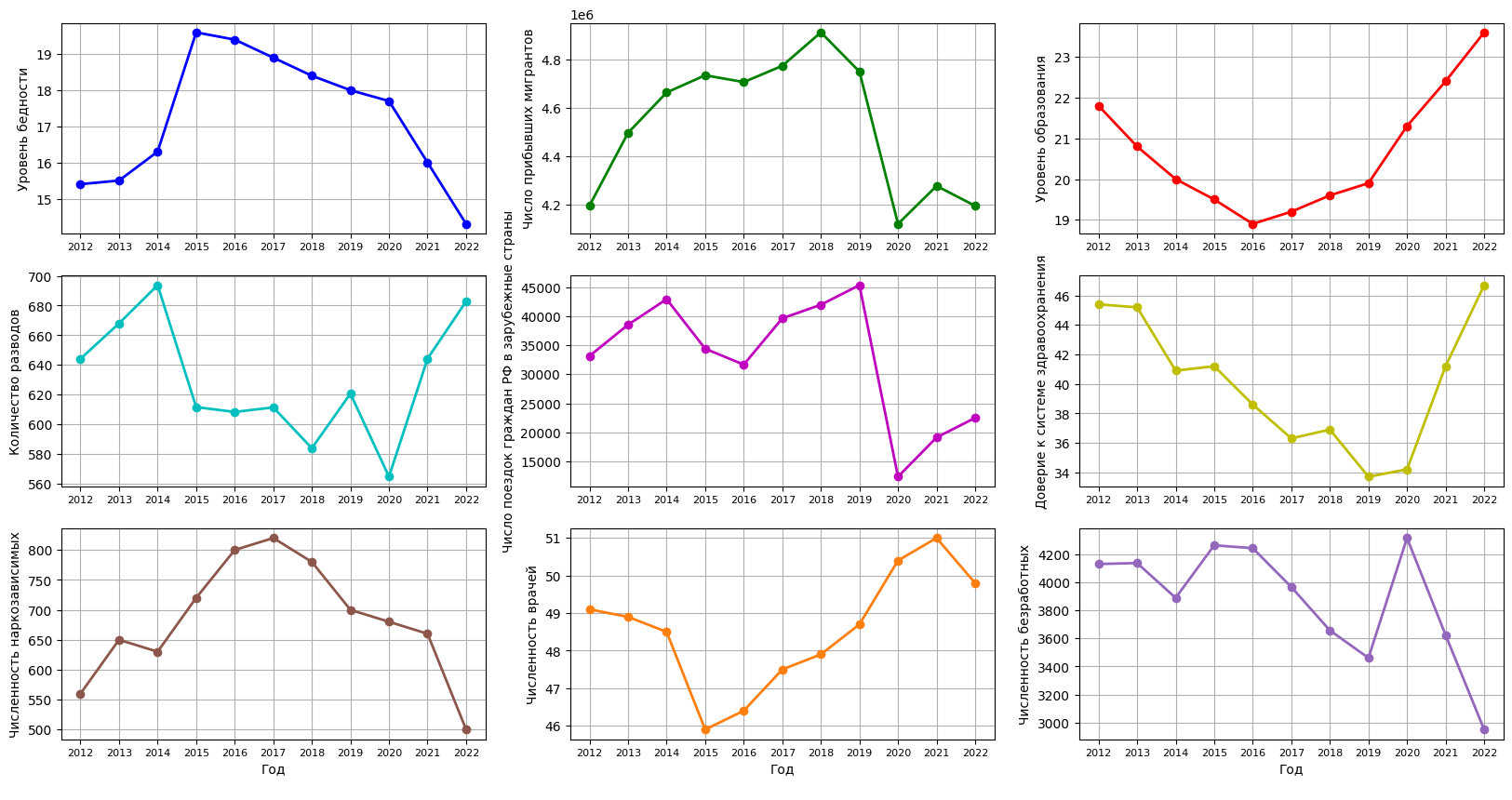


Рисунок 7 – Динамика количества разводов в России

Аналогично численности мигрантов следует предположить зависимость распространения ВИЧ-инфекции в России от численности россиян, путешествующих в зарубежные страны. Ведь во время туристических поездок люди могут подцепить вирус. Например, некоторые люди могут использовать инъекционные наркотики во время поездок, особенно в странах, где это более доступно. К тому же, во время отдыха или путешествия люди могут быть менее осторожными и не придавать должного значения профилактическим мерам.

При анализе графика динамики числа выездных туристских поездок граждан России в зарубежные страны (Рисунок 8) можно увидеть несильную связь между распространением ВИЧ и числом выездных туристских поездок. Однако, все-таки можно предположить, что с увеличением поезд увеличивается и число новых заражений ВИЧ в России.

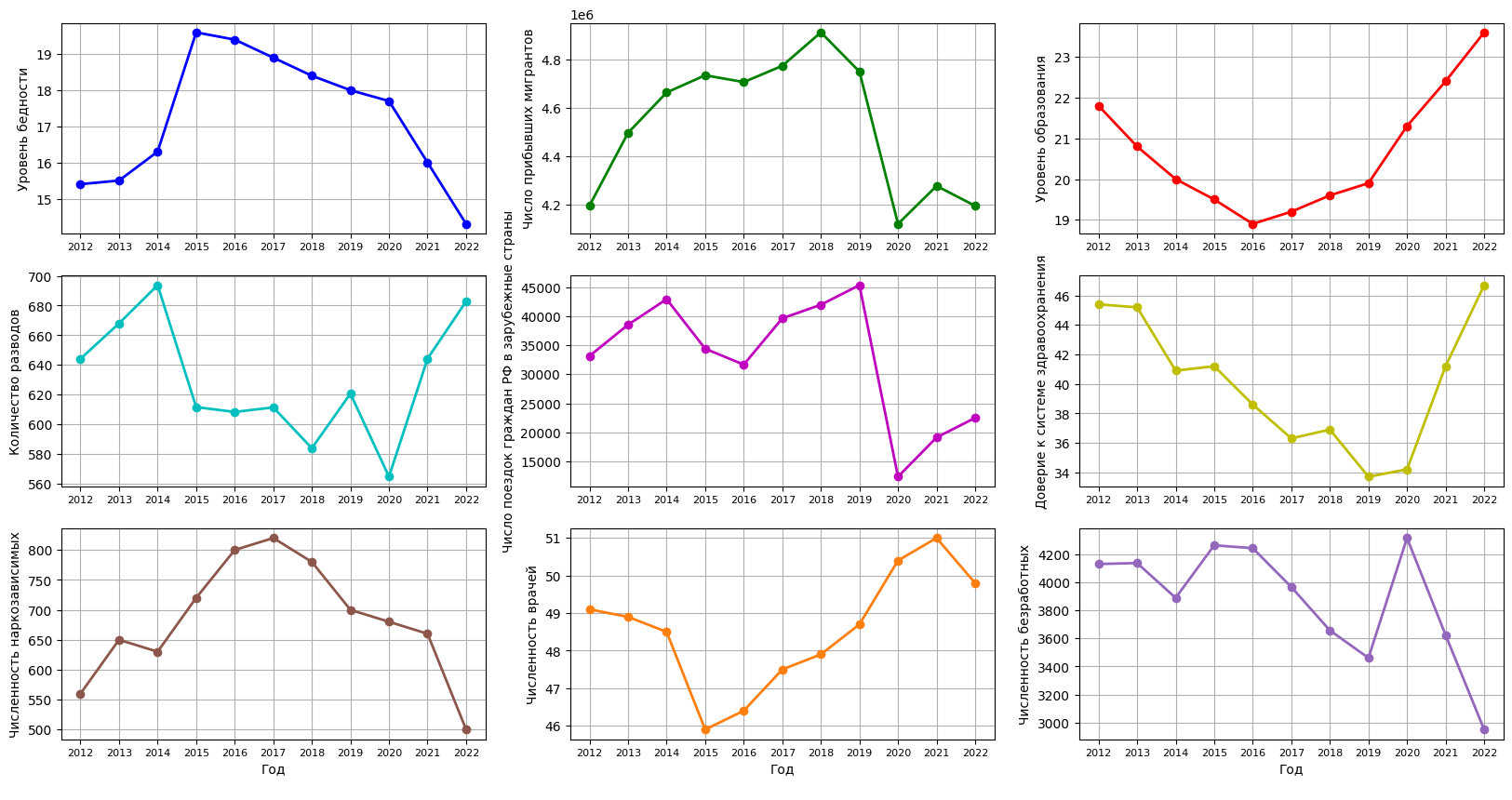


Рисунок 8 – Динамика числа выездных туристских поездок граждан России в зарубежные страны

Уровень доверия россиян к системе здравоохранения может оказывать значительное влияние на распространение ВИЧ в России: если уровень доверия к системе здравоохранения низок, люди могут быть менее склонны обращаться за помощью и информацией о ВИЧ. Это приведет к недостаточной осведомленности о методах предотвращения заражения ВИЧ и снизит доступность тестирования на ВИЧ, что в свою очередь поспособствует распространению инфекции.

Низкое доверие к системе здравоохранения также может создавать психологический барьер для людей, которые нуждаются в помощи или лечении ВИЧ. Они могут опасаться стигматизации, дискриминации или непрофессионального отношения со стороны медицинских работников, что может оттолкнуть их от обращения за помощью.

По графику динамики уровня доверия россиян к системе здравоохранения (Рисунок 9) видно, что уровень доверия с 2012 года по 2019 постепенно снижался (в 2019 достиг своего пика), а к 2022 значительно возрос, став самым большим показателем за 11 лет. Исходя из этого, можно сделать вывод о достаточно сильной зависимости между распространением ВИЧ-инфекции и уровнем доверия к системе здравоохранения.

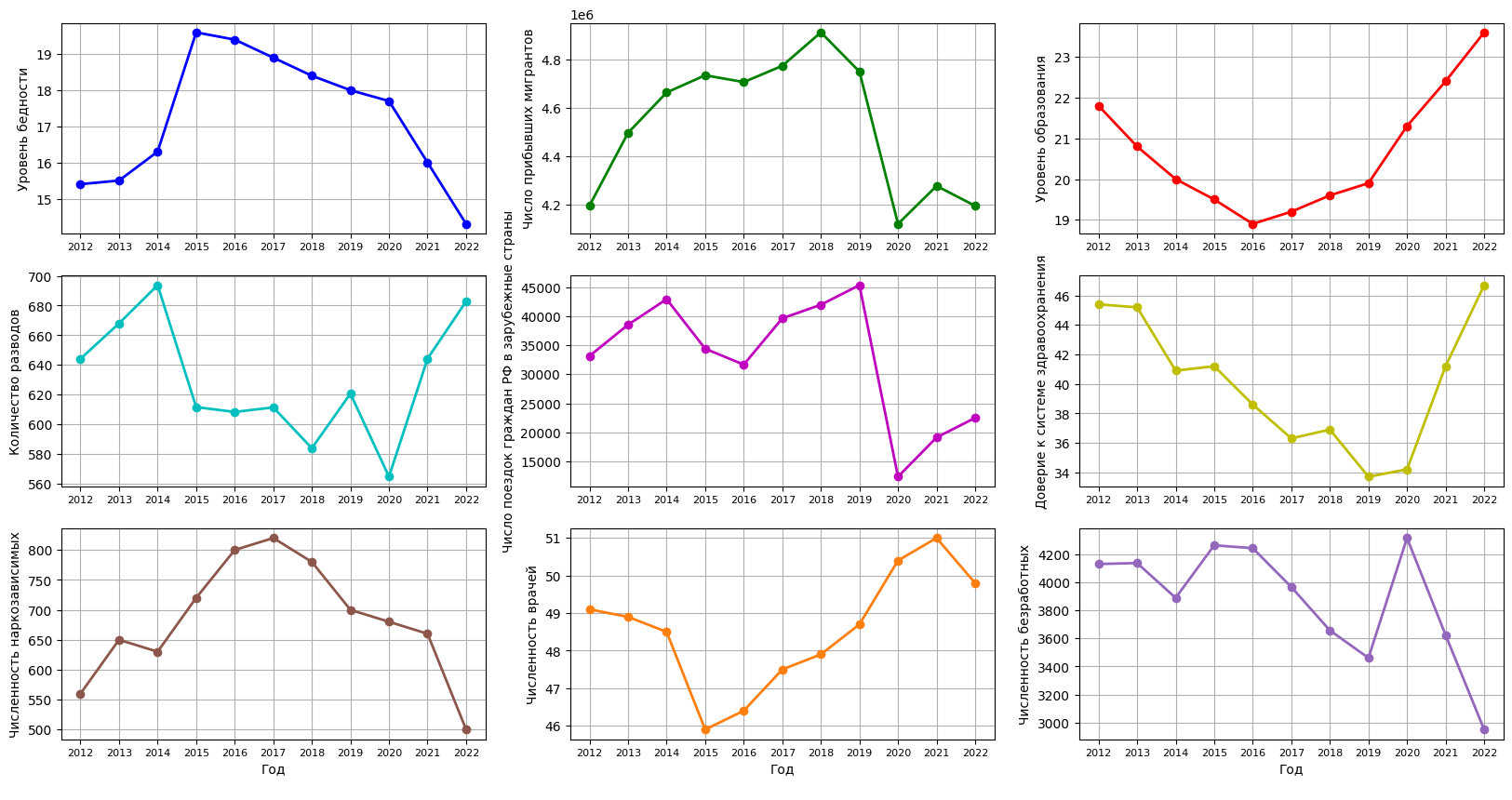


Рисунок 9 – Динамика уровня доверия россиян к системе здравоохранения

Наиболее важным факторов развития ВИЧ является численность наркозависимых людей в России, так как очень большой процент заражаемости ВИЧ у наркоманов, применяющих ПАВ инъекционно, особенно через нестерильные иглы. Использование общих игл и шприцев может способствовать передаче инфекции.

Так как точную статистику наркоманов в России определить невозможно, то для анализа данной зависимости следует использовать данные о численности наркоманов, стоящих на медицинском учете.

Таким образом, численность наркозависимых, стоящих на медицинском учете, оказывает непосредственное влияние на распространение ВИЧ в России. Это можно определить и исходя из графика динамики численности наркозависимых, стоящих на медицинском учете (Рисунок 10). С 2012 года и до 2017 наблюдался значительный рост числа наркозависимых, стоящих на медицинском учете в России. На 2017 год пришелся пик численности, аналогично и с ВИЧ-инфекцией, а после 2017 года численность наркоманов стремительно падала. Поэтому, численность наркоманов является одним из главных фактором распространения ВИЧ в России.

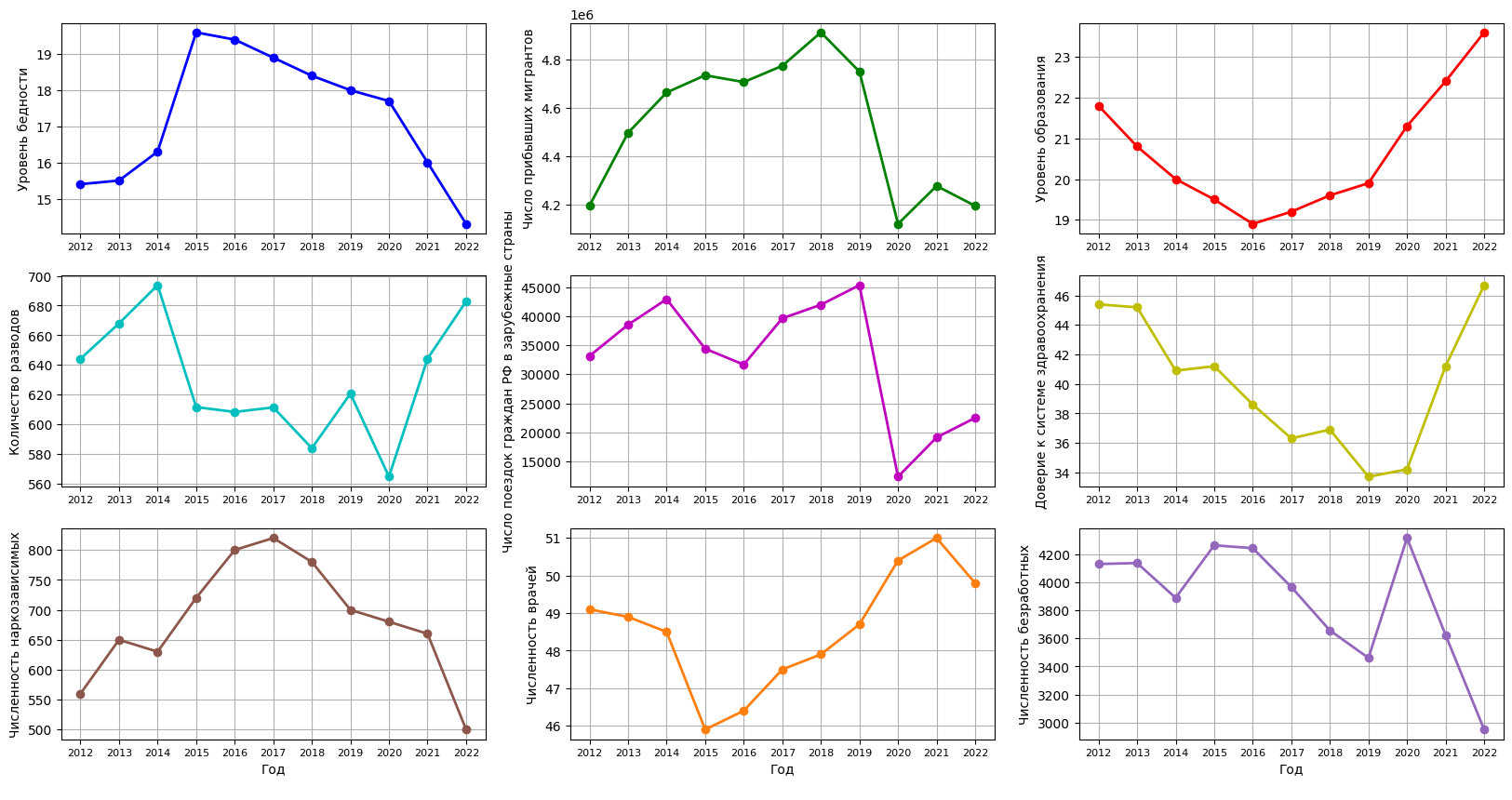


Рисунок 10 – Динамика численности наркозависимых, стоящих на медицинском учете

Численность врачей всех специальностей также оказывает влияние на распространение ВИЧ в России. Чем больше численность врачей, тем больше доступ к тестированию на ВИЧ и консультированию по профилактике и лечению этого заболевания, что способствует выявлению большего числа случаев инфицирования и предоставлению необходимой информацию по профилактике и лечению. Врачи также могут принимать участие в предоставлении доступа к антивирусной терапии для ВИЧ-инфицированных пациентов, что поможет снизить вероятность передачи инфекции другим людям.

