Санкт–Петербургский государственный университет

Мотренко Светлана Андреевна

Выпускная квалификационная работа

Статистический анализ влияния социально-экономических показателей на рождаемость в России

Уровень образования: бакалавриат
Направление 01.03.02 «Прикладная математика и информатика»
Основная образовательная программа СВ.5005.2017
«Прикладная математика, фундаментальная информатика и программирование»

Научный руководитель:
старший преподаватель, кафедра математической теории игр и статистических решений, кандидат физ.-мат. наук
Кумачева Сурия Шакировна

Рецензент:
доцент, кафедра математического моделирования энергетических систем, кандидат физ.-мат. наук
Балыкина Юлия Ефимовна

Санкт-Петербург
2021 г.

Содержание

Введение

Вот уже несколько десятилетий демографы и экономисты работают над решением проблемы рождаемости в России, желая за счет этого повысить естественный прирост и избежать сокращения населения. Данная проблема актуальна в большинстве стран, так как для каждого государства важно его будущее развитие и процветание, а рождаемость в данном случае является одним из ключевых факторов, будучи источником новой экономической и военной мощи страны, обеспечивающим развитость государства, а также его безопасность.

Для определения приоритетных направлений и формирования эффективной демографической политики государства, поддерживающей устойчивое развитие всей экономической системы, необходимо принять во внимание и исследовать ряд факторов и условий. Для того чтобы получить объективную картину функционирования социально-экономических процессов, оценивать влияние факторов необходимо не только по отдельным составляющим, но и учитывая их совместное взаимодействие. В зависимости от соотношения значимости показателей, от взаимообусловленности их воздействия можно прогнозировать различные результаты подобного влияния [1].

Демографическое развитие, в основном, характеризуют такие показатели, как воспроизводство населения, уровень рождаемости и смертности, а также миграционный прирост [2]. Для России характерна достаточно сложная демографическая ситуация, обусловленная низкой рождаемостью и высокой смертностью, которая наблюдается в любой возрастной группе. Из этого следует, что во внимание также стоит принять оценки состояния здоровья и общий уровень здравоохранения.

Несмотря на совершенствование современных технологий во всех сферах деятельности, основным общественно-необходимым элементом экономических отношений является человек. Таким образом, уровень развития государства во всех сферах напрямую взаимосвязан с людьми. Это значит, что воспроизводство населения является важной составляющей для государства и его дальнейшего развития в целом [3].

Актуальность

Актуальность проблем рождаемости как социально-демографического процесса обусловлена обострением противоречий между его динамикой и объективной потребностью общества в воспроизводстве населения, особенностями протекания демографических и социальных процессов на современном этапе развития общества, который характеризуется глубоким демографическим кризисом.

Необходимо подчеркнуть: проблема рождаемости продолжает беспокоить правительство РФ, о чем свидетельствует введение в 2007 году программы материнского капитала и ее дальнейшие изменения, приуроченные к повышению рождаемости в стране.

На рисунке 1 представлен график изменения суммарного коэффициента рождаемости с 1995 года. На нем хорошо видно, что с введением новой политики в 2007 году коэффициент начал расти, но через несколько лет резко начал падать.