Таким образом, из графика динамики численности врачей всех специальностей (Рисунок 11) видно, что численность врачей всех специальностей оказывает влияние на распространение ВИЧ в России: увеличение численности врачей и улучшение их доступности для населения способствует уменьшению распространения ВИЧ.

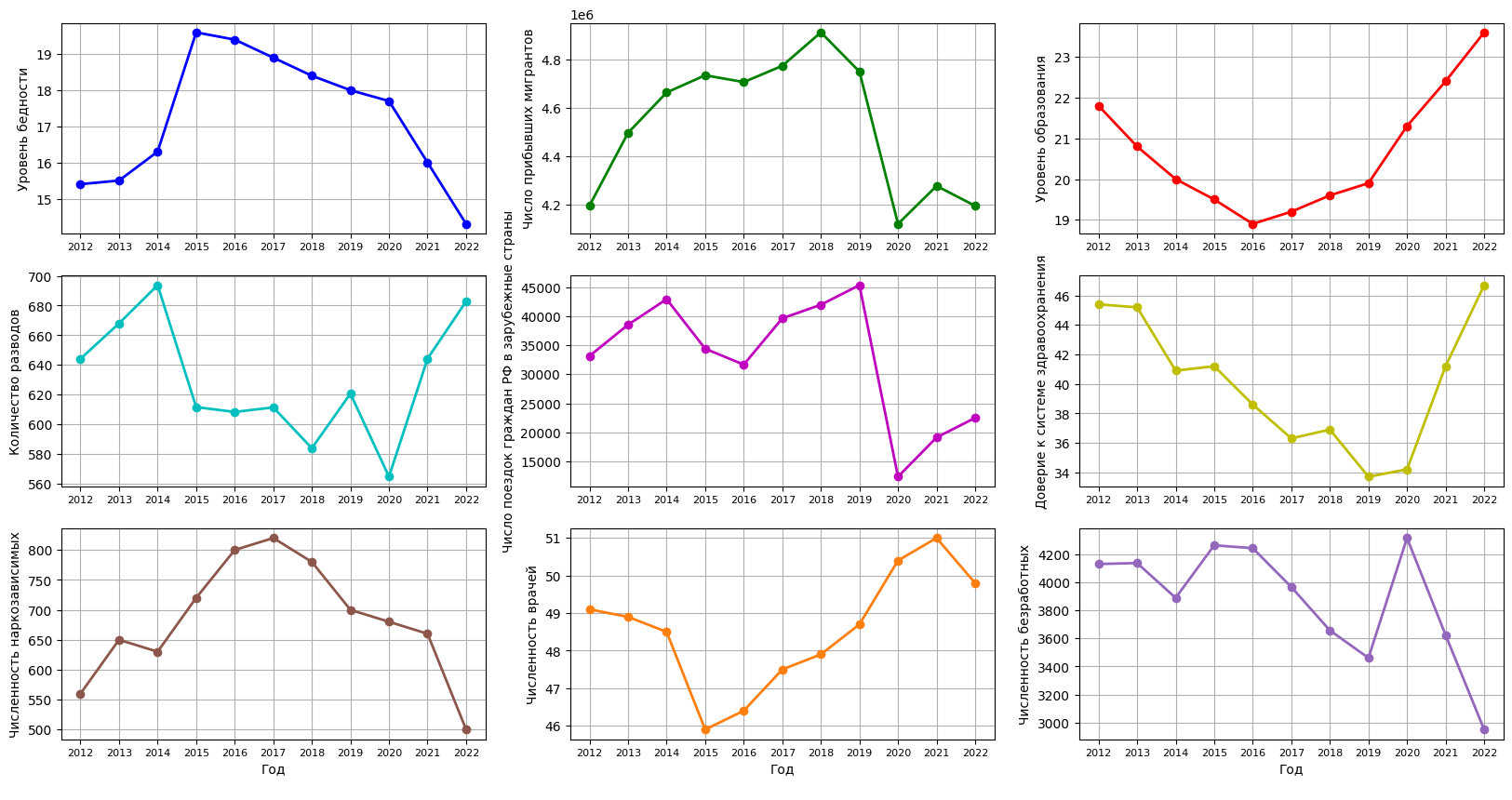


Рисунок 11 – Динамика численности врачей всех специальностей

Кроме того, можно предположить, что численность безработных также оказывает влияние на распространение ВИЧ в России. Это объясняется несколькими причинами:

* Люди, находящиеся без работы, могут столкнуться с повышенным риском инфицирования ВИЧ из-за более высокой вероятности вовлечения в рискованные поведенческие практики, а также употребления наркотиков.
* Безработные люди могут сталкиваться с ограничениями в доступе к медицинской помощи, включая тестирование на ВИЧ, консультирование и получение антивирусной терапии. Это может привести к недооценке своего статуса ВИЧ-инфицированного и увеличению риска передачи инфекции другим людям.
* Безработные люди могут сталкиваться с ухудшением своего социально-экономического положения, что может привести к увеличению уязвимости перед ВИЧ из-за недостатка доступных средств для обеспечения себя и своей семьи.

Однако если посмотреть на график динамики численности безработных (Рисунок 12), то можно сделать вывод, что данный показатель практически не влияет на развитие ВИЧ в России. Безработица, конечно, может являться причиной вовлечения некоторых людей в антисоциальную жизнь, что способствует большому риску заражения ВИЧ, но никак не является влиятельным фактором для распространения ВИЧ-инфекции.

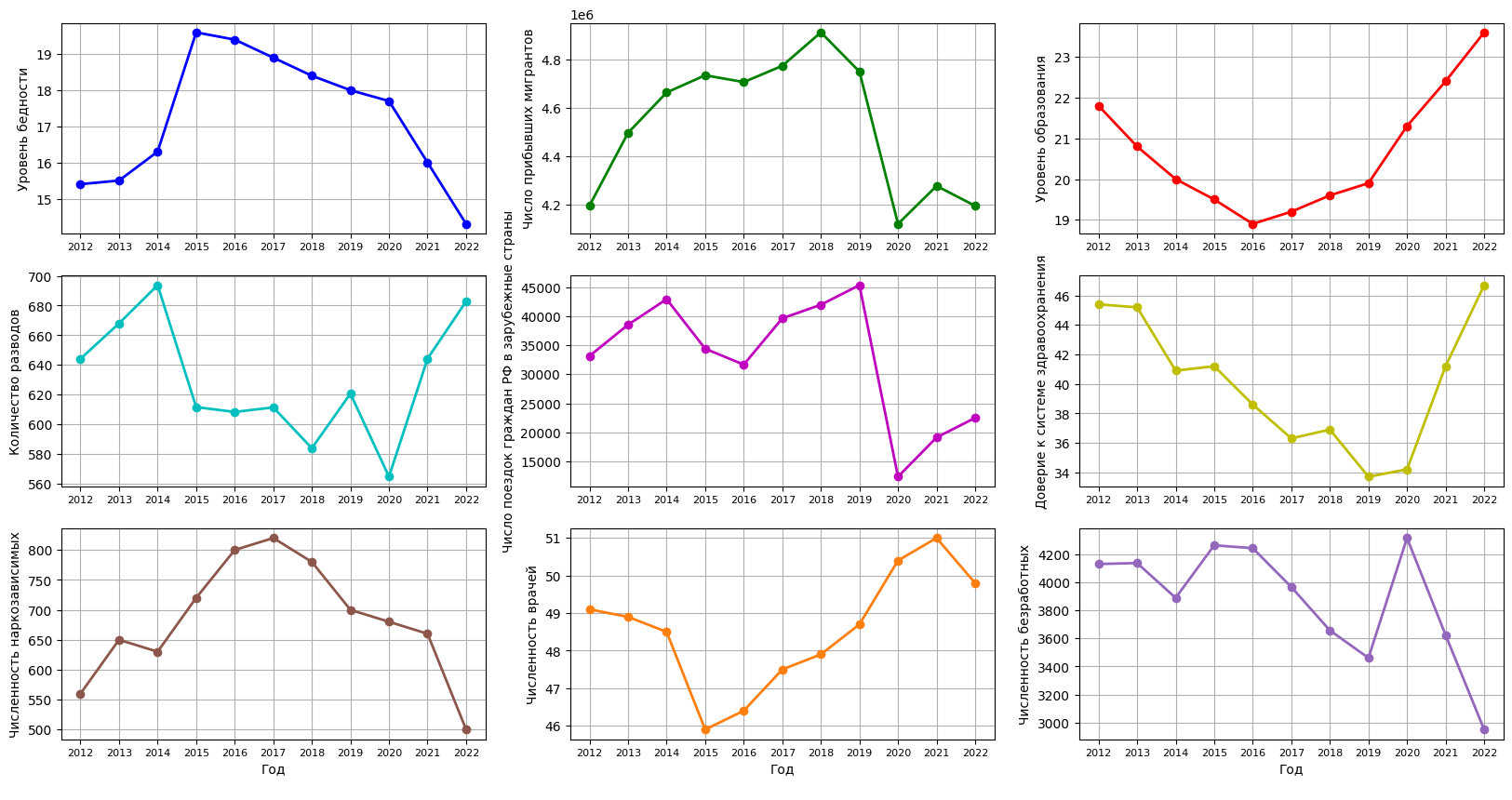


Рисунок 12 – Динамика численности безработных

Все данные, на основе которых были построены графики, были взяты из статистических сборников Росстата.

На основе проведенного анализа, можно сделать вывод, что наиболее влиятельными факторами на развитие ВИЧ в России являются численность мигрантов, уровень образования, уровень доверия россиян к системе здравоохранения и численность наркозависимых, стоящих на медицинском учете.

1. Постановка задачи моделирования

Данное моделирование основывается на учёте факторов, влияющих на распространение ВИЧ в России, таких как уровень бедности, миграционные потоки, уровень образования, социально-экономические изменения, демографические показатели, здравоохранение, данные по наркозависимости, общественные мнения и другие социальные и экономические факторы. Хотя снижение числа новых случаев заражения ВИЧ-инфекцией является желательным результатом, нельзя игнорировать сложность и многообразие факторов, влияющих на этот показатель.

С учетом этого, задача моделирования развития ВИЧ в России заключается в построении математической модели, которая позволит прогнозировать число новых случаев заражения ВИЧ-инфекцией в будущем с целью принятия мер, направленных на предотвращение распространения инфекции и улучшение эффективности программ по ее профилактике и лечению.

Модель прогноза должна учитывать динамику факторов и их взаимодействие.

1. Моделирование статистических зависимостей
2. Формализация и классификация переменных

Входные переменные – факторы, влияющие на распространение ВИЧ-инфекции в России:

* x1: численность населения с денежными доходами ниже границы бедности, млн. чел.;
* x2: число прибывших в страну людей (мигрантов), чел.;
* x3: валовой коэффициент охвата населения образовательными программами среднего профессионального образования и высшего образования, %;
* x4: количество разводов, тыс. ед.;
* x5: число выездных туристских поездок граждан России в зарубежные страны, тыс. чел.;
* x6: уровень доверия россиян к системе здравоохранения, %;
* x7: численность наркозависимых, стоящих на медицинском учете, тыс. чел.;
* x8: численность врачей всех специальностей на 10000 человек населения;
* x9: численность безработных, тыс. чел.

Выходная переменная:

* y: количество новых случаев ВИЧ-инфицированных, чел.

Параметры:

* θ0 – номинальное количество новых случаев заражения ВИЧ;
* θ1 – коэффициент, характеризующий среднее количество новых случаев заражения ВИЧ, приходящийся на единицу численности населения с денежными доходами ниже границы бедности;
* θ2 – коэффициент, характеризующий среднее количество новых случаев заражения ВИЧ, приходящийся на одного мигранта;
* θ3 – коэффициент, характеризующий среднее количество новых случаев заражения ВИЧ, приходящийся на единицу валового коэффициента охвата образовательными программами среднего профессионального образования и высшего образования;
* θ4 – коэффициент, характеризующий среднее количество новых случаев заражения ВИЧ, приходящийся на один развод;
* θ5 – коэффициент, характеризующий среднее количество новых случаев заражения ВИЧ, приходящийся на одну выездную туристскую поездку гражданина России в зарубежные страны;
* θ6 – коэффициент, характеризующий среднее количество новых случаев заражения ВИЧ, приходящийся на единицу уровня доверия россиян к системе здравоохранения;
* θ7 – коэффициент, характеризующий среднее количество новых случаев заражения ВИЧ, приходящийся на одного наркозависимого, стоящего на медицинском учете;
* θ8 – коэффициент, характеризующий среднее количество новых случаев заражения ВИЧ, приходящийся на одного врача;
* θ9 – коэффициент, характеризующий среднее количество новых случаев заражения ВИЧ, приходящийся на одного безработного.

1. Проверка гипотезы о нормальном распределении

В статистическом анализе данных важным этапом является проверка гипотезы о нормальности распределения переменных. Распределение выходной величины y (количество новых случаев ВИЧ-инфицированных, чел.) играет ключевую роль в понимании структуры данных и может влиять на выбор статистических методов для дальнейшего анализа.

Нормальным распределением (распределением Гаусса) – называется распределение вероятностей непрерывной случайной величины, описываемое плотностью

Чтобы задать нормальное распределение, достаточно знать два параметра: α – математическое ожидание и σ – среднее квадратическое отклонение нормального распределения.

Для проверки гипотезы о нормальном распределении числа новых случаев заражения ВИЧ-инфекцией за период с 2012 по 2022 год воспользуемся правилом трех сигм и критерием Пирсона.

Для начала составим интервальный ряд распределения (Таблица №5), в котором левая граница – 2011 год (если быть точнее, то 31.12.2011), правая граница – 2022 год (31.12.2022), а длина интервала – 1 год.

Таблица №5 – Интервальный ряд распределения зарегистрированных больных ВИЧ

|  |  |
| --- | --- |
| **Период** | **Зарегистрировано больных ВИЧ** |
| 2011-2012 | 69280 |
| 2012-2013 | 77896 |
| 2013-2014 | 85252 |
| 2014-2015 | 88040 |
| 2015-2016 | 89155 |
| 2016-2017 | 104402 |
| 2017-2018 | 101345 |
| 2018-2019 | 94668 |
| 2019-2020 | 88154 |
| 2020-2021 | 71019 |
| 2021-2022 | 63150 |

Теперь определим асимметрию по формуле:

Асимметрия указывает на отличие кривой распределения случайной величины от нормальной кривой.

Полученное значение: -0.157

И эксцесс по формуле:

Эксцесс указывает на остроту вершины кривой распределения по сравнению с нормальной кривой.

Полученное значение: -1.016

Перейдем к проверке гипотезы о нормальном распределении с помощью правила трёх сигм.

Правило трех сигм:

Случайная величина подчинена распределению N(a, σ), тогда:

* около 68% ее реализаций лежат в интервале (a − σ, a + σ)
* около 95% ее реализаций лежат в интервале (a − 2σ, a + 2σ)
* около 99,7% ее реализаций лежат в интервале (a − 3σ, a + 3σ)

Иначе говоря, если случайная величина распределена нормально, то абсолютная величина ее отклонения от математического ожидания не превосходит утроенного среднего квадратического отклонения.

По правилу трех сигм получилось, что:

* В пределах 1 сигмы находится 56,11% данных;
* В пределах 2 сигм ~100% данных;
* В пределах 3 сигм охвачено ~100% данных.

Данный анализ показал, что распределение числа новых случаев заражения ВИЧ-инфекцией за период с 2012 по 2022 год не является нормальным, так как гипотеза о нормальном распределении отвергается.

На основе данных из таблицы №5 построим гистограмму (Рисунок 13).

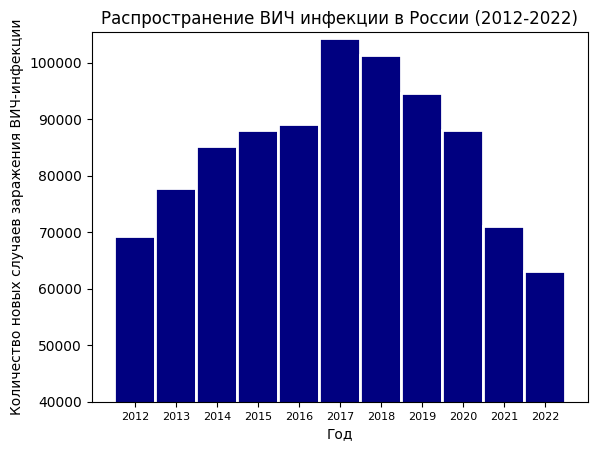


Рисунок 13 – Гистограмма распределения новых случаев заражения ВИЧ-инфекции

С использованием интервального ряда (Таблица №5) проверим гипотезу о нормальном распределении по критерию Пирсона при уровне значимости α = 0.05. Для этого определим теоретическое число заболеваний ВИЧ-инфекцией, характерное для функции нормального распределения (Таблица №6).

Число степеней свободы:

По критерию Пирсона:

Наблюдаемое значение: 128306.07

Критическое значение: 15.51

Так как 15.51 значительно меньше, чем 128306.07 (критическое значение меньше наблюдаемого), то гипотеза о нормальном распределении отвергается.

Таблица №6 – Теоретическое число заболеваний ВИЧ

|  |  |  |
| --- | --- | --- |
| **Период** | **Число заболеваний ВИЧ** | **Теоретическое число заболеваний ВИЧ** |
| 2011-2012 | 69280 | 30637.137978 |
| 2012-2013 | 77896 | 51035.311193 |
| 2013-2014 | 85252 | 75850.853154 |
| 2014-2015 | 88040 | 100581.313004 |
| 2015-2016 | 89155 | 118998.437587 |
| 2016-2017 | 104402 | 125612.351503 |
| 2017-2018 | 101345 | 118301.584173 |
| 2018-2019 | 94668 | 99406.756290 |
| 2019-2020 | 88154 | 74526.096250 |
| 2020-2021 | 71019 | 49850.322394 |
| 2021-2022 | 63150 | 29750.529353 |

Осуществим выравнивание статистического ряда с помощью нормального закона, найдем точечные оценки нормального распределения методом моментов и построим кривую Гаусса, аппроксимировав ею гистограмму частот (Рисунок 14).

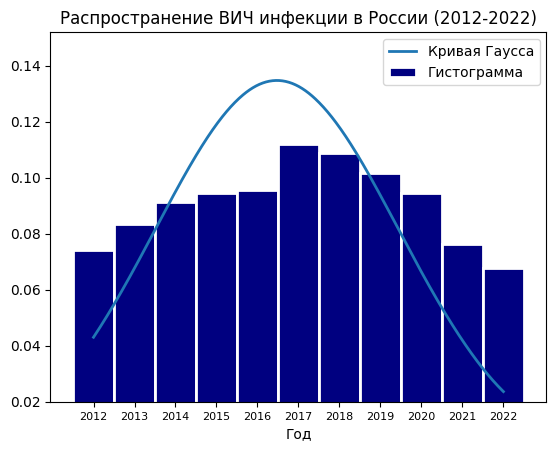


Рисунок 14 – Кривая Гаусса

Исходя из гистограммы с кривой Гаусса (Рисунок 14), можно убедиться в отсутствии нормального распределения числа новых случаев заражения ВИЧ-инфекцией за период с 2012 по 2022 год.

1. Корреляционный анализ

Корреляционный анализ — это метод исследования, направленный на выявление степени взаимосвязи между двумя или более переменными. Этот метод используется для оценки, насколько изменения в одной переменной соотносятся с изменениями в другой. Ключевой мерой в корреляционном анализе является коэффициент корреляции, который измеряет степень линейной зависимости между двумя переменными.

Ковариация (корреляционный момент) – мера линейной зависимости случайных величин, определяется по формуле:

Коэффициент корреляции Пирсона был введен для устранений недостатка ковариации. Он также отражает линейную зависимость между двумя величинами:

Где – средние квадратические отклонения по x и y.

Изменяется в пределах от -1 до +1, причем чем ближе коэффициент корреляции к 1, тем сильнее зависимость между переменными. Значение 0 указывает на отсутствие линейной связи.

Положительная корреляция (прямая) возникает при одновременном изменении двух переменных величин в одинаковых направлениях (в положительном или отрицательном). Корреляция отрицательная (обратная). Если изменения одной величины приводит противоположному изменению другой.

Для оценки силы связи коэффициентов корреляции используется шкала Чеддока. Она определяет тесноту связи следующим образом:

0 – полное отсутствие связи;

0 – 0.3 – очень слабая;

0.3 – 0.5 – слабая;

0.5 – 0.7 – средняя;

0.7 – 0.9 – высокая;

0.9 – 1 – очень высокая.

Вычислим коэффициент корреляции Пирсона (Таблица №7) для каждого фактора x1…x9, от которого зависит чисто новых заражений ВИЧ в РФ (пояснение факторов указано в п. 2.1). Тем самым проведем анализ связи между числом новых заражений ВИЧ и факторами.

Таблица №7 – Анализ зависимости заражаемости ВИЧ инфекции в России от ряда факторов

|  |  |  |
| --- | --- | --- |
| **Факторы** | **Коэффициент корреляции** | **Сила связи** |
| Численность населения с денежными доходами ниже границы бедности | 0.83 | Высокая |
| Число прибывших в страну людей (мигрантов) | 0.79 | Высокая |
| Валовой коэффициент охвата образовательными программами среднего  профессионального образования и высшего образования | -0.86 | Высокая |
| Количество разводов | -0.63 | Средняя |
| Число выездных туристских поездок граждан России в зарубежные страны | 0.51 | Средняя |
| Уровень доверия россиян к системе здравоохранения | -0.82 | Высокая |
| Численность наркозависимых, стоящих на медицинском учете | 0.88 | Высокая |
| Численность врачей всех специальностей на 10000 человек населения | -0.55 | Средняя |
| Численность безработных | 0.29 | Очень слабая |

Таким образом, связь между всеми факторами, кроме численности безработных, и числом новых заражений ВИЧ высокая или средняя. А вот связь между численностью безработных и числом новых заражений ВИЧ – очень слабая (0.29). Это указывает на то, что данный параметр почти не оказывает влияние на распространение ВИЧ в России.

Самая сильная прямая связь наблюдается между числом новых заражений ВИЧ и параметром «Численность наркозависимых, стоящих на медицинском учете». Коэффициент корреляции – 0.88. Получается, что при увеличении численности наркоманов, стоящих на медицинском учете, увеличивается и число новых заражений ВИЧ-инфекцией в России.

Также высокое значение коэффициента корреляции имеет обратная связь между числом новых заражений ВИЧ и валовым коэффициентом охвата образовательными программами среднего профессионального образования и высшего образования – -0.86. Получается, что чем выше уровень образованных людей в стране, тем ниже численность новых заражений ВИЧ.

Близкое значение имеет сила связи между числом новых заражений ВИЧ и численностью населения с денежными доходами ниже границы бедности (0.83) – прямая, а также уровнем доверия россиян к системе здравоохранения (-0.82) – обратная.

Также высокий коэффициент корреляции присутствует между числом новых заражений ВИЧ и числом мигрантов – 0.79. При этом связь прямая, а это значит, что с большим потоком мигрантов в России возрастет и заболеваемость ВИЧ.

Связь между числом новых заражений ВИЧ и параметром «Количество разводов» обратная (-0.63), однако является средней, то есть данный параметр оказывает не такое сильное влияние на изменение числа новых заражений ВИЧ-инфекцией в России, как предыдущие.

Это же относится и к параметрам «Число выездных туристских поездок граждан России в зарубежные страны», «Численность врачей всех специальностей на 10000 человек населения». Сила связи между числом новых заражений ВИЧ и этими параметрами близкая. Для числа выездных туристских поездок граждан России в зарубежные страны – прямая (0.51), а для численности врачей всех специальностей на 10000 человек населения – обратная (-0.55).

Таким образом, большинство факторов оказывают сильное влияние на распространение ВИЧ-инфекции в России. Тем не менее, каждый из этих факторов, хоть и по-разному, но влияет на распространение ВИЧ, поэтому при разработке модели прогнозирования очень важно учитывать каждый из них.

Построим корреляционные поля для зависимости между числом новых случаев заражения ВИЧ и каждым параметром (Рисунок 15).

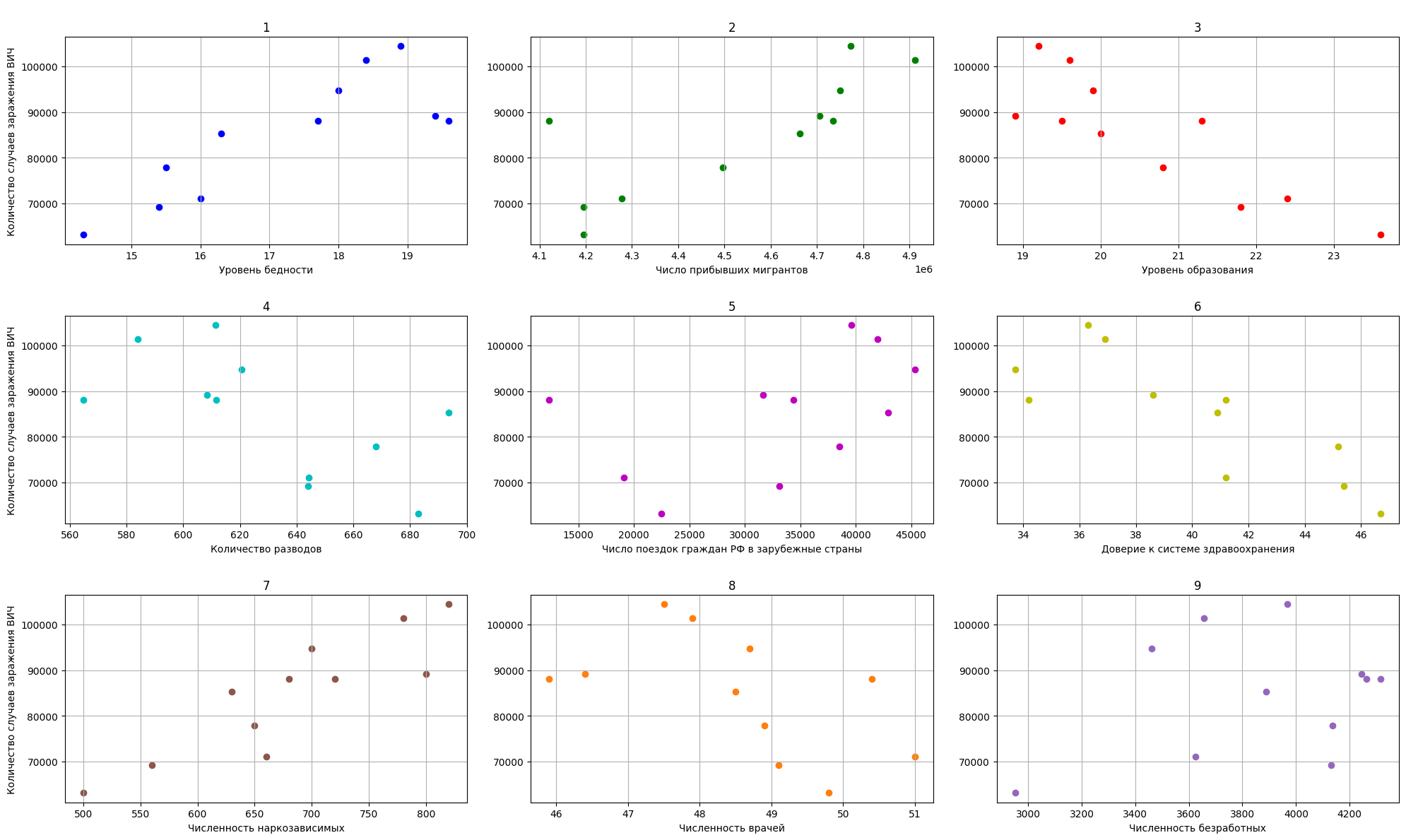


Рисунок 15 – Корреляционные поля для параметров x1…x9

1. Построение регрессионной модели

В контексте современного анализа данных и научных исследований, разработка регрессионных моделей выступает важным инструментом для понимания взаимосвязей между переменными. Такие модели не только позволяют описывать эти взаимосвязи, но и предоставляют возможность создавать прогнозы, основанные на имеющихся данных.

В этом разделе осуществим структурную идентификацию модели, то есть подберем лучшие аппроксимирующие зависимости для переменных, и параметрическую идентификацию.

1. Структурная идентификация модели

Для начала осуществим линейную аппроксимацию зависимости между величинами xi и y. Для этого найдем коэффициенты a и b.

Линейная зависимость:

Формулы для расчета коэффициентов:

Также найдем коэффициент детерминации. Коэффициент достоверности аппроксимации (детерминации) измеряет долю дисперсии зависимой переменной, объясненную моделью. Он принимает значения от 0 до 1 и выражает, какую часть изменчивости зависимой переменной можно объяснить изменениями в независимой переменной.

Коэффициент достоверности аппроксимации R2 вычисляется по формуле:

Таблица №8 – Линейная зависимость между величинами xi и y

|  |  |  |
| --- | --- | --- |
| **Факторы** | **Линейная зависимость между xi и y** | **Коэффициент детерминации** |
| X1 | y(x1) = 6094.935 \* x - 20239.022 | 0.69 |
| X2 | y(x2) = 0.037 \* x - 82375.640 | 0.62 |
| X3 | y(x3) = -7669.190 \* x + 243024.290 | 0.75 |
| X4 | y(x4) = -207.308 \* x + 215428.308 | 0.40 |
| X5 | y(x5) = 0.626 \* x + 64185.040 | 0.26 |
| X6 | y(x6) = -2402.269 \* x + 180916.355 | 0.67 |
| X7 | y(x7) = 118.768 \* x + 3781.851 | 0.78 |
| X8 | y(x8) = -4608.372 \* x + 308517.475 | 0.30 |
| X9 | y(x9) = 9.168 \* x + 49221.404 | 0.08 |

Полученные уравнения линейной зависимости и коэффициенты детерминации представлены в таблице №8.

Теперь осуществим структурную идентификацию модели для каждого фактора, определив вид наилучшей аппроксимирующей прямой.

Таблица №9 – Виды зависимостей

|  |  |
| --- | --- |
| **Номер структуры** | **Структура** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

Наилучшей структурой для описания зависимости между y и x1 оказалась структура под номером 4 (Таблица №9).

Осуществим линеаризацию модели: .

Замена переменных: ;

Линейный вид зависимости:

В соответствии с МНК:

Переход к нелинейной модели:

Вычислим также коэффициент достоверности аппроксимации (детерминации) между величинами u и v.

Коэффициент детерминации: 0.73

Таким образом, коэффициент детерминации показывает, что 73% вариации числа новых случаев заражения ВИЧ-инфекцией обусловлено вариацией уровня бедности в РФ за период с 2012 по 2022 год.

На основе полученных данных построим графически линейную и нелинейную аппроксимацию на корреляционном поле (Рисунок 16). Таким образом, можно подтвердить, что полученная нелинейная модель лучше описывает зависимость между переменными y и x1.

Аналогичным образом, проведем структурную идентификацию для остальных факторов.

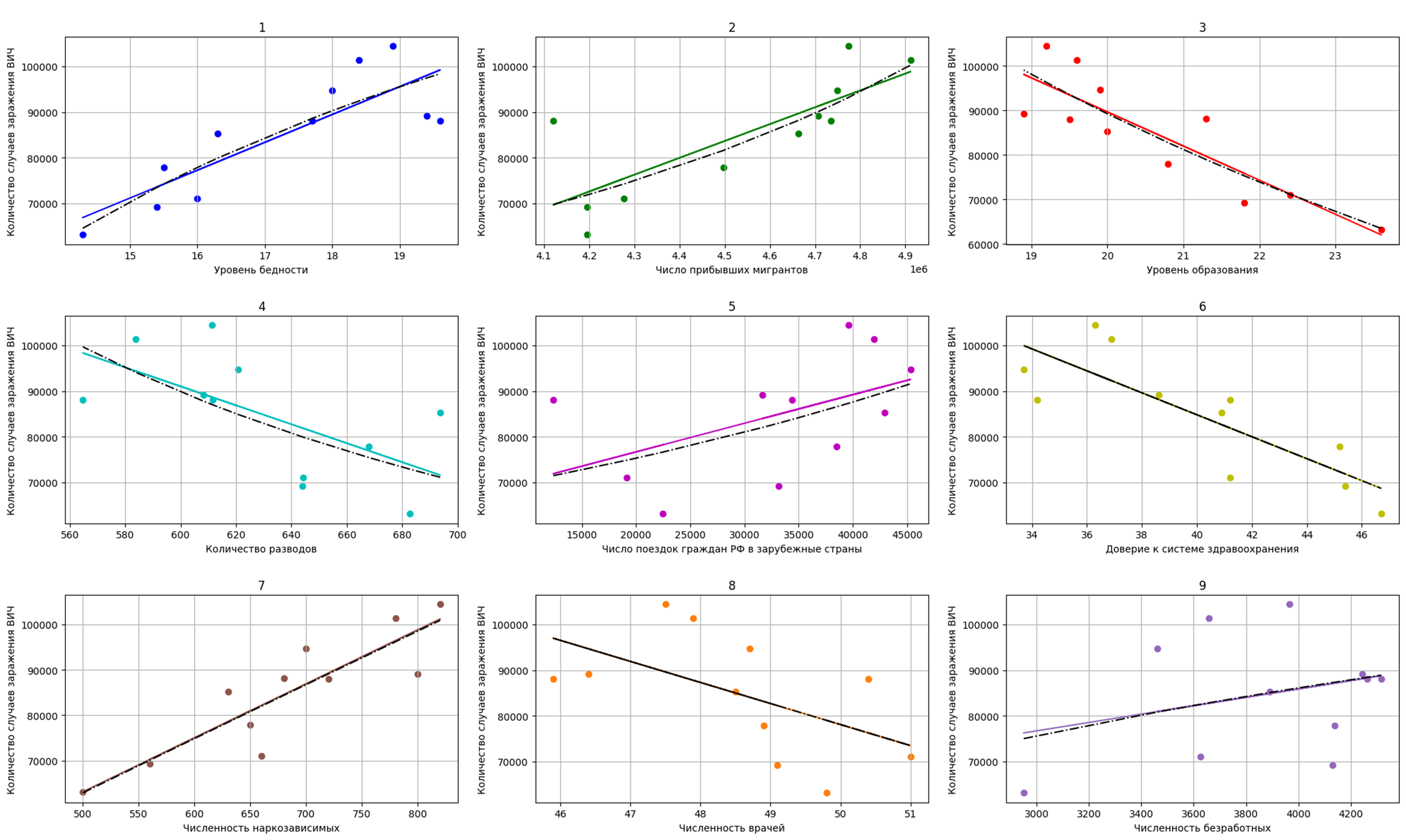


Рисунок 16 – Зависимость между числом новых случаев заражения ВИЧ и уровнем бедности в России

Лучшей структурой для описания зависимости между y и x2 является структура 5.

Осуществим линеаризацию модели:

Замена переменных: ;

Линейный вид зависимости:

В соответствии с МНК:

Переход к нелинейной модели:

Коэффициент детерминации: 0.61

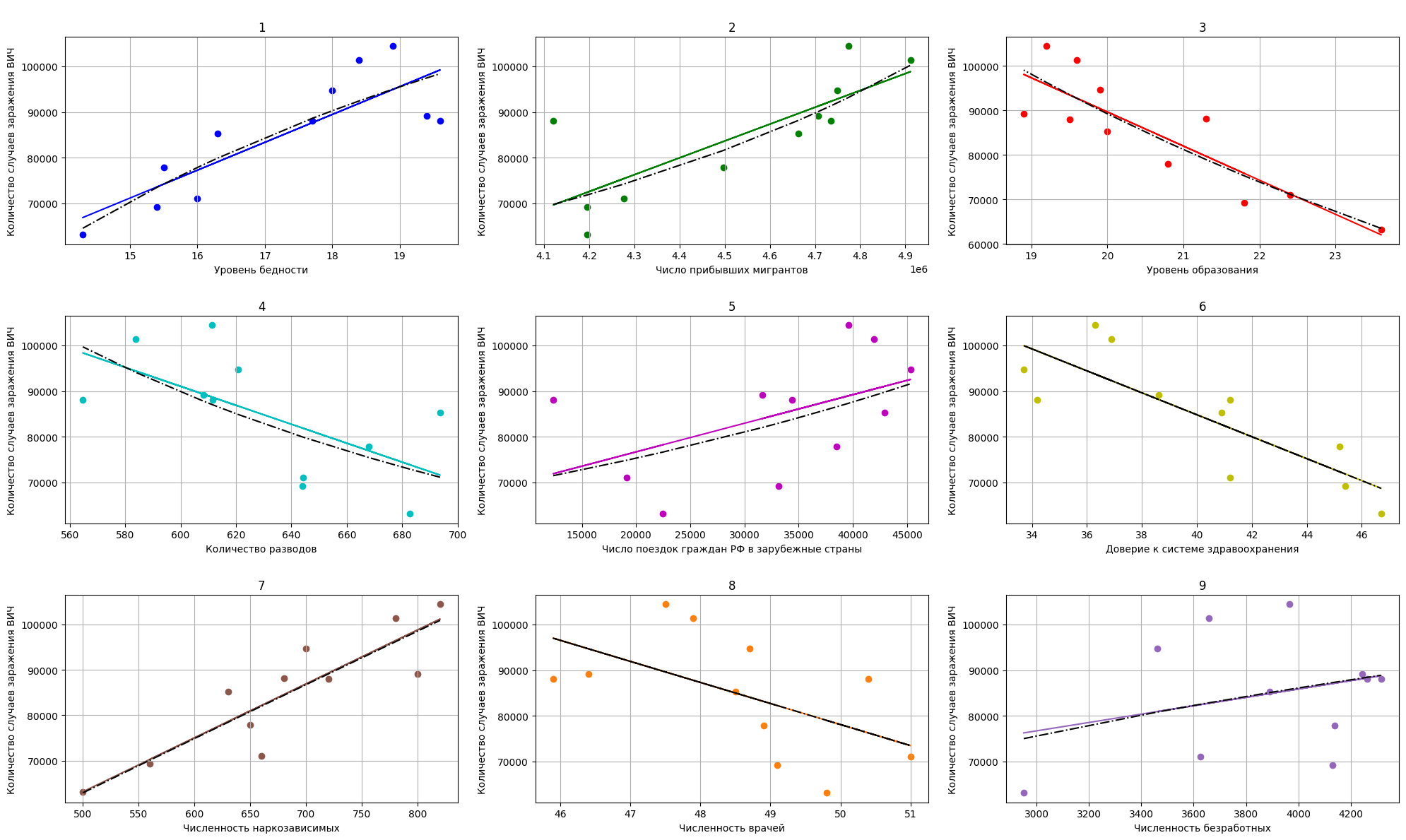


Рисунок 17 – Зависимость между числом новых случаев заражения ВИЧ и числом мигрантов

Таким образом, коэффициент детерминации показывает, что 61% вариации числа новых случаев заражения ВИЧ-инфекцией обусловлен вариацией числа мигрантов за период с 2012 по 2022 год.