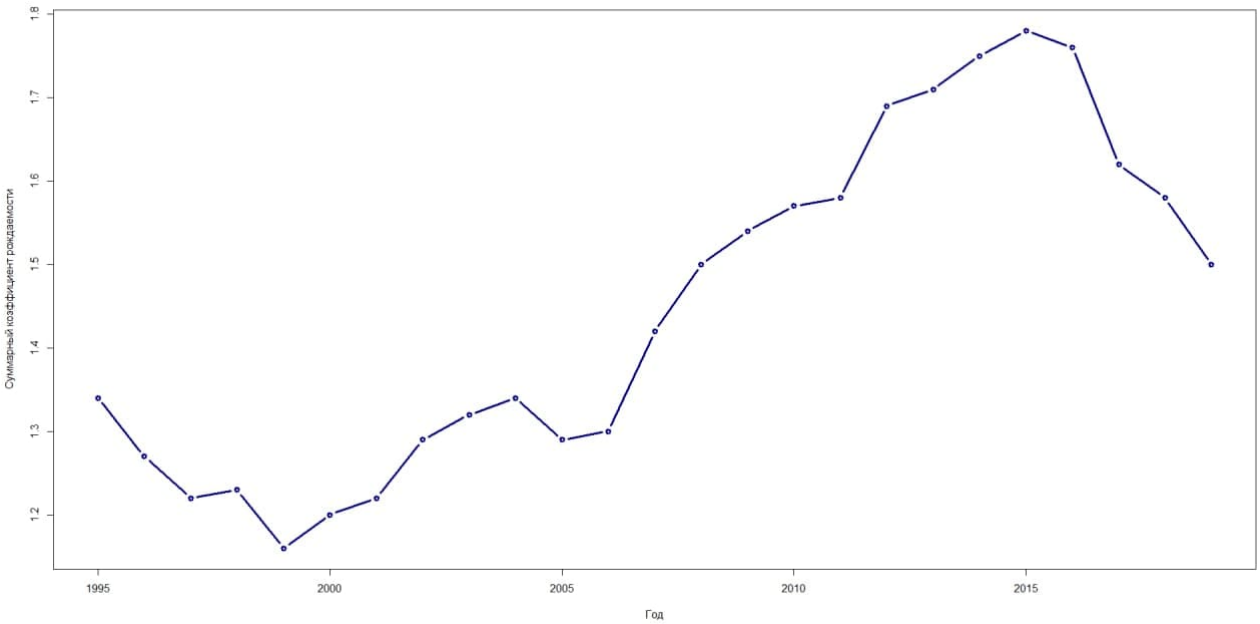

Рис. 1: Суммарный коэффициент рождаемости в России по годам

В 2020 году суммарный коэффициент рождаемости в РФ показал минимальные значения за последние 13 лет и составил 1,49 ребенка на одну женщину. Такая стабилизация наступила после сокращения суммарного

коэффициента рождаемости на 16% при 1,5 в 2019 г. При этом уровень, на котором закрепился суммарный коэффициент рождаемости в последние два года, является самым низким с 2009 г., т. е. почти за весь период действия программы выплат федерального материнского капитала (начались в 2007 г.)[4].

Проблема низкой рождаемости не нова как в науке, так и на практике, но при этом в каждом конкретном обществе она отличается своей специфичностью, которая углубляется на уровне региональной дифференциации. Отсюда вытекает необходимость изучения локальных особенностей развития населения, как на уровне Российской Федерации, так и на уровне ее региональных единиц.

Для понимания полной картины изменения рождаемости в стране необходимо узнать, от каких факторов может зависеть коэффициент рождаемости. Именно поэтому целью данной работы является выявление социально-экономических факторов, влияющих на уровень рождаемости по регионам РФ, а также построение модели множественной линейной регрессии для получения прогноза дальнейшей динамики рождаемости на основе отобранных факторов.

Обзор литературы

Авторы сборника «Население России 2007» [22] исследуют изменения показателей рождаемости в 2007 году в связи с введением новой государственной политики в сфере здравоохранения (политика материнского капитала), а также анализируют взаимосвязь коэффициентов рождаемости с изменением других социально-демографических факторов.

Лучшее понимание эффекта государственной политики дают работы [23, 24], которые были выпущены 2013 и 2009 годах соответственно. В них было изучено влияние политики материнского капитала на рождаемость в России. Одним из значимых результатов в данных статьях является факт того, что эффект от политики, предложенной государством, кратковременный и неэффективный. Если государство нацелено на глобальное решение проблемы, то следует учитывать влияние социально-экономических факторов, которые, в свою очередь, имеют тесную связь с медико-демографическими показателями. Об этом в работе [25] рассказывают авторы и утверждают, что неблагоприятные тенденции, характерные для демографических процессов в РФ, во многом связаны с низкой эффективностью мер по социально-экономическому развитию.

Влияние различных социально-экономических факторов на рождаемость на основе множественной линейной регрессии многократно исследовалось ранее. Среди работ стоит отметить статьи [3, 7, 8], в которых проблематика связана именно с демографией России в последние годы. В основном в статьях была использована модель множественной линейной регрессии. Именно эти работы и легли в основу проводимого исследования. Основными отличиями являются расширение спектра факторов, которые, возможно, имеют влияние на рождаемость, а также формирование выборки независимых наблюдений по различным данным для регионов России.

Благодаря работам [11, 26], были изучены различные показатели рождаемости и выбрана зависимая переменная.

Инструментарий регрессионного и факторного анализа был изучен по учебникам и пособиям [13], [14], [15], [16], [17], [18], [19], [20], [21].