На основе полученных данных построим графически линейную и нелинейную аппроксимацию на корреляционном поле (Рисунок 17).

Структурная идентификация для моделирования зависимости между y и x3 принимает вид: . Осуществим линеаризацию модели.

Замена переменных: ;

Линейный вид зависимости:

В соответствии с МНК:

Переход к нелинейной модели:

Коэффициент детерминации: 0.74

Таким образом, коэффициент детерминации показывает, что 74% вариации числа новых случаев заражения ВИЧ-инфекцией обусловлено вариацией уровня образования в РФ за период с 2012 по 2022 год.

На основе полученных данных построим графически линейную и нелинейную аппроксимацию на корреляционном поле (Рисунок 18).

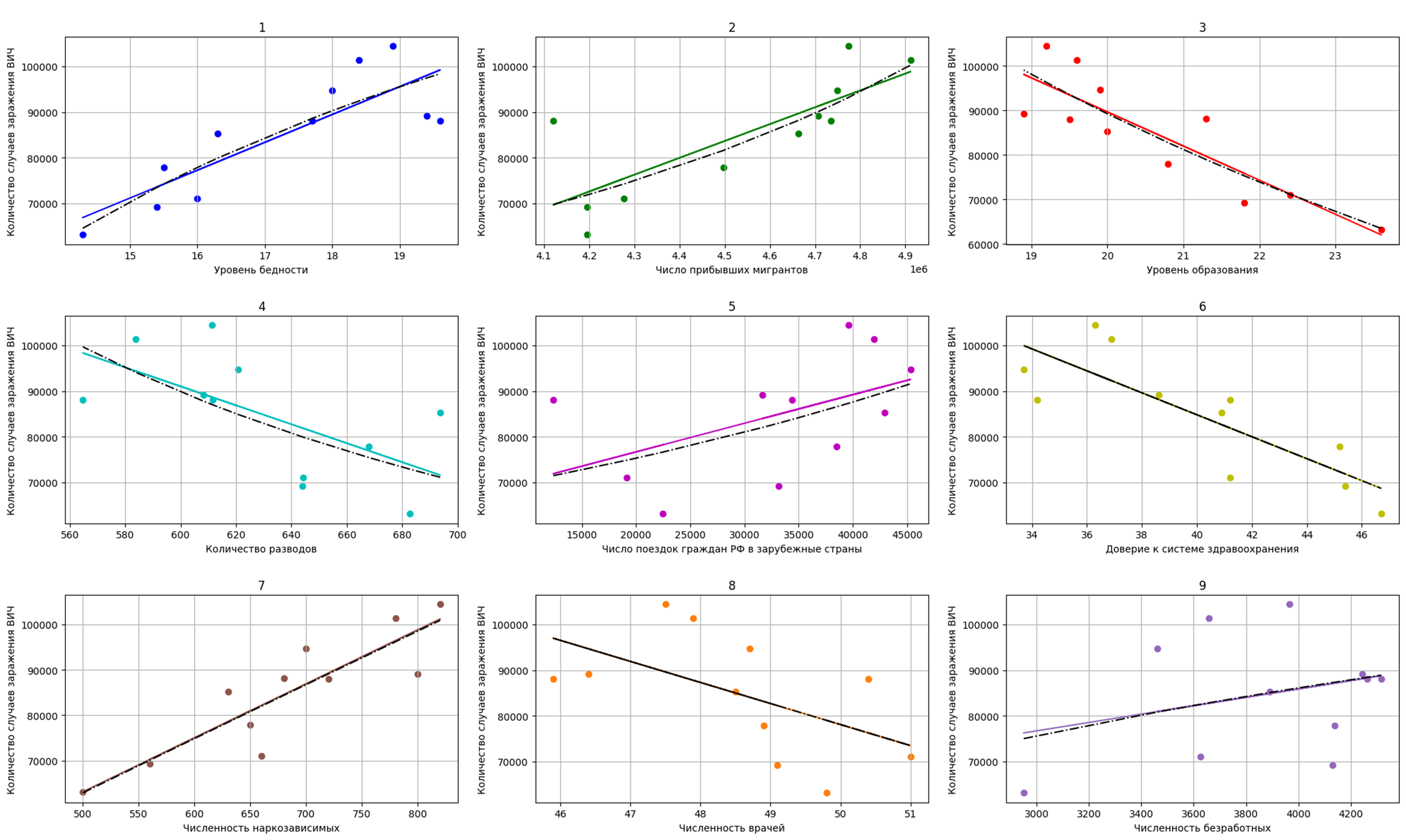


Рисунок 18 – Зависимость между числом новых случаев заражения ВИЧ и уровнем образования в РФ

Структурная идентификация для моделирования зависимости между y и x4 принимает вид: . Осуществим линеаризацию модели.

Замена переменных: ;

Линейный вид зависимости:

В соответствии с МНК:

Переход к нелинейной модели:

Коэффициент детерминации: 0.4

Таким образом, коэффициент детерминации показывает, что 40% вариации числа новых случаев заражения ВИЧ-инфекцией обусловлено вариацией количества разводов в РФ за период с 2012 по 2022 год.

На основе полученных данных построим графически линейную и нелинейную аппроксимацию на корреляционном поле (Рисунок 19).

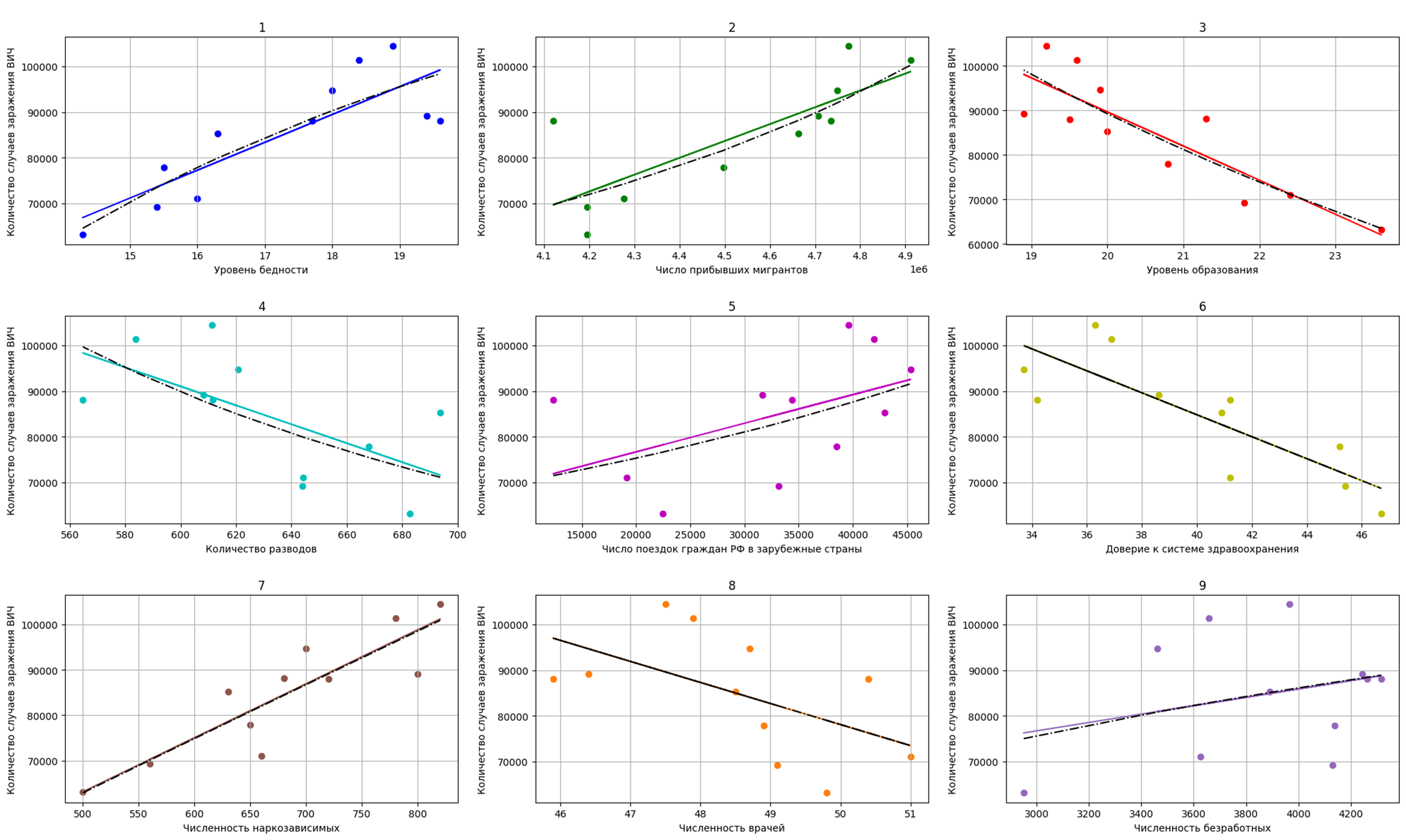


Рисунок 19 – Зависимость между числом новых случаев заражения ВИЧ и количеством разводов

Аналогичная структурная идентификация и для моделирования зависимости между y и x5: . Осуществим линеаризацию модели.

Замена переменных: ;

Линейный вид зависимости:

В соответствии с МНК:

Переход к нелинейной модели:

Коэффициент детерминации: 0.25

Таким образом, коэффициент детерминации показывает, что 25% вариации числа новых случаев заражения ВИЧ-инфекцией обусловлено вариацией числа поездок граждан РФ в зарубежные страны в РФ за период с 2012 по 2022 год.

На основе полученных данных построим графически линейную и нелинейную аппроксимацию на корреляционном поле (Рисунок 20).

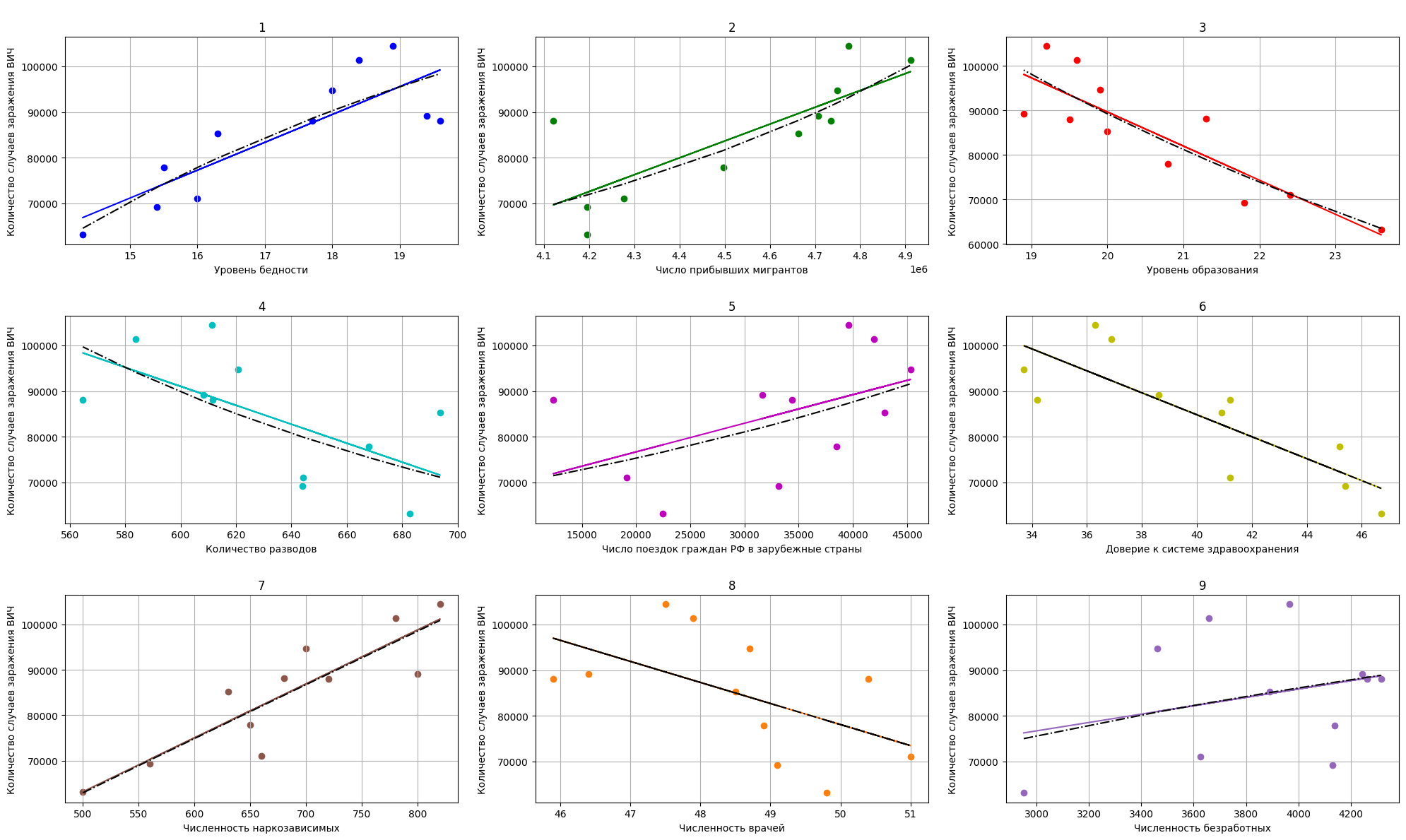


Рисунок 20 – Зависимость между числом новых случаев заражения ВИЧ и числом поездок граждан РФ в зарубежные страны

Структурная идентификация для моделирования зависимости между y и x6 принимает вид: – линейная зависимость.

Линейная модель имеет вид:

Коэффициент детерминации: 0.67.

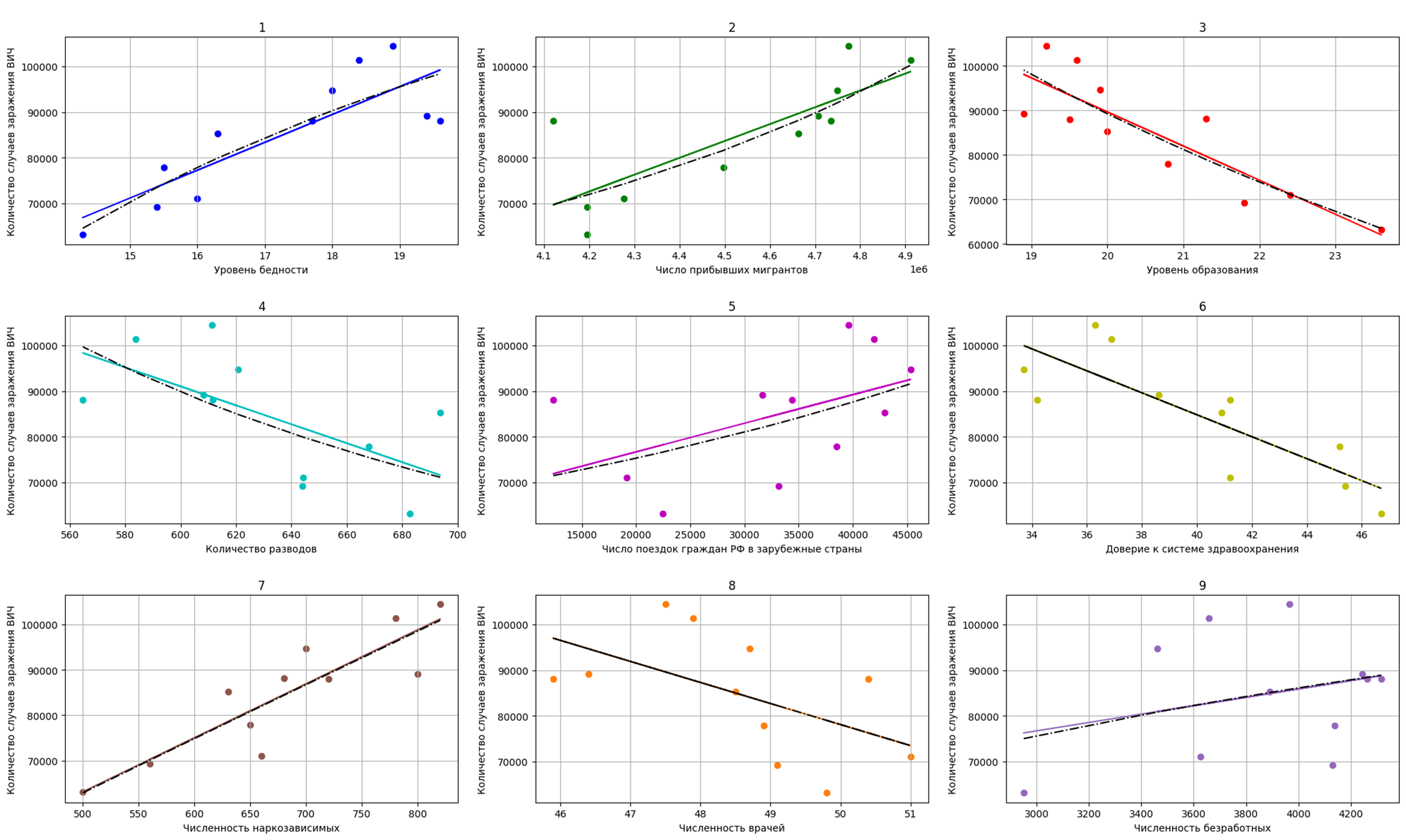


Рисунок 21 – Зависимость между числом новых случаев заражения ВИЧ и уровнем доверия к системе здравоохранения

Таким образом, коэффициент детерминации показывает, что 67% вариации числа новых случаев заражения ВИЧ-инфекцией обусловлено вариацией уровня доверия к системе здравоохранения в РФ за период с 2012 по 2022 год.

На основе полученных данных построим графически линейную аппроксимацию на корреляционном поле (Рисунок 21).

Структурная идентификация для моделирования зависимости между y и x7 принимает вид: . Осуществим линеаризацию модели.

Замена переменных: ;

Линейный вид зависимости:

В соответствии с МНК:

Переход к нелинейной модели:

Коэффициент детерминации: 0.8

Таким образом, коэффициент детерминации показывает, что 80% вариации числа новых случаев заражения ВИЧ-инфекцией обусловлено вариацией численности наркоманов, стоящих на учете в медицинских учреждениях в РФ за период с 2012 по 2022 год.

На основе полученных данных построим графически линейную и нелинейную аппроксимацию на корреляционном поле (Рисунок 22).

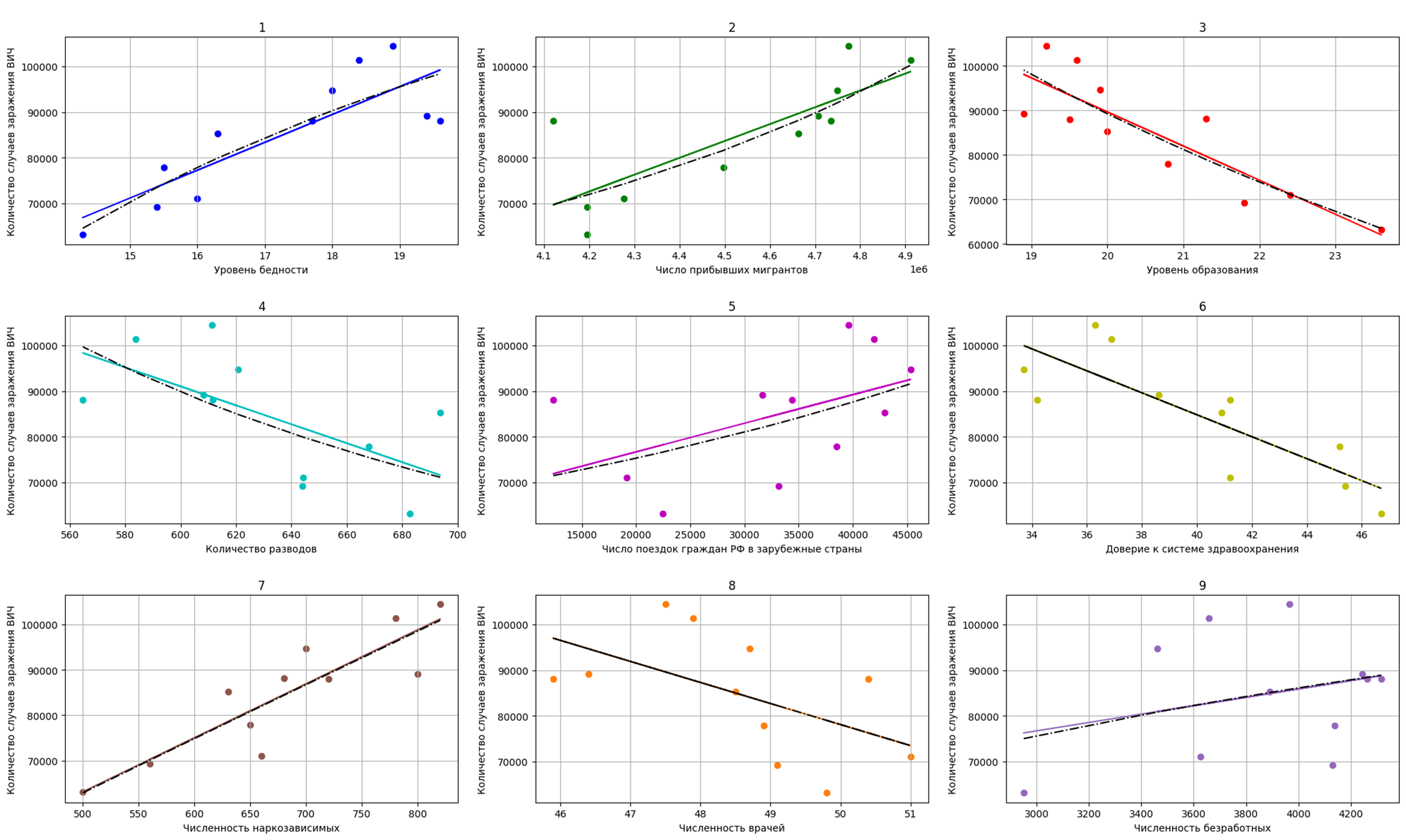


Рисунок 22 – Зависимость между числом новых случаев заражения ВИЧ и численностью наркозависимых, стоящих на учете в медицинских учреждениях

Структурная идентификация для моделирования зависимости между y и x8 принимает вид: – линейная зависимость.

Линейная модель имеет вид:

Коэффициент детерминации: 0.3.

Таким образом, коэффициент детерминации показывает, что 30% вариации числа новых случаев заражения ВИЧ-инфекцией обусловлено вариацией численности врачей всех специальностей в РФ за период с 2012 по 2022 год.

На основе полученных данных построим графически линейную аппроксимацию на корреляционном поле (Рисунок 23).

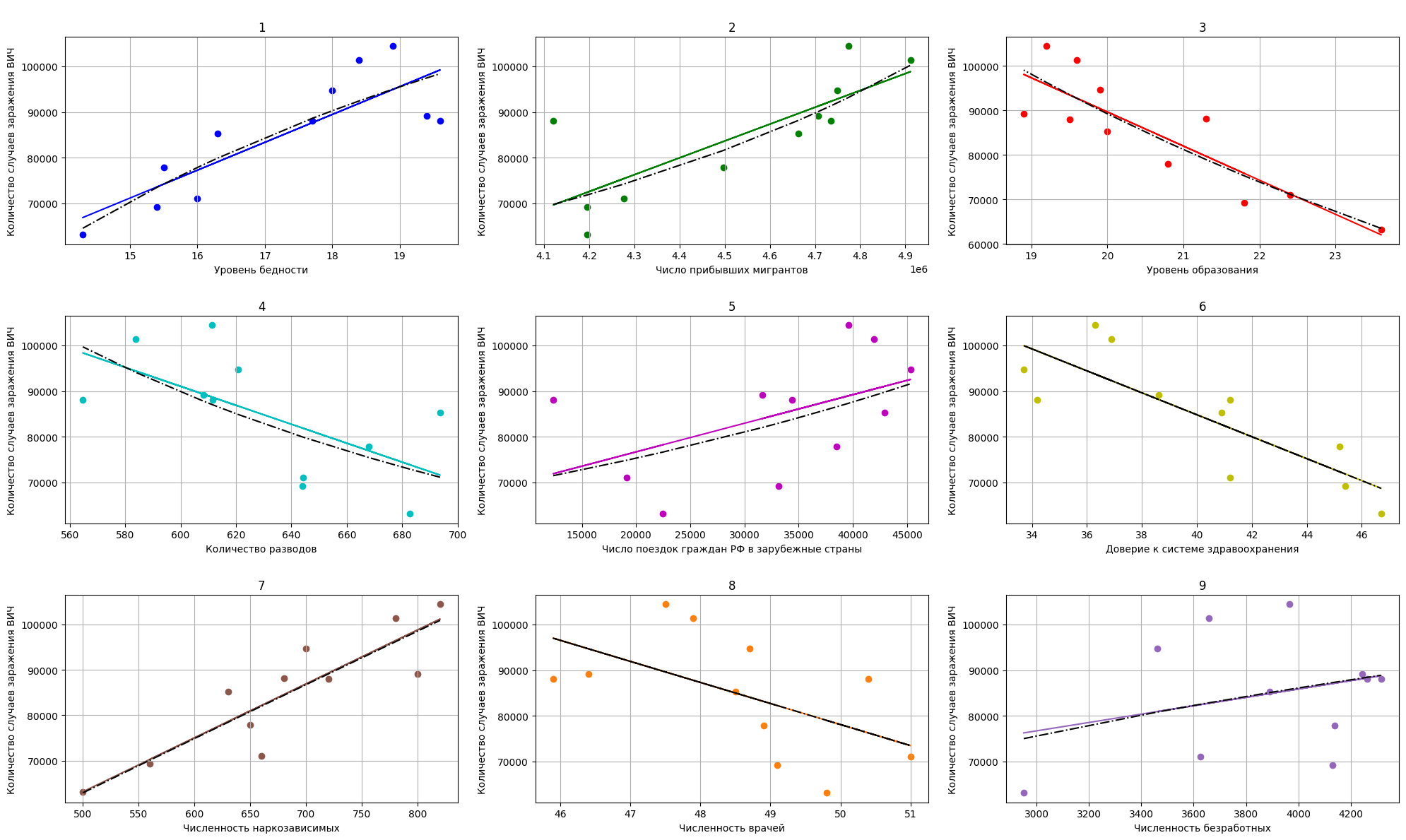


Рисунок 23 – Зависимость между числом новых случаев заражения ВИЧ и численностью врачей всех специальностей

Структурная идентификация для моделирования зависимости между y и x9 принимает вид: . Осуществим линеаризацию модели.

Замена переменных: ;

Линейный вид зависимости:

В соответствии с МНК:

Переход к нелинейной модели:

Коэффициент детерминации: 0.1

Таким образом, коэффициент детерминации показывает, что 10% вариации числа новых случаев заражения ВИЧ-инфекцией обусловлено вариацией численности безработных в РФ за период с 2012 по 2022 год.

На основе полученных данных построим графически линейную и нелинейную аппроксимацию на корреляционном поле (Рисунок 24).

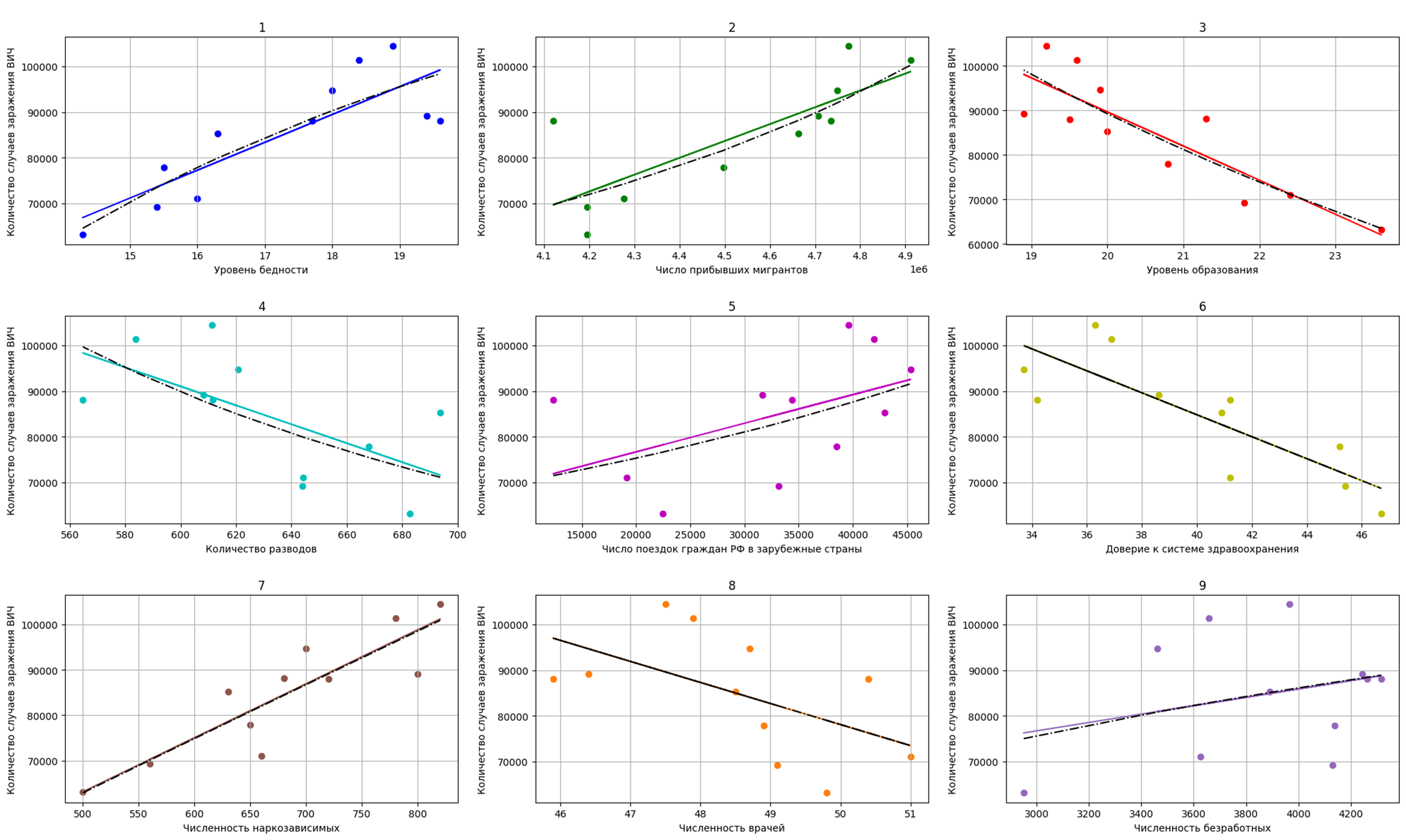


Рисунок 24 – Зависимость между числом новых случаев заражения ВИЧ и численностью безработных

1. Параметрическая идентификация модели

На основе данных за 2012-2022 года проведем многофакторный корреляционно-регрессионный анализ, чтобы выявить факторы, оказывающие наибольшее влияние на число новых случаев заражения ВИЧ-инфекцией в России.

Для начала составим матрицу парных корреляций (Таблица №10).

Таблица №10 – Корреляционная матрица влияния факторов на развитие ВИЧ-инфекции в России

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **y** | **x1** | **x2** | **x3** | **x4** | **x5** | **x6** | **x7** | **x8** | **x9** |
| **y** | 1 | 0.83 | 0.79 | -0.86 | -0.63 | 0.51 | -0.82 | 0.88 | -0.55 | 0.29 |
| **x1** | 0.83 | 1 | 0.69 | -0.86 | -0.74 | 0.26 | -0.72 | 0.9 | -0.73 | 0.47 |
| **x2** | 0.79 | 0.69 | 1 | -0.86 | -0.21 | 0.81 | -0.43 | 0.74 | -0.76 | 0.08 |
| **x3** | -0.86 | -0.86 | -0.86 | 1 | 0.43 | -0.64 | 0.58 | -0.86 | 0.81 | -0.52 |
| **x4** | -0.63 | -0.74 | -0.21 | 0.43 | 1 | 0.18 | 0.75 | -0.67 | 0.22 | -0.41 |
| **x5** | 0.51 | 0.26 | 0.81 | -0.64 | 0.18 | 1 | -0.12 | 0.33 | -0.57 | -0.02 |
| **x6** | -0.82 | -0.72 | -0.43 | 0.58 | 0.75 | -0.12 | 1 | -0.71 | 0.14 | -0.18 |
| **x7** | 0.88 | 0.9 | 0.74 | -0.86 | -0.67 | 0.33 | -0.71 | 1 | -0.6 | 0.42 |
| **x8** | -0.55 | -0.73 | -0.76 | 0.81 | 0.22 | -0.57 | 0.14 | -0.6 | 1 | -0.39 |
| **x9** | 0.29 | 0.47 | 0.08 | -0.52 | -0.41 | -0.02 | -0.18 | 0.42 | -0.39 | 1 |

По полученным данным можно сделать вывод, что среди факторов наибольшее влияние на число новых случаев инфицирования ВИЧ в России оказывает фактор x7 – численность наркозависимых, стоящих на медицинском учете.

Чтобы наиболее полно оценить степень влияния факторов на результативный признак, проведем регрессионный анализ. Построим линейную регрессионную модель, используя МНК-алгоритм.

Регрессионная модель имеет линейную структуру:

Чтобы найти параметры (коэффициенты при xi), воспользуемся методом наименьших квадратов. Метод наименьших квадратов (МНК) помогает минимизировать сумму квадратов разностей между фактическими значениями зависимой переменной и значениями, предсказанными моделью.

Таким образом, при применении МНК было получено следующее уравнение регрессии:

Коэффициент детерминации: 0.99.

Анализ полученного уравнения регрессии позволяет сделать следующие выводы о влиянии факторов на число новых случаев заражения ВИЧ:

* Среднее количество новых случаев заражения ВИЧ, приходящихся на одного мигранта, составляет 0.1 случая.
* Среднее количество новых случаев заражения ВИЧ, приходящихся на одного наркозависимого, стоящего на медицинском учете, составляет 176.6 случаев.
* Среднее количество новых случаев заражения ВИЧ, приходящихся, на одного врача на 100000 населения, составляет -24512.13. Следовательно, при увеличении численности врачей случаев заражения ВИЧ будет меньше.
* Остальные факторы также оказывают свое влияние на число новых случаев заражения ВИЧ в соответствующем уравнении регрессии.

Коэффициенты эластичности показывают, на сколько процентов в среднем изменится отклик с увеличением фактора xi на 1% от своего среднего уровня при фиксированном положении других факторов.

Для линейной модели коэффициент эластичности равен:

Определим коэффициенты эластичности для полученной модели (Таблица №11).

Таблица №11 – Коэффициенты эластичности

|  |  |
| --- | --- |
| **Фактор** |  |
| Численность населения с денежными доходами ниже границы бедности | -4.2 |
| Число прибывших в страну людей (мигрантов) | 5.3 |
| Валовой коэффициент охвата образовательными программами среднего  профессионального образования и высшего образования | 10.24 |
| Количество разводов | 0.96 |
| Число выездных туристских поездок граждан России в зарубежные страны | 0.025 |
| Уровень доверия россиян к системе здравоохранения | -4.56 |
| Численность наркозависимых, стоящих на медицинском учете | 1.42 |
| Численность врачей всех специальностей на 10000 человек населения | -14.04 |
| Численность безработных | 2.6 |

Из выше представленного анализа можно сделать вывод, что взаимодействие исследуемых показателей непосредственно и переменные имеют свойство оказывать друг на друга воздействие.

Таким образом, данные факторы могут оказывать значительное влияние на уровень распространения ВИЧ и требуют дальнейшего исследования и анализа в контексте предупреждения и контроля заболевания.