Все расчеты и вычисления были реализованы с помощью языка про-

граммирования для статистической обработки данных — R, который был изучен при помощи [12], [16] и [29].

С помощью изученных на основе указанной литературы методов был проведен факторный анализ влияния на показатели рождаемости, построены регрессионные модели, проведен их сравнительный анализ и сделаны выводы. Математическим результатам была дана интерпретация на популяризированном общедоступном языке. Также была дана оценка эффективности построенных моделей и разбор вспомогательных характеристик.

Постановка задачи

Чтобы решить задачу, поставленную выше, требуется выполнить несколько пунктов:

- выявить и отобрать социально-экономические факторы, имеющие возможное влияние на рождаемость;
- сформировать выборку по регионам РФ за 2018 и 2019 годы;
- выполнить предварительную обработку данных;
- провести множественный регрессионный анализ на данных 2018 года;
- оценить качество полученных моделей;
- провести факторный анализ;
- проверить качество построенных моделей на данных 2019 года.

Глава 1. Формирование и обработка данных

1.1 Формирование выборки

Для реализации поставленных в работе целей были выдвинуты основанные на анализе изученных работ [3], [6], [7], [8], [9], [11] предположения о составе факторов, влияющих на показатели рождаемости:

- S (Salary) — среднемесячная номинальная начисленная заработная плата работников организаций (в рублях);
- B (Birthrate) — число родившихся на 1000 человек населения;
- P (Population) — численность населения (в тысячах человек);
- LE (Life Expectancy) — ожидаемая продолжительность жизни при рождении (в годах);
- TFR (Total Fertility Rate) — суммарный коэффициент рождаемости (в количестве детей на одну женщину);
- MR (Migration Rate) — коэффициент миграционного прироста на 10000 человек населения;
- TMR (Total Marriage rate) — общий коэффициент брачности на 1000 человек;
- HPI (Healthcare Performance Index) — индекс эффективности систем здравоохранения в регионе;
- GRP (Gross Regional Product) — валовый региональный продукт на душу населения;
- ER (Employment Rate of population) — уровень занятости населения в процентах;
- MD (Marriages, Divorces) — количество разводов на 1000 браков;
- QLI (Quality of Life Index) — индекс качества жизни;

- MF (Male to Female ratio) — соотношение мужчин и женщин (количество женщин на 1000 мужчин);
- UP (Urban Population) — удельный вес городского населения в общей численности населения (в процентах).

Поиск производился по общедоступным статистическим базам. В основном все факторы были взяты с сайта федеральной службы государственной статистики — Росстат [5], за исключением QLI([27]) и HPI([28]). На основе данных факторов была составлена выборка по 86 регионам Российской Федерации за 2018 год. Кроме того, была составлена аналогичная выборка за 2019 год, чтобы провести сравнение прогнозируемого и реального суммарного коэффициента рождаемости (далее — СКР).

1.2 Обработка данных

При обработке данных было обнаружено большое количество выбросов. Для наглядного представления обратимся к графикам «ящик с усами», которые представлены в приложении дипломной работы (Рисунок 18). На них хорошо видно, что аномальные значения присутствуют почти у каждого фактора. При работе с выбросами важно понимать природу их происхождения. Зачастую такие резкие отклонения возникают в результате случайного просчета, неправильного чтения показаний измерительного прибора, случайного сдвига запятой в десятичной записи числа и т.д. В нашем случае они отражают природу совокупности и могут иметь большую значимость, поэтому было принято решение не убирать наблюдения с аномальными значениями [17].

Так как уровни в Москве и Санкт-Петербурге сильно отличаются от других регионов России, то было принято решение сделать две выборки и сравнить полученные модели. Первая выборка включала в себя Москву и Санкт-Петербург, вторая — не включала.

Глава 2. Регрессионный анализ

В настоящей главе будет описана методология построения моделей и представлен регрессионный анализ данных, проведенный на их основе.

2.1 Выбор зависимого фактора

Один из самых важных моментов при построении модели множественной регрессии — это выбор зависимого фактора. Так как наша задача состоит в том, чтобы узнать, что влияет на рождаемость, зависимым фактором, очевидно, могут выступать:

1. Суммарный коэффициент рождаемости;
2. Общий коэффициент рождаемости (далее — ОКР);
3. Абсолютное число родившихся за год.