1. Исследование модели
2. Анализ статистической значимости уравнения регрессии

Анализ статистической значимости уравнения регрессии способствует выявлению влияния независимых переменных на зависимую. Статистическая значимость уравнения регрессии указывает на то, насколько уравнение лучше объясняет изменчивость зависимой переменной, чем просто среднее значение этой переменной.

Статистическая значимость уравнения регрессии определяется условием:

Оценка статистической значимости уравнения регрессии выполняется с помощью F-критерия Фишера.

F-критерий для проверки нулевой гипотезы:

Уравнение регрессии является статистически значимым, если:

1. попадает в критическую область при заданном уровне значимости α, то есть ;
2. уровень значимости , для которого является критической точкой меньше заданного уровня значимости α, то есть .

Проведем статистический анализ значимости, подсчитав, критерий Фишера. Для этого осуществим дисперсионный анализ регрессионной модели (Таблица №12).

Таблица №12 – Дисперсионный анализ регрессионной модели

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ***df*** | ***S*** |  | ***F*** | **Значимость *F*** |
| Факторная | 9 | 1729022653.54 | 192149378.28 | 13.34 | 0.2096 |
| Остаточная | 1 | 14404001.37 | 14404001.37 |  |  |
| **Общая** | 10 | 1743426654.91 |  |  |  |

На основании проведенного анализа статистической значимости уравнения регрессии можно сделать следующий вывод: F-критерий, равный 13.34, имеет соответствующую значимость 0.2096.

Обычно, в статистических тестах используется уровень значимости 0.05. В данном случае, так как значимость F больше 0.05 (0.2096>0.05), то нет основания отвергнуть нулевую гипотезу о том, что уравнение регрессии является статистически не значимым. Таким образом, на основе имеющихся данных нельзя отнести полученное уравнение регрессии к статистически значимому для объяснения изменчивости зависимой переменной.

1. Анализ статистической значимости коэффициентов уравнения регрессии

Данный анализ позволяет оценить, насколько каждый из коэффициентов уравнения является статистически значимым, а также выявить его влияние на исследуемый процесс.

Для оценки значимости коэффициентов уравнения используется t-критерий Стьюдента. В качестве критерия выбирается случайная величина Tj, распределенная по закону Стьюдента с n – m – 1 степенями свободы. В данном случае количество степеней свободы – 1.

Для проверки формулируются гипотезы: (коэффициент незначим), (коэффициент значим).

Коэффициент θj статистически значим, то есть значимо отличается от нуля (принимается гипотеза H1 на уровне значимости α), если:

* + 1. Tj попадает в критическую область при заданном уровне значимости α, то есть:
    2. уровень значимости αTj, для которого Tj является критической точкой (P-значение) меньше заданного уровня значимости α: αTj < α.

Исследуем статистическую значимость коэффициентов модели, а также найдем интервальные оценки коэффициентов регрессии с надежностью 0.95 (Таблица №13).

Таблица 13 – Статистическая значимость коэффициентов регрессии

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | *P-*Значение |  |  |
| Y-пересечение | 276982.87 | 311917.65 | 0.888 | 0.538 | -3686306.61 | 4240272.35 |
| x1 | -20678.4 | 8587.99 | -2.408 | 0.251 | -129791.35 | 88434.55 |
| x2 | 0.1 | 0.045 | 2.213 | 0.27 | -0.469 | 0.667 |
| x3 | 42051.63 | 20236.59 | 2.078 | 0.286 | -215078.6 | 299181.86 |
| x4 | 129.337 | 121.05 | 1.068 | 0.479 | -1409.41 | 1667.461 |
| x5 | 0.065 | 0.422 | 0.154 | 0.903 | -5.3 | 5.431 |
| x6 | -9652.92 | 3682.92 | -2.621 | 0.232 | -56448.79 | 37142.95 |
| x7 | 176.6 | 77.25 | 2.286 | 0.262 | -804.9 | 1158.1 |
| x8 | -24512.13 | 10479.75 | -2.339 | 0.257 | -157669.95 | 108645.69 |
| x9 | 56.85 | 26.15 | 2.174 | 0.274 | -275.41 | 389.12 |

По таблице Стьюдента (0,05; 1) = 12,7, при значимости 0,05 (1-0.95).

Для парных коэффициентов -фактическое: . Поскольку , то нет оснований отвергнуть нулевую гипотезу, а значит коэффициент уравнения регрессии является статистически незначимым.

Аналогично для остальных коэффициентов уравнения регрессии. Все они являются статистически незначимыми, так как .

1. Исследование мультиколлинеарности факторов

Мультиколлинеарность модели множественной регрессии – наличие высокой взаимной коррелированности между факторами. Такая высокая коррелированность может привести к нестабильным оценкам коэффициентов и затруднить толкование результатов.

Наличие мультиколлинеарности выявляется путем анализа:

* Матрицы парных коэффициентов корреляции факторов

Если пары переменных имеют высокие коэффициенты корреляции (> 0,7), в модели наблюдается явление мультиколлинеарности.

* Определителя матрицы парных корреляций

Если факторы не коррелированы между собой, матрица парных корреляций является единичной матрицей, и ее определитель равен 1.

Следовательно, чем ближе определитель к нулю, тем сильнее мультиколлинеарность факторов и наоборот.

Для выявления мультиколлинеарности изобразим корреляционную матрицу на тепловой карте (Рисунок 25). Чем выше будет коэффициент корреляции между факторами, тем ближе к темному оттенку на тепловой карте будет цвет их зависимости.

Таким образом, если пары переменных имеют высокие коэффициенты корреляции (> 0,7), то их нельзя будет включать в одну модель.

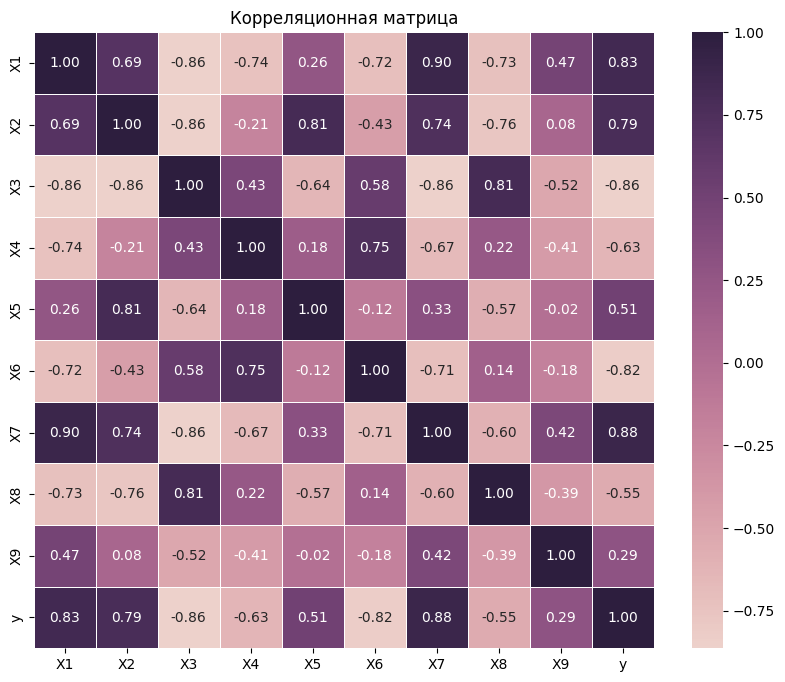


Рисунок 25 – Тепловая карта корреляционной матрицы

Если анализировать корреляционную матрицу, то можно сделать вывод, что между многими факторами, которые могут быть включены в модель, существует высокая связь, а значит модель – мультиколлиниарна.

Следовательно, необходимо провести пошаговую регрессию, чтобы в результате получилась наилучшая регрессионная модель.

1. Применения шагового регрессионного анализа для улучшения модели

Шаговый регрессионный анализ – это метод пошагового включения и исключения факторов с целью построения наилучшей модели регрессии. Основной задачей является определение наименее значимых факторов и создание оптимальной модели, улучшающей прогнозные способности. Он реализуется двумя способами: добавлением факторов или удалением факторов.

Применим шаговый регрессионный анализ для улучшения модели путем удаления факторов.

Убираем из модели переменные, имеющие наименьший коэффициент корреляции с выходной величиной – факторы x5, x8, x9. Получим:

Рассчитаем значимость коэффициентов (Таблица №14) и проведем дисперсионный анализ (Таблица №15).

Таблица №14 – Значимость коэффициентов

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  | *P-*Значение |
| Y-пересечение | 202180.89 | 148771,81 | 1.359 | 0.246 |
| x1 | -2603.81 | 2744.96 | -0.949 | 0.397 |
| x2 | 0.018 | 0.012 | 1.490 | 0.210 |
| x3 | -3226.93 | 3144.94 | -1.026 | 0.363 |
| x4 | -87.96 | 88.911 | -0.989 | 0.379 |
| x6 | -1161.07 | 573.427 | -2.025 | 0.113 |
| x7 | 18.46 | 43 | 0.429 | 0.690 |

Коэффициент детерминации: 0.95

Исправленный коэффициент детерминации: 0.86

Таблица №15 – Дисперсионный анализ

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ***df*** | ***S*** |  | ***F*** | **Значимость *F*** |
| Факторная | 6 | 1649141804.8 | 274840338.1 | 11.66 | 0.0163 |
| Остаточная | 4 | 94284850.11 | 23571212.53 |  |  |
| **Общая** | 10 | 1743426654.9 |  |  |  |

В результате сравнения фактического F-критерия (F=11,66) и табличного значения F-критерия (Fкр(0,05; 6; 4) = 6.16), можно сделать вывод, что F > Fкр, то есть нулевая гипотеза отвергается и уравнение является статистически значимым.

Теперь уберем из модели x7. Получим:

Рассчитаем значимость коэффициентов (Таблица №16) и проведем дисперсионный анализ (Таблица №17).

Таблица №16 – Значимость коэффициентов

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  | *P-*Значение |
| Y-пересечение | 221056.97 | 130033.51 | 1.7 | 0.15 |
| x1 | -2387.39 | 2468.38 | -0.967 | 0.378 |
| x2 | 0.02 | 0.011 | 1.835 | 0.126 |
| x3 | -3609.75 | 2758.93 | -1.308 | 0.248 |
| x4 | -100.22 | 77.02 | -1.301 | 0.25 |
| x6 | -1189.64 | 521.03 | -2.283 | 0.071 |

Коэффициент детерминации: 0.94

Исправленный коэффициент детерминации: 0.886

Таблица №17 – Дисперсионный анализ

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ***df*** | ***S*** |  | ***F*** | **Значимость *F*** |
| Факторная | 5 | 1644797671.18 | 329026289.8 | 16.68 | 0.0039 |
| Остаточная | 5 | 98628983.73 | 19725796.75 |  |  |
| **Общая** | 10 | 1743426654.9 |  |  |  |

Табличное значение критерия Фишера: 5.05

F > Fкр, а значит, что нулевая гипотеза отвергается и уравнение является статистически значимым.

Далее убираем параметр x1:

Рассчитаем значимость коэффициентов (Таблица №18) и проведем дисперсионный анализ (Таблица №19).

Таблица №18 – Значимость коэффициентов

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  | *P-*Значение |
| Y-пересечение | 123432.496 | 81527.4 | 1.514 | 0.181 |
| x2 | 0.018 | 0.011 | 1.7 | 0.140 |
| x3 | -2010.31 | 2196.475 | -0.915 | 0.395 |
| x4 | -48.689 | 55.318 | -0.880 | 0.413 |
| x6 | -1204.65 | 517.986 | -2.326 | 0.059 |

Коэффициент детерминации: 0.93

Исправленный коэффициент детерминации: 0.889

Таблица №19 – Дисперсионный анализ

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ***df*** | ***S*** |  | ***F*** | **Значимость *F*** |
| Факторная | 4 | 1626345016.72 | 406585502,3 | 20.836 | 0.0012 |
| Остаточная | 6 | 117081638.19 | 19513606.37 |  |  |
| **Общая** | 10 | 1743426654.9 |  |  |  |

Табличное значение критерия Фишера: 4.53

F > Fкр, а это значит, что нулевая гипотеза отвергается и уравнение является статистически значимым.

Теперь следует убрать параметр x4:

Рассчитаем значимость коэффициентов (Таблица №20) и проведем дисперсионный анализ (Таблица №21).

Таблица №20 – Значимость коэффициентов

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  | *P-*Значение |
| Y-пересечение | 131195.347 | 79705,557 | 1.646 | 0.144 |
| x2 | 0.015 | 0.01 | 1.512 | 0.174 |
| x3 | -2577.595 | 2065.716 | -1.248 | 0.252 |
| x6 | -1508.48 | 379.929 | -3.97 | 0.005 |

Коэффициент детерминации: 0.92

Исправленный коэффициент детерминации: 0.89

Таблица №21 – Дисперсионный анализ

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ***df*** | ***S*** |  | ***F*** | **Значимость *F*** |
| Факторная | 3 | 1611227889.91 | 537076368,4 | 28.4385 | 0.0003 |
| Остаточная | 7 | 132198764.999 | 18885537.86 |  |  |
| **Общая** | 10 | 1743426654.9 |  |  |  |

Табличное значение критерия Фишера: 4.34

Дальнейшее удаление факторов приводит к ухудшению модели, так что шаговый регрессионный анализ следует завершить.

Полученное уравнение является статистически значимым, так как нулевая гипотеза о статистической незначимости отвергается по критерию Фишера

По таблице Стьюдента (0,05; 6) = 2,45, при значимости 0,05. Для парных коэффициентов -фактическое: . Поскольку , то нулевая гипотеза отклоняется, а значит, что коэффициент уравнения регрессии является статистически значимым.

Однако, остальные коэффициенты уравнения регрессии статистически незначимы, потому что . Поэтому могут возникнуть сомнения относительно необходимости включения этих предикторов в модель, так как эти предикторы не имеют статистически значимого влияния на зависимую переменную.

В целом, включение статистически незначимых предикторов в регрессионную модель может привести к ухудшению ее прогностической способности и к неправильным выводам о важности факторов. Однако, если продолжать убирать параметры из модели, то исправленный коэффициент детерминации начнет уменьшаться.

1. Программная реализация и численное исследование результатов моделирования
2. Обоснование выбора и описание программного обеспечения

Для разработки модели прогнозирования распространения ВИЧ-инфекции в России был выбран язык программирования Python. Выбор данного языка обусловлен его широкой популярностью, наличием богатого набора библиотек для анализа данных и статистики, а также возможностью легкой интеграции с другими инструментами и технологиями.

Программное обеспечение на основе Python с использованием специальных библиотек предоставляет широкий набор функций для анализа данных, построения статистических моделей и визуализации результатов. Для работы с большим объемом данных были использованы библиотеки pandas и numpy, предоставляющие мощные инструменты для манипуляции и анализа данных.

Также была использована библиотека scipy для выполнения статистических расчетов, включая тесты на нормальность, корреляцию, и библиотека matplotlib с помощью которой можно было построить различные диаграммы и графики, тем самым визуализировав данные.

Для построения регрессионных моделей использовалась библиотека statsmodels. Кроме того, для выполнения отдельных задач использовался Microsoft Excel, что позволило провести дополнительный анализ данных, создать отчеты и диаграммы. Изначально все таблицы с данными были загружены в Excel, затем считаны с использованием специальных методов Python.

В целом, выбор языка программирования Python и указанных библиотек обусловлен их удобством, гибкостью и возможностью проведения разнообразных статистических и аналитических расчетов, необходимых для создания модели прогнозирования развития ВИЧ-инфекции в России, а использование Microsoft Excel дополнило возможности анализа данных.

1. Описание основных модулей программы

Для упрощения расчетов и оптимизации кода в программе были созданы специальные методы:

* Метод *find\_p()*: служит для проверки гипотезы о нормальном распределении по правилу трех сигм.
* Метод *Pearson()*: вычисляет среднее значение, стандартное отклонение, плотность вероятности нормального распределения, а также критическое значение. Используется для используется для анализа и оценки соответствия данных нормальному распределению с использованием критерия Хи-квадрат. Она позволяет проверить степень соответствия набора данных нормальному распределению в заданных границах.
* Метод *ys()*: данная функция принимает список значений xs и списки x\_list, y\_list. Затем значения x\_list и y\_list сортируются по возрастанию, после чего происходит интерполяция значений ys в зависимости от xs. Используется для определения структуры наилучшей аппроксимирующей кривой.
* Метод *Pearson\_coeff()*: принимает списки значений x и y, после чего вычисляет коэффициент корреляции Пирсона между ними. Используется для вычисления коэффициента корреляции между двумя наборами числовых данных. Коэффициент корреляции Пирсона показывает степень линейной зависимости между двумя переменными.
* Метод *coefficient()*: используется для вычисления коэффициента корреляции Пирсона, коэффициента детерминации и коэффициентов линейного уравнения на основе переданных двух наборов данных. Осуществляет анализ линейной зависимости между двумя переменными и позволяет оценить степень взаимосвязи между ними.
* Метод *model()*: создает DataFrame из переданных данных, что может быть полезно для визуализации и анализа данных. Позволяет наглядно представить эмпирическую модель или набор данных и использовать для дальнейшего анализа.

В целом код программы может быть разбит на модули относительно содержания данной работы:

* Анализ с помощью статистических показателей;
* Исследование нормального распределения;
* Корреляционный анализ;
* Структурная идентификация модели;
* Параметрическая идентификация модели;
* Анализ статистической значимости уравнения регрессии;
* Анализ статистической значимости коэффициентов уравнения регрессии;
* Исследование мультиколлинеарности факторов;
* Применение шагового регрессионного анализа.

Полный код программы представлен в приложении №1.

1. Численное исследование результатов моделирования

В результате произведенного анализа получилось, что уравнение зависимости числа новых случаев заражения ВИЧ-инфекцией в России выглядит следующим образом:

Получается, что основными факторами, от которых зависит распространение ВИЧ-инфекции в России являются:

* x2: число прибывших в страну людей (мигрантов);
* x3: валовой коэффициент охвата населения образовательными программами среднего профессионального образования и высшего образования;
* x6: уровень доверия россиян к системе здравоохранения.

Следовательно, на основе полученного уравнения можно сделать вывод о том, что на 1 мигранта приходится 0.015 новых случаев заражения ВИЧ-инфекцией в России. Поэтому с увеличением численности мигрантов в нашей стране возрастает и процент заражаемости ВИЧ.

При этом увеличение валового коэффициента охвата населения образовательными программами среднего профессионального образования и высшего образования поспособствует уменьшению числа случаев новых заражений ВИЧ-инфекцией на 2577.595 случаев. Этот фактор оказывает существенное влияние на распространение вируса таким образом, что чем выше уровень образованных людей в стране, тем меньше будет новых случаев заражений. Однако, коэффициент θ3 не является статистически значимым, поэтому влияние данного фактора на распространение ВИЧ-инфекции в России может оказаться спорным. За последние 11 лет с 2012 по 2022 год статистика по данному показателю и числу новых случаев заражения-ВИЧ показывает сильную обратную связь: пока число новых случаев заражения росло, валовой коэффициент охвата населения образовательными программами среднего профессионального образования и высшего образования падал. Следовательно, данный фактор все-таки не стоит исключать из модели прогнозирования.

Статистически значимым является коэффициент θ6, а значит уровень доверия россиян к системе здравоохранения будет оказывать наиболее существенное влияние на число новых случаев заражения ВИЧ-инфекцией. Получается, что при увеличении доверия россиян к системе здравоохранения на 1%, число новых случаев заражения ВИЧ в России уменьшается на 1508,48.

Данное уравнение является статистически значимым по критерию Фишера при уровне значимости .

Статистическая значимость уравнения по критерию Фишера при уровне значимости α=0,05 говорит о том, что данное уравнение регрессии может быть использовано для прогнозирования числа новых случаев заражения ВИЧ-инфекцией в России.

Исходя из множественного скорректированного коэффициента детерминации, равного 0,92, можно сделать вывод о том, что уравнение регрессии, описывающее зависимость числа новых случаев заражения ВИЧ-инфекцией в России от факторов x2, x3 и x6, имеет высокую предсказательную способность. Это означает, что примерно 92% изменчивости в числе новых случаев заражения ВИЧ-инфекцией в России может быть объяснено исследуемыми факторами.

Таким образом, наличие весьма высокой корреляционной связи между результативными и факторными показателями указывает на то, что уравнение регрессии хорошо объясняет изменения в числе новых случаев заражения ВИЧ-инфекцией в России и может быть использовано для прогнозирования этого показателя.

# ЗАКЛЮЧЕНИЕ

Анализ динамики изменения числа новых случаев заражений-ВИЧ инфекцией за период с 2012 по 2022 год показал, что на данный момент наблюдается тенденция к снижению числа новых случаев заражения. Несмотря на это, Россия продолжает являться страной с высоким процентом ВИЧ-инфицированных, а статус распространения вируса в России многими экспертами оценивается как эпидемия. Именно поэтому так важно выявить зависимость влияния различных факторов на число новых заражений ВИЧ для прогнозирования и предотвращения массового заражения населения в дальнейшем.

В ходе выполнения данной исследовательской работы был проведен корреляционно-регрессионный анализ, который позволил выявить степень влияния различных факторов на динамику распространения ВИЧ-инфекции в России.

На основании полученных данных была построена регрессионная модель, позволяющая прогнозировать число новых случаев заражения ВИЧ-инфекцией от изменения различных факторов, таких как численность мигрантов в стране, охват образованием населения и уровень доверия к системе здравоохранения.

Таким образом, полученные результаты могут быть использованы для разработки эффективных мер по предотвращению распространения ВИЧ-инфекции и улучшению ситуации с эпидемией в России.

# Список используемой литературы

Официальный сайт завода «Электросталь» https://elsteel.ru/

* + 1. Данные по ВИЧ-инфекции в России [электронный ресурс] // HIVrussia. URL: <http://www.hivrussia.info/dannye-po-vich-infektsii-v-rossii/> (дата обращения: 02.12.2023)
    2. Эпидемиология ВИЧ-инфекции [электронный ресурс] // Wikidata. URL: <https://www.wikidata.ru-ru.nina.az/%D0%AD%D0%BF%D0%B8%D0%B4%D0%B5%D0%BC%D0%B8%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D1%8F_%D0%92%D0%98%D0%A7-%D0%B8%D0%BD%D1%84%D0%B5%D0%BA%D1%86%D0%B8%D0%B8.html> (дата обращения: 13.12.2023)
    3. Уровень безработицы в России [электронный ресурс] // gogov.ru. URL: <https://gogov.ru/articles/unemployment-rate> (дата обращения: 16.12.2023)
    4. Уровень жизни в России [электронный ресурс] // gogov.ru. URL: <https://gogov.ru/articles/standard-of-living> (дата обращения: 17.12.2023)
    5. Численность и миграция населения Российской Федерации [электронный ресурс] // Росстат. URL: <https://rosstat.gov.ru/compendium/document/13283> (дата обращения: 14.12.2023)
    6. Образование в цифрах [электронный ресурс] // НИУ ВШЭ. URL: <https://www.hse.ru/primarydata/oc> (дата обращения: 19.12.2023)
    7. Здравоохранение [электронный ресурс] // Росстат. URL: <https://rosstat.gov.ru/folder/13721> (дата обращения: 18.12.2023)
    8. Российский Статистический Ежегодник 2021 [электронный ресурс] // Росстат. URL: <https://rosstat.gov.ru/bgd/regl/b21_13/Main.htm> (дата обращения: 16.12.2023)
    9. Российский Статистический Ежегодник 2022 [электронный ресурс] // Росстат. URL: <https://rosstat.gov.ru/storage/mediabank/Ejegodnik_2022.pdf> (дата обращения: 20.12.2023)
    10. Показатели преступности [электронный ресурс] // Имиджинфинанс. URL: <https://www.iminfin.ru/areas-of-analysis/criminality/pokazateli-prestupnosti?territory=45000000> (дата обращения: 18.12.2023)

ПРИЛОЖЕНИЕ 1

# %%

import pandas as pd

import numpy as np

from scipy import stats

import scipy.stats as stats

import math

import matplotlib.pyplot as plt

from scipy.interpolate import interp1d

from scipy.stats import kurtosis, skew, normaltest, chi2, norm, t, spearmanr, f

import colorsys

import statsmodels.api as sm

import seaborn as sns

from statsmodels.formula.api import ols

# %%

def find\_p(dict, a, k, sigma):

  fa, fb = 0, 0

  for key, value in dict.items():

      if key < (a - k\*sigma):

          fa += value

      if key < (a + k\*sigma):

          fb += value

  return ((fb-fa)\*100)

def Pearson(dict, n, h):

    sumAv, sumD, i = 0, 0, 0

    nt = []

    middle = []

    for key, value in dict.items():

        list2 = key.split('-')

        middle.append((int(list2[0]) + int(list2[1]))/2)

        sumAv += middle[i]\*value

        i += 1

    avg2 = sumAv/n

    i = 0

    for key, value in dict.items():

        sumD += (middle[i]-avg2)\*\*2 \* value

        i +=1

    sigma1 = math.sqrt(sumD/n)

    Xin, i = 0, 0

    for key, value in dict.items():

        list2 = key.split('-')

        zi = (middle[i]-avg2)/sigma1

        density = norm.pdf(zi, loc=0, scale=1)

        nt.append(h\*n\*density/sigma1)

        Xin += (value- nt[i])\*\*2/nt[i]

        i += 1

    k = i - 3

    critical\_value = chi2.ppf(1 - 0.05, k)

    return [nt, Xin, critical\_value]

def ys(xs, x\_list, y\_list):

    sorted\_values = sorted(zip(x\_list, y\_list))

    x\_list, y\_list = map(list, zip(\*sorted\_values))

    for i in range(0, len(x\_list)-1):

        if (xs > x\_list[i]) and (xs < x\_list[i+1]):

            ys\_result = y\_list[i] + (y\_list[i+1] - y\_list[i]) \* (xs - x\_list[i]) / (x\_list[i+1] - x\_list[i])

        if (xs == x\_list[i]):

            ys\_result = y\_list[x\_list.index(xs)]

    return ys\_result

def Pearson\_coeff(x, y):

    xy = []

    for i in range(0, len(x)):

        xy.append(x[i] \* y[i])

    avg\_x = np.mean(x)

    avg\_y = np.mean(y)

    avg\_xy = np.mean(xy)

    sigma\_x = np.std(x)

    sigma\_y = np.std(y)

    return (avg\_xy - avg\_x\*avg\_y)/(sigma\_x\*sigma\_y)

def coefficient(u\_, v\_):

    r1 = Pearson\_coeff(u\_, v\_)

    print(f"Коэффициент корреляции Пирсона: {r1}")

    print(f"Коэффициент детерминации: {r1\*\*2}")

    a1 = r1\*(np.std(v\_)/np.std(u\_))

    b1 = np.mean(v\_)-a1\*np.mean(u\_)

    print(f"Коэффициенты линеаризованного уравнения: a1 = {a1} и b1 = {b1}")

    return [a1, b1]

def model(nonlinear\_y, x\_list, nonlinear\_model, new\_x):

    sorted\_values = sorted(zip(nonlinear\_y, x\_list))

    nonlinear\_y, x\_list = map(list, zip(\*sorted\_values))

    df = pd.DataFrame({'Значения нелинейной зависимости y': nonlinear\_y, 'X': x\_list})

    nonlinear\_model.append(nonlinear\_y.copy())

    new\_x.append(x\_list.copy())

    nonlinear\_y.clear()

    print(df)

    print('\n')

# %%

df = pd.read\_excel('Статистика ВИЧ.xlsx')

result\_dict = df.set\_index(df.columns[0]).to\_dict()[df.columns[1]]

plt.plot(df['Год'], df['Зарегистрировано новых заболевших, чел.'], marker='o', color="navy", linewidth= 2)

plt.grid(True)

plt.xlabel('Год')

plt.ylabel('Количество новых случаев заражения ВИЧ-инфекции')

plt.title('Распространение ВИЧ инфекции в России (2012-2022)')

plt.xticks(df['Год'], df['Год'], fontsize=8)

for i, txt in enumerate(df['Зарегистрировано новых заболевших, чел.']):

    plt.text(df['Год'][i]+0.3, df['Зарегистрировано новых заболевших, чел.'][i]-750, f"{df['Зарегистрировано новых заболевших, чел.'][i]}", ha='left', va='bottom', fontsize=6)

plt.show()

# %%

hiv\_cases = list(result\_dict.values())

max\_value = max(hiv\_cases)

min\_value = min(hiv\_cases)

range\_value = max\_value - min\_value

median = np.median(hiv\_cases)

n = len(hiv\_cases)

sum1 = 0

total = 932361

#total = sum(hiv\_cases)

avg1 = total/n

for i in range (0, len(hiv\_cases)):

    sum1 += (hiv\_cases[i]-avg1)\*\*2

disp = sum1/n

print(f'Среднее:  {avg1: .2f}')

print(f'Дисперсия: {disp: .2f}')

print(f'Среднее квадратичное отклонения: {math.sqrt(disp): .2f}')

print(f'Коэффициент вариации: {math.sqrt(disp)/avg1\*100: .2f}%')

print("Медиана:", median)

print("Максимальное значение:", max\_value)

print("Минимальное значение:", min\_value)

print("Размах:", range\_value)

for key in result\_dict:

  if result\_dict[key] == max(result\_dict.values()):

    Mo = key

print('Мода:', Mo, 'Значение:', max(result\_dict.values()))

print('Общее число заболевших за период с 2012 по 2022:', total)

# %%

hiv\_coef = []

for i in range (0, len(hiv\_cases)):

    c = int(df['Зарегистрировано новых заболевших, чел.'][i])

    h = int(df['Численность населения РФ'][i])

    hiv\_coef.append(c\*100000/h)

df\_k = pd.DataFrame({'Год': df['Год'],  'Уровень заболеваемости на 100 тыс. населения' : hiv\_coef})

print(df\_k)

# %%

count = 1

sum, summt, sumc, summtt = 0, 0, 0, 0

for key, value in result\_dict.items():

    if count == 1:

        sum += value

        y = value

    else:

        sum += value

        print(f"Показатели динамики на {key} год: \n")

        print(f"Базисный абсолютный прирост: {value-y}")

        print(f"Цепной абсолютный прирост: {value-last}")

        print (f"Базисный темп роста: {value\*100/y}%")

        print (f"Цепной темп роста: {value\*100/last}%")

        print (f"Базисный темп прироста: {(value-y)\*100/y}%")

        print (f"Цепной темп прироста: {(value-last)\*100/last}% \n")

        r = value - y

        d = value/y

    last = value

    summt += (count)\*value

    summtt += (count)\*\*2

    sumc += count

    count += 1

# %%

T = pow(d, 1/(n-1))

p = sum/(n)

print(f"Средний уровень ряда: {p: .2f}")

print(f"Средний абсолютный прирост: {r/(n-1)}")

print(f"Средний темп роста: {T\*100: .2f} %")

print(f"Средний темп прироста: {T\*100-100: .2f} %")

# %%

sa = last + r/(n-1)

print(f"Прогнозирование по среднему абсолютному приросту на 2023 год: {sa}")

st = T\*last

print(f"Прогнозирование по среднему темпу роста на 2023 год: {st: .2f}")

# %%

coeff = np.polyfit(df['Год'], df['Зарегистрировано новых заболевших, чел.'], 2)

year = np.arange(2012, 2024)

val = coeff[0]\*year\*\*2 + coeff[1]\*year + coeff[2]

prog = coeff[0]\*2023\*\*2 + coeff[1]\*2023 + coeff[2]

print(coeff[2])

print(f"Прогнозирование методом аналитического выравнивания на 2023 год: {prog: .2f}")

plt.plot(df['Год'], df['Зарегистрировано новых заболевших, чел.'], marker='o', color="navy", linewidth= 2)

plt.grid(True)

plt.xlabel('Год')

plt.ylabel('Количество новых случаев заражения ВИЧ-инфекцией')

plt.title('Прогнозирование методом аналитического выравнивания')

plt.xticks(year, year, fontsize=8)

for i, txt in enumerate(df['Зарегистрировано новых заболевших, чел.']):

    plt.text(df['Год'][i]+0.3, df['Зарегистрировано новых заболевших, чел.'][i]-750, f"{df['Зарегистрировано новых заболевших, чел.'][i]}", ha='left', va='bottom', fontsize=6)

cubic\_interpolation\_model = interp1d(year, val, kind = "cubic")

X\_=np.linspace(year[0], year[len(year)-1], 500)

Y\_=cubic\_interpolation\_model(X\_)

plt.plot(X\_, Y\_)

plt.show()

# %%

years = list(result\_dict.keys())

means\_min = int(min(years))

means\_max = math.ceil(max(years))

h = 1

x = 2011

new\_interval, align = {}, {}

align[x] = None

for key, value in result\_dict.items():

  key = str(x) + '-' + str(int(x + h))

  new\_interval[key] = value/total

  x += h

  align[x] = None

df\_inter1 = pd.DataFrame({'Период': new\_interval.keys(),  'Зарегистрировано больных ВИЧ' : df['Зарегистрировано новых заболевших, чел.']})

print(df\_inter1.to\_string(index=False))

plt.bar(df['Год'], df['Зарегистрировано новых заболевших, чел.'], color='navy', width=1, edgecolor="white", linewidth=2)

plt.xticks(df['Год'], df['Год'], fontsize=8)

plt.xlabel('Год')

plt.ylim(40000, max(df['Зарегистрировано новых заболевших, чел.'])+1000)

plt.ylabel('Количество новых случаев заражения ВИЧ-инфекции')

plt.title('Распространение ВИЧ инфекции в России (2012-2022)')

plt.show()

# %%

sum3, sum4 = 0, 0

prob = {}

for key, value in result\_dict.items():

    prob[value] = result\_dict[key]/total

    sum3 += (value-avg1)\*\*3

    sum4 += (value-avg1)\*\*4

sigma = math.sqrt(disp)

As = sum3/(n\*sigma\*\*3)

Ek = sum4/(n\*sigma\*\*4)-3

print(f'Ассиметрия: {As: .3f}')

print(f'Эксцесс: {Ek: .3f}')

print(f"{find\_p(prob, avg1, 1, sigma):.2f}%")

print(f"{find\_p(prob, avg1, 2, sigma)}%")

print(f"{find\_p(prob, avg1, 3, sigma)}%")

n\_inter = {}

for key, value in new\_interval.items():

    n\_inter[key] = value\*total

results1 = Pearson(n\_inter, total, h)

df\_inter1 = pd.DataFrame({'Период': n\_inter.keys(),  'Число заболеваний' : n\_inter.values(), 'Теоретическое число заболеваний' : results1[0]})

print(df\_inter1['Теоретическое число заболеваний'].to\_string(index=False))

print(f"Наблюдаемое значение: {results1[1]}")

print(f"Критическое значение: {results1[2]}")

print("\nВывод:")

if results1[1] < results1[2]:

    print("Нет оснований отвергнуть гипотезу о нормальном распределения")

else:

    print("Гипотеза о нормальном распределения отвергается")

# %%

m, sD = 0, 0

for key, value in new\_interval.items():

    list3 = key.split('-')

    m += ((int(list3[0]) + int(list3[1]))/2)\*value

for key, value in new\_interval.items():

    list3 = key.split('-')

    sD += (((int(list3[0]) + int(list3[1]))/2)-m)\*\*2 \* value

sigma3 = math.sqrt(sD)

for key, value in align.items():

    align[key] = norm.pdf(key, loc=m, scale=sigma3)

df\_align = pd.DataFrame({'x': align.keys(),  'f(x)' : align.values()})

print(df\_align)

years\_1=list(align.keys())

y=list(align.values())

cubic\_interpolation\_model = interp1d(years\_1, y, kind = "cubic")

X\_=np.linspace(years[0], years[len(years)-1], 500)

Y\_=cubic\_interpolation\_model(X\_)

plt.bar(years, new\_interval.values(), color='navy', width=1, edgecolor="white", linewidth=2, align='center', label="Гистограмма")

plt.plot(X\_, Y\_, linewidth= 2, label="Кривая Гаусса")

plt.title('Распространение ВИЧ инфекции в России (2012-2022)')

plt.xticks(df['Год'], df['Год'], fontsize=8)

plt.xlabel('Год')

plt.ylim(0.02, max(new\_interval.values())+0.04)

plt.legend()

plt.show()

# %%

corrected\_sum = 0

for i in range(n):

    corrected\_sum += (int(hiv\_cases[i]) - avg1)\*\*2/(n-1)

corrected\_sigma = math.sqrt(corrected\_sum)

t\_crit = stats.t.ppf((1-(1-0.95)/2), n-1)

delta = corrected\_sigma\*t\_crit/math.sqrt(n)

left = avg1 - delta

right = avg1 + delta

print(f"Доверительный интервал для оценки мат. ожидания: {left} < a < {right}")

# %%

factors = pd.read\_excel('Факторы.xlsx')

data = factors.values.tolist()

x\_cols = ['X1', 'X2', 'X3', 'X4', 'X5', 'X6', 'X7', 'X8', 'X9']

x = factors[x\_cols].values.T

y = factors['y'].values

avg\_x, sigma\_x, r, rho = [], [], [], []

for i in range(0, len(x)):

    avg\_x.append(np.mean(x[i]))

    sigma\_x.append(np.std(x[i]))

avg\_y = np.mean(y)

sigma\_y = np.std(y)

for i in range(0, len(x)):

    r.append(Pearson\_coeff(x[i], y))

    print(f'Коэффициент корреляции Пирсона для параметра X{i+1}: {r[i]:.2f}')

    rho.append(spearmanr(x[i], y)[0])

    print(f'Коэффициент ранговой корреляции Спирмена для параметра X{i+1}: {rho[i]:.2f}\n')