Все коэффициенты описывают схожие понятия, но имеют немного разное значение и интерпретацию. Чтобы лучше понять различия, дадим определения вышеперечисленным показателям:

- Суммарный коэффициент рождаемости показывает то число детей, которые были бы рождены в среднем одной женщиной на протяжении всей ее жизни при условии сохранения уровня рождаемости во всех возрастах неизменным и именно таким, каков он на момент расчета коэффициента. Рассчитывается как сумма возрастных коэффициентов рождаемости, умноженная на 5 (если возрастные коэффициенты по 5-летним группам; если они по однолетним группам, то умножения не делается) и деленная на 1000, так как возрастные коэффициенты рождаемости рассчитываются на 1000 женщин, а суммарный коэффициент — на одну.
- Общий коэффициент рождаемости представляет собой число родившихся в расчете на 1000 человек населения. Рассчитывается путем деления абсолютного числа родившихся на среднегодовую общую численность населения и умножения полученного результата на 1000.

- Абсолютное число родившихся представляет собой общее число детей, родившихся живыми.

Общий коэффициент рождаемости больше подходит в качестве показателя, чем абсолютное число родившихся, потому что в нем учитывается количество людей, проживающих на территории. Однако, этот показатель для серьезного анализа рождаемости не совсем удачен. Стоит заметить, что реально в процессе деторождения принимает участие не все население, а только женщины фертильного возраста. Поэтому, чем выше будет доля этих женщин в общей численности населения, тем, при прочих равных условиях, будет выше и общий коэффициент рождаемости [11]. Поэтому более предпочтительным является суммарный коэффициент рождаемости, который в текущем исследовании и рассматривается в качестве зависимой переменной.

В Главе 3.3 также были построены модели на основе общего коффициента рождаемости, чтобы проверить наличие отличий в значимых факторах при влиянии на различные коэффициенты рождаемости.

2.2 Методология построения моделей

В процедуре построения множественной регрессионной модели правильный отбор факторов весьма важен. Подходы к отбору факторов на основе показателей корреляции могут быть разными. Есть два наиболее применимых метода, которые дают хорошие результаты:

- метод исключения — на первом шаге строится уравнение регрессии с полным набором факторов, а затем, после исключения коллинеарных факторов, отбираются регрессоры, имеющие наибольшее влияние на изменение результативного признака; менее значимые факторы при этом исключаются.

- метод включения — заключается в поэтапном введении новых факторов в регрессионную модель.

Поскольку в нашем случае мы рассматриваем всего 12 факторов, влияние которых будем проверять, то можем воспользоваться методом исключения и сначала включить все указанные факторы, а далее убрать незначимые.

Модель множественной линейной регрессии описывается уравнением:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_m x_{im} + \epsilon, \tag{2.1}$$

где коэффициенты β_j, $j = 1, 2, \ldots, m$ — неизвестные параметры, характеризующие среднее изменение результата с изменением фактора x_j на единицу при неизменном значении других факторов. x_{ij} — значения объясняющих факторов, ϵ_i — ненаблюдаемая случайная компонента, j — номер переменной, i — номер наблюдения [16].

Модель наблюдений (2.1) можно записать в матричном виде:

$$Y = X\beta_0 + \epsilon,$$

где

$$\begin{aligned} Y &= (y_1, y_2, ..., y_n)^T, \\ \beta &= (\beta_0, \beta_1, ..., \beta_k)^T, \\ \epsilon &= (\epsilon_1, \epsilon_2, ..., \epsilon_n)^T, \\ X &= (X_0, X_1, ..., X_k), \end{aligned}$$

где

$$\begin{aligned} X_0 &= (1, 1, ..., 1)^T, \\ X_r &= (x_{1r}, x_{2r}, ..., x_{nr})^T \end{aligned}$$

$r = 1, ..., m$, вектора, содержащие все наблюдения признака r.

2.3 Метод наименьших квадратов

Одним из наиболее распространенных методов оценки неизвестных параметров $\beta_0, \beta_1, \beta_2, ..., \beta_m$ в линейной модели (2.1) является метод наименьших квадратов (МНК) [13, 16]. Задача в том, чтобы найти такие оценки неизвестных параметров β, чтобы вектор $X\beta$ как можно лучше объяснял значения наблюдаемого показателя, т.е. вектор Y. Следовательно, надо выбирать такое β, чтобы минимизировать расстояние между векторами или квадрат длины отклонения $Y - X\beta$.