# %%

fig, axs = plt.subplots(3, 3, figsize=(20, 10))

base\_colors = ['b', 'g', 'r', 'c', 'm', 'y', 'tab:brown', 'tab:orange', 'tab:purple']

x\_labels = ['Уровень бедности',

'Число прибывших мигрантов',

'Уровень образования',

'Количество разводов',

'Число поездок граждан РФ в зарубежные страны',

'Доверие к системе здравоохранения',

'Численность наркозависимых',

'Численность врачей',

'Численность безработных']

for i in range(9):

    row = i // 3

    col = i % 3

    axs[row, col].plot(factors['Год'], x[i], marker='o', color=base\_colors[i], linewidth= 2)

    axs[row, col].grid(True)

    if (i>5):

        axs[row, col].set\_xlabel('Год\n')

    axs[row, col].set\_ylabel(x\_labels[i])

    axs[row, col].set\_xticks(factors['Год'], factors['Год'], fontsize=8)

plt.show()

# %%

fig, axs = plt.subplots(3, 3, figsize=(20, 12))

for i in range(9):

    row = i // 3

    col = i % 3

    axs[row, col].grid(True)

    axs[row, col].scatter(x[i], y, color=base\_colors[i])

    axs[row, col].set\_xlabel(x\_cols[i])

    axs[row, col].set\_ylabel('y')

    axs[row, col].set\_xlabel(x\_labels[i])

    if (col % 3 == 0):

        axs[row, col].set\_ylabel('Количество случаев заражения ВИЧ')

    else:

        axs[row, col].set\_ylabel('')

    axs[row, col].set\_title(f'\n{i+1}')

plt.tight\_layout()

plt.show()

# %%

a, b = [], []

for i in range(0, len(x)):

    a.append(r[i]\*(sigma\_y/sigma\_x[i]))

    b.append(avg\_y-r[i]\*(sigma\_y\*avg\_x[i]/sigma\_x[i]))

    print(f"Y(X{i+1}) = {a[i]:.3f} \* X + {b[i]:.3f}")

    print(f"Коэффициент детерминации: {r[i]\*\*2:.2f}\n")

# %%

cond\_x1, cond\_x2, cond\_x3, cond\_y1, cond\_y2, cond\_y3,= [], [], [], [], [], []

for i in range(0, len(x)):

    x1 = min(x[i])

    xn = max(x[i])

    y1 = y[x[i].tolist().index(x1)]

    yn = y[x[i].tolist().index(xn)]

    cond\_x1.append((x1+xn)/2)

    cond\_x2.append(math.sqrt(x1\*xn))

    cond\_x3.append((2\*x1\*xn)/(x1+xn))

    cond\_y1.append((y1+yn)/2)

    cond\_y3.append((2\*y1\*yn)/(y1+yn))

    cond\_y2.append(math.sqrt(y1\*yn))

delta = []

for i in range(0, len(x)):

    delta = []

    delta.append(abs(cond\_y1[i] - ys(cond\_x1[i], x[i].tolist(), y.tolist())))

    delta.append(abs(cond\_y2[i] - ys(cond\_x2[i], x[i].tolist(), y.tolist())))

    delta.append(abs(cond\_y2[i] - ys(cond\_x1[i], x[i].tolist(), y.tolist())))

    delta.append(abs(cond\_y1[i] - ys(cond\_x3[i], x[i].tolist(), y.tolist())))

    delta.append(abs(cond\_y3[i] - ys(cond\_x1[i], x[i].tolist(), y.tolist())))

    delta.append(abs(cond\_y3[i] - ys(cond\_x3[i], x[i].tolist(), y.tolist())))

    delta.append(abs(cond\_y1[i] - ys(cond\_x2[i], x[i].tolist(), y.tolist())))

    print(f"Номер структуры для зависимости y от X{i+1}: {delta.index(min(delta))+1}")

    delta.clear()

# %%

nonlinear\_model, new\_x = [], []

for i in range(0, len(x)):

    print(f'Для X{i+1}:')

    if ((i == 1) or (i == 3) or (i == 4)):

        v, nonlinear\_y = [], []

        u = x[i]

        for j in range(0, len(x[i])):

            v.append(1/(y[j]))

        coeff = coefficient(u, v)

        print(f"Y(x) = {coeff[0]} \* X + {coeff[1]}")

        for j in range(0, len(x[i])):

            nonlinear\_y.append(1/(coeff[0]\*(x[i][j])+coeff[1]))

        model(nonlinear\_y, x[i], nonlinear\_model, new\_x)

    elif ((i == 0) or (i == 2)):

        u, nonlinear\_y = [], []

        v = y

        for j in range(0, len(x[i])):

            u.append(1/(x[i][j]))

        coeff = coefficient(u, v)

        print(f"Y(x) = {coeff[0]} \* X + {coeff[1]}")

        for j in range(0, len(x[i])):

            nonlinear\_y.append(coeff[1] + coeff[0]/(x[i][j]))

        model(nonlinear\_y, x[i], nonlinear\_model, new\_x)

    elif ((i == 5) or (i == 7)):

        nonlinear\_y = []

        coeff = coefficient(x[i], y.tolist())

        print(f"Y(x) = {coeff[0]} \* X + {coeff[1]}")

        for j in range(0, len(x[i])):

            nonlinear\_y.append(coeff[0]\*x[i][j] + coeff[1])

        print('Зависимость линейная')

        df = pd.DataFrame({'Значения нелинейной зависимости y': nonlinear\_y, 'X': x[i]})

        print(df)

        print('\n')

        nonlinear\_model.append(nonlinear\_y.copy())

        new\_x.append(x[i].tolist().copy())

    elif (i == 8):

        u, nonlinear\_y = [], []

        v = y

        for j in range(0, len(x[i])):

            u.append(math.log(x[i][j]))

        coeff = coefficient(u, v)

        print(f"Y(x) = {coeff[0]} \* X + {coeff[1]}")

        for j in range(0, len(x[i])):

            nonlinear\_y.append(coeff[0]\*math.log(x[i][j])+coeff[1])

        model(nonlinear\_y, x[i], nonlinear\_model, new\_x)

    elif (i == 6):

        u, v, nonlinear\_y = [], [], []

        for j in range(0, len(x[i])):

            v.append(math.log10(y[j]))

            u.append(math.log10(x[i][j]))

        coeff = coefficient(u, v)

        print(f"Y(x) = {coeff[0]} \* X + {coeff[1]}")

        for j in range(0, len(x[i])):

            nonlinear\_y.append(math.pow(10, coeff[1])\*math.pow(x[i][j], coeff[0]))

        model(nonlinear\_y, x[i], nonlinear\_model, new\_x)

# %%

fig, axs = plt.subplots(3, 3, figsize=(20, 12))

linear\_model = []

linear\_model\_d = []

for i in range (0, len(x)):

    for j in range (0, len(x[i])):

        linear\_model\_d.append(a[i]\*x[i][j] + b[i])

    linear\_model.append(linear\_model\_d.copy())

    linear\_model\_d.clear()

for i in range(9):

    row = i // 3

    col = i % 3

    axs[row, col].scatter(x[i], y, color=base\_colors[i])

    axs[row, col].set\_xlabel(x\_cols[i])

    axs[row, col].set\_ylabel('y')

    axs[row, col].grid(True)

    axs[row, col].set\_xlabel(x\_labels[i])

    axs[row, col].plot(x[i], linear\_model[i], label='Линейная аппроксимация', color=base\_colors[i])

    axs[row, col].plot(new\_x[i], nonlinear\_model[i], label='Нелинейная аппроксимация', color = 'black', linestyle='-.')

    axs[row, col].set\_ylabel('Количество случаев заражения ВИЧ')

    axs[row, col].set\_title(f'\n{i+1}')

plt.tight\_layout()

plt.show()

# %%

param = pd.read\_excel('Факторы.xlsx')

param.drop('Год', axis=1, inplace=True)

data = param.values

correlation\_matrix = param.corr(method = 'pearson')

print("Матрица парных корреляций:")

print(correlation\_matrix.to\_string())

X = data[:, :-1]

Y = data[:, -1]

X = np.insert(X, 0, 1, axis=1)

XTX = np.dot(X.T, X)

XTX\_inv = np.linalg.inv(XTX)

XTY = np.dot(X.T, Y)

beta = np.dot(XTX\_inv, XTY)

print(beta)

print(f"\nУравнение регрессии: y' = {beta[0]: .2f} +{(-1\*beta[1]): .2f}\*X1 +{beta[2]: .2f}\*X2 +{beta[3]: .2f}\*X3 +{beta[4]: .3f}\*X4 +{beta[5]: .2f}\*X5 -{(-1\*beta[6]): .2f}\*X6 +{beta[7]: .2f}\*X7 -{-1\*beta[8]: .2f}\*X8 +{beta[9]: .2f}\*X9")

# %%

y\_mean = np.mean(Y)

SST = np.sum((Y - y\_mean)\*\*2)

Y\_pred = np.dot(X, beta)

SSE = np.sum((Y - Y\_pred)\*\*2)

r2 = 1 - (SSE / SST)

print(f"Коэффициент детерминации: {r2:.2f}")

bet = beta.tolist()

e = []

for i in range (0, len(bet)-1):

    e.append(beta[i+1]\*np.mean(x[i])/np.mean(y))

elasticity = pd.DataFrame({'Параметры': x\_labels, 'Коэффициенты эластичности:': e})

print(elasticity.to\_string(index = False))

# %%

data = {'y': y, 'x1': x[0], 'x2': x[1], 'x3': x[2], 'x4': x[3], 'x5': x[4], 'x6': x[5], 'x7': x[6], 'x8': x[7], 'x9': x[8]}

df = pd.DataFrame(data)

model = ols('y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9', data=df).fit()

f\_statistic = model.fvalue

print(f'F-критерий: {f\_statistic:.4f}')

p\_value = model.f\_pvalue

print(f'Значимость F: {p\_value:.4f}')

critical\_value = f.ppf(1 - 0.05, 9, 1)

print("Табличное значение критерия Фишера:", critical\_value)

anova\_table = sm.stats.anova\_lm(model)

print("Число степеней свободы (остаточное):", anova\_table.iloc[-1, 0])

print("Факторная S:", anova\_table.iloc[:-1, 1].sum())

print("Остаточная S:", anova\_table.iloc[-1, 1])

print("Общая S:", anova\_table.iloc[:10, 1].sum())

print("Остаточная S^2:", anova\_table.iloc[-1, 2])

# %%

print(model.summary())

results\_summary = model.summary()

std\_err = results\_summary.tables[1]

std\_err\_df = pd.DataFrame(std\_err.data)

print(std\_err\_df[5])

# %%

plt.figure(figsize=(10, 8))

sns.heatmap(correlation\_matrix, cmap=sns.cubehelix\_palette(as\_cmap=True), annot=True, linewidths=.5, fmt=".2f")

plt.title('Корреляционная матрица')

plt.show()

# %%

data = {'y': y, 'x1': x[0], 'x2': x[1], 'x3': x[2], 'x4': x[3], 'x6': x[5], 'x7': x[6]}

df = pd.DataFrame(data)

X = sm.add\_constant(df[['x1', 'x2', 'x3', 'x4', 'x6', 'x7']])

Y = df['y']

model = sm.OLS(Y, X).fit()

print(model.params.to\_string())

y\_mean = np.mean(Y)

SST = np.sum((Y - y\_mean)\*\*2)

Y\_pred = np.dot(X, model.params)

SSE = np.sum((Y - Y\_pred)\*\*2)

r2 = 1 - (SSE / SST)

print(f"Коэффициент детерминации: {r2:.2f}")

print(f"Исправленный коэффициент детерминации: {1-10/(11-6-1)\*(1-r2)}")

results\_summary = model.summary()

std\_err = results\_summary.tables[1]

std\_err\_df = pd.DataFrame(std\_err.data)

print(std\_err\_df.to\_string())

# %%

data = {'y': y, 'x1': x[0], 'x2': x[1], 'x3': x[2], 'x4': x[3], 'x6': x[5], 'x7': x[6]}

df = pd.DataFrame(data)

model = ols('y ~ x1 + x2 + x3 + x4 + x6 + x7', data=df).fit()

f\_statistic = model.fvalue

print(f'F-критерий: {f\_statistic:.4f}')

p\_value = model.f\_pvalue

print(f'Значимость F: {p\_value:.4f}')

critical\_value = f.ppf(1 - 0.05, 6, 4)

print("Табличное значение критерия Фишера:", critical\_value)

anova\_table = sm.stats.anova\_lm(model)

print("Число степеней свободы (остаточное):", anova\_table.iloc[-1, 0])

print("Факторная S:", anova\_table.iloc[:-1, 1].sum())

print("Остаточная S:", anova\_table.iloc[-1, 1])

print("Общая S:", anova\_table.iloc[:10, 1].sum())

print("Остаточная S^2:", anova\_table.iloc[-1, 2])

# %%

data = {'y': y, 'x1': x[0], 'x2': x[1], 'x3': x[2], 'x4': x[3], 'x6': x[5]}

df = pd.DataFrame(data)

X = sm.add\_constant(df[['x1', 'x2', 'x3', 'x4', 'x6']])

Y = df['y']

model = sm.OLS(Y, X).fit()

print(model.params.to\_string())

y\_mean = np.mean(Y)

SST = np.sum((Y - y\_mean)\*\*2)

Y\_pred = np.dot(X, model.params)

SSE = np.sum((Y - Y\_pred)\*\*2)

r2 = 1 - (SSE / SST)

print(f"Коэффициент детерминации: {r2:.2f}")

print(f"Исправленный коэффициент детерминации: {1-10/(11-5-1)\*(1-r2)}")

results\_summary = model.summary()

std\_err = results\_summary.tables[1]

std\_err\_df = pd.DataFrame(std\_err.data)

print(std\_err\_df.to\_string())

# %%

data = {'y': y, 'x1': x[0], 'x2': x[1], 'x3': x[2], 'x4': x[3], 'x6': x[5]}

df = pd.DataFrame(data)

model = ols('y ~ x1 + x2 + x3 + x4 + x6', data=df).fit()

f\_statistic = model.fvalue

print(f'F-критерий: {f\_statistic:.4f}')

p\_value = model.f\_pvalue

print(f'Значимость F: {p\_value:.4f}')

critical\_value = f.ppf(1 - 0.05, 5, 5)

print("Табличное значение критерия Фишера:", critical\_value)

anova\_table = sm.stats.anova\_lm(model)

print("Число степеней свободы (остаточное):", anova\_table.iloc[-1, 0])

print("Факторная S:", anova\_table.iloc[:-1, 1].sum())

print("Остаточная S:", anova\_table.iloc[-1, 1])

print("Общая S:", anova\_table.iloc[:10, 1].sum())

print("Остаточная S^2:", anova\_table.iloc[-1, 2])

# %%

data = {'y': y, 'x2': x[1], 'x3': x[2], 'x4': x[3], 'x6': x[5]}

df = pd.DataFrame(data)

X = sm.add\_constant(df[['x2', 'x3', 'x4', 'x6']])

Y = df['y']

model = sm.OLS(Y, X).fit()

print(model.params.to\_string())

y\_mean = np.mean(Y)

SST = np.sum((Y - y\_mean)\*\*2)

Y\_pred = np.dot(X, model.params)

SSE = np.sum((Y - Y\_pred)\*\*2)

r2 = 1 - (SSE / SST)

print(f"Коэффициент детерминации: {r2:.2f}")

print(f"Исправленный коэффициент детерминации: {1-10/(11-4-1)\*(1-r2)}")

results\_summary = model.summary()

std\_err = results\_summary.tables[1]

std\_err\_df = pd.DataFrame(std\_err.data)

print(std\_err\_df.to\_string())

# %%

data = {'y': y, 'x2': x[1], 'x3': x[2], 'x4': x[3], 'x6': x[5]}

df = pd.DataFrame(data)

model = ols('y ~ x2 + x3 + x4 + x6', data=df).fit()

f\_statistic = model.fvalue

print(f'F-критерий: {f\_statistic:.4f}')

p\_value = model.f\_pvalue

print(f'Значимость F: {p\_value:.4f}')

critical\_value = f.ppf(1 - 0.05, 4, 6)

print("Табличное значение критерия Фишера:", critical\_value)

anova\_table = sm.stats.anova\_lm(model)

print("Число степеней свободы (остаточное):", anova\_table.iloc[-1, 0])

print("Факторная S:", anova\_table.iloc[:-1, 1].sum())

print("Остаточная S:", anova\_table.iloc[-1, 1])

print("Общая S:", anova\_table.iloc[:10, 1].sum())

print("Остаточная S^2:", anova\_table.iloc[-1, 2])

# %%

data = {'y': y, 'x2': x[1], 'x3': x[2], 'x6': x[5]}

df = pd.DataFrame(data)

X = sm.add\_constant(df[['x2', 'x3', 'x6']])

Y = df['y']

model = sm.OLS(Y, X).fit()

print(model.params.to\_string())

y\_mean = np.mean(Y)

SST = np.sum((Y - y\_mean)\*\*2)

Y\_pred = np.dot(X, model.params)

SSE = np.sum((Y - Y\_pred)\*\*2)

r2 = 1 - (SSE / SST)

print(f"Коэффициент детерминации: {r2:.2f}")

print(f"Исправленный коэффициент детерминации: {1-10/(11-3-1)\*(1-r2)}")

results\_summary = model.summary()

std\_err = results\_summary.tables[1]

std\_err\_df = pd.DataFrame(std\_err.data)

print(std\_err\_df.to\_string())

# %%

data = {'y': y, 'x2': x[1], 'x3': x[2], 'x6': x[5]}

df = pd.DataFrame(data)

model = ols('y ~ x2 + x3 + x6', data=df).fit()

f\_statistic = model.fvalue

print(f'F-критерий: {f\_statistic:.4f}')

p\_value = model.f\_pvalue

print(f'Значимость F: {p\_value:.4f}')

critical\_value = f.ppf(1 - 0.05, 3, 7)

print("Табличное значение критерия Фишера:", critical\_value)

anova\_table = sm.stats.anova\_lm(model)

print("Число степеней свободы (остаточное):", anova\_table.iloc[-1, 0])

print("Факторная S:", anova\_table.iloc[:-1, 1].sum())

print("Остаточная S:", anova\_table.iloc[-1, 1])

print("Общая S:", anova\_table.iloc[:10, 1].sum())

print("Остаточная S^2:", anova\_table.iloc[-1, 2])