В таком случае будем минимизировать функцию:

$$(Y - X\beta)^T(Y - X\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_k x_{ik})^2$$

Оценки метода наименьших квадратов $\tilde{\beta}$ неизвестных параметров β имеют вид:

$$\tilde{\beta} = (X^T X)^{-1} X^T Y$$

Тогда уравнение линейной регрессии имеет следующий вид:

$$\tilde{Y} = X\tilde{\beta} = X(X^T X)^{-1} X^T Y,$$

а вектор остатков рассчитывается по формуле:

$$\tilde{\epsilon} = Y - \tilde{Y}.$$

Чтобы оценки были качественныминужно проверить условие теоремы Гаусса-Маркова:

1. $E\epsilon_i = 0, i = 1, ..., n$ (математическое ожидание последовательности случайных величин равно нулю),

2. $E(\epsilon_i \epsilon_j) = 0, i \neq j$ (отсутствие автокорреляции),

3. $E\epsilon_i^2 = \sigma^2, i = 1, ..., n$ (постоянство дисперсий последовательности случайных величин — гомоскедастичность),

тогда оценки метода наименьших квадратов $\tilde{\beta}$ являются наилучшими линейными несмещенными оценками [16].

Еще одной предпосылкой МНК является условие о нормальности распределения случайных величин: если оно выполняется, то мы можем проводить оценку параметров регрессии, используя статистики Стьюдента и Фишера. Стоит отметить, что оценки регрессии, найденные с применением МНК, обладают хорошими свойствами даже при отсутствии нормального распределения остатков[13].

2.4 Верификация модели

Чтобы оценить качество модели множественной регрессии, требуется выполнение нескольких постулатов.

1. Особо важное значение имеет проверка гипотез статистической значимости найденных оценок $\tilde{\beta}_j, j = 0, 1, 2..., k$. $H_0 : \tilde{\beta}_j = 0$: если данная гипотеза принимается, то мы не имеем права считать, что наш коэффициент значим на уровне значимости α.

2. В качестве эффективности уравнения регрессии наиболее часто используют коэффициент детерминации R^2 или квадрат коэффициента корреляции между наблюдаемыми значения показателя $Y = (y_1, ..., y_n)^T$ и значениями эмпирической функции $\tilde{Y} = (\tilde{y}_1, ..., \tilde{y}_n)^T$, т.е.

$$R^2 = \frac{(\sum_{i=1}^{n}(y_i - \bar{y})(\tilde{y}_i - \bar{y}))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2}$$

   R^2 фиксирует долю объясненной вариации результативного признака за счет рассматриваемых в регрессии k факторов. R^2 изменяется от 0 до 1, следовательно, если $R^2 = 1$, то линейная регрессия идеально точно соответствует наблюдениям, а чем ближе к 0, тем больше полученные значения функции $\tilde{Y}$ отличаются от значений истинного Y [13, 16].

3. Стандартное нормальное распределение остатков. Как уже было сказано выше, чтобы к проверке оценок параметров можно было применить статистики Стьюдента и Фишера, необходимо, чтобы остатки имели нормальное распределение.

4. Чрезвычайно важным является требование относительно матрицы исследуемых факторов. Она должна быть свободна от мультиколлинеарности. Наличие мультиколлинеарности может исказить правильную экономическую интерпретацию параметров регрессии[13].

5. Проверка выполнимости предпосылок МНК по теореме Гаусса-Маркова.

2.5 Проверка параметров регрессии

Последовательно проверим выполнение всех перечисленных постулатов регрессионного анализа для исследуемой модели.

1. Качество уравнения и оценка тесноты связи факторов с признаком. Чтобы оценить общее качество уравнения регрессии, наиболее часто используют коэффициент детерминации R^2. Эта величина при умножении на 100% показывает, на сколько процентов изменения результативного признака объясняются изменением факторных признаков, включенных в модель. R^2 рассчитывается как квадрат индекса множественной корреляции, который, в свою очередь, является оценкой тесноты связи факторов с исследуемым признаком.

2. Значимость коэффициентов регрессии.
   Для проверки данной гипотезы будем смотреть на результат работы функции lm из пакета $stats$. Применение данной функции выводит таблицу остатков, а точнее, разности между данными и предсказанными значениями Y. Также дает нам оценки параметров уравнения линейной регрессии $\tilde{\beta}_j, j = 1, ..., k$, их стандартные ошибки $\tilde{\beta}_j / t_{\beta_j}, j = 1, ..., k$, значения статистик $t_{\beta_j}, j = 1, ..., k$ для проверки гипотез $H_0 : \beta_j = 0, j = 1, ..., k$. В столбце $Pr(> |t|)$ записаны соответствующие значения p-value, по которым можно сделать вывод

о значимости или незначимости коэффициента. Если p-value больше уровня значимости, то нулевая гипотеза принимается и делается вывод о незначимости коэффициента. В противном случае, нулевая гипотеза отвергается и коэффициент признается значимым.

3. Нормальность распределения остатков.
   Нормальность распределения $H_0 : \epsilon_i \sim N(0, 1)$ посмотрим на гистограмме частот, а также определим с помощью теста Шапиро-Уилка (shapiro.test). Будем использовать именно этим критерием, так как его мощность выше мощности непараметрических критериев согласия типа Колмогорова, Крамера-Мизеса-Смирнова и других. Уровень значимости возьмем равный $\alpha = 0.01$.

4. Математическое ожидание остатков равно нулю.
   Данную предпосылку проверяем с помощью критерия Стьюдента (t.test), сформулируем гипотезу: $H_0 : E\epsilon_i = 0, i = 1, ..., n$, при уровне значимости $\alpha = 0.01$.

5. Мультиколлинеарность.
   Для определения мультиколлинеарности будем использовать коэффициент вздутия дисперсии (vif — рассчитывает коэффициент для каждого фактора). Он позволяет оценить увеличение дисперсии заданного коэффициента регрессии, происходящее из-за высокой корреляции данных.
   $$VIF_j = \frac{1}{1 - R_j^2},$$
   где R_j^2 — коэффициент детерминации j-го признака относительно остальных. Пороговое значение возьмем VIF = 10, если VIF $>$ 10, то такой результат свидетельствует о мультиколлинеарности относительно остальных признаков нашего набора.

6. Гомоскедастичность:
   Для проверки на гомоскедастичность $H_0 : E\epsilon_i^2 = \sigma^2, i = 1, ..., n$ воспользуемся двумя тестами для более точной проверки. Будем применять тест Бройша-Пагана (bptest) и Голдфельда-Квандта (gqtest),

так как они учитывают количество наблюдений в выборке и поэтому являются точным. Уровень значимости α возьмем равный 0.01.

7. Автокорреляция:
   Автокорреляцию $H_0 : E(\epsilon_i \epsilon_j) = 0, i \neq j$ (отсутствие автокорреляции) между остатками текущих и предыдущих наблюдений проверяем тестом Бройша-Годфри (bgtest). Преимущество теста Бройша—Годфри по сравнению с тестом Дарбина—Уотсона заключается в том, что он проверяется с помощью статистического критерия, между тем как тест Дарбина—Уотсона содержит зону неопределенности для значений статистики d. Уровень значимости возьмем равный $\alpha = 0.01$.

Глава 3. Построение и верификация моделей множественной регрессии

3.1 Построение моделей по первой выборке

Модель 1.1:

$$\begin{aligned} TFR = 7.326 - 0.000006632 * S + 0.0009635 * MR \\ -0.004533 * TMR - 0.005606 * HPI \\ +0.00003356 * P - 0.02113 * LE \\ -0.001906 * UP - 0.002085 * MF \\ -0.001626 * ER - 0.0013 * MD \\ +0.0000001537 * GRP - 0.009092 * QLI \end{aligned}$$

Построенная модель включала в себя все предложенные факторы и строилась на выборке, куда входили все регионы, включая Москву и Санкт-Петербург.

1. Коэффициент детерминации: $R^2 = 0.7997$.

2. Значимость коэффициентов.
   В данной модели, следуя описанной вышеметодике, были признаны значимыми коэффициенты при факторах:
   S — заработная плата
   MR — миграционный прирост
   HPI— индекс качества здравоохранения
   P — количество населения
   MF — соотношение количества женщин и мужчин
   MD — соотношение браков и разводов
   GRP — валовый региональный продукт
   QLI — индекс качества жизни
   β_0 — свободный член

   Так как мы используем метод исключения, то для построения модели

№1.2 все факторы с незначимыми коэффициентами будут исключены.

3. Нормальность распределения остатков.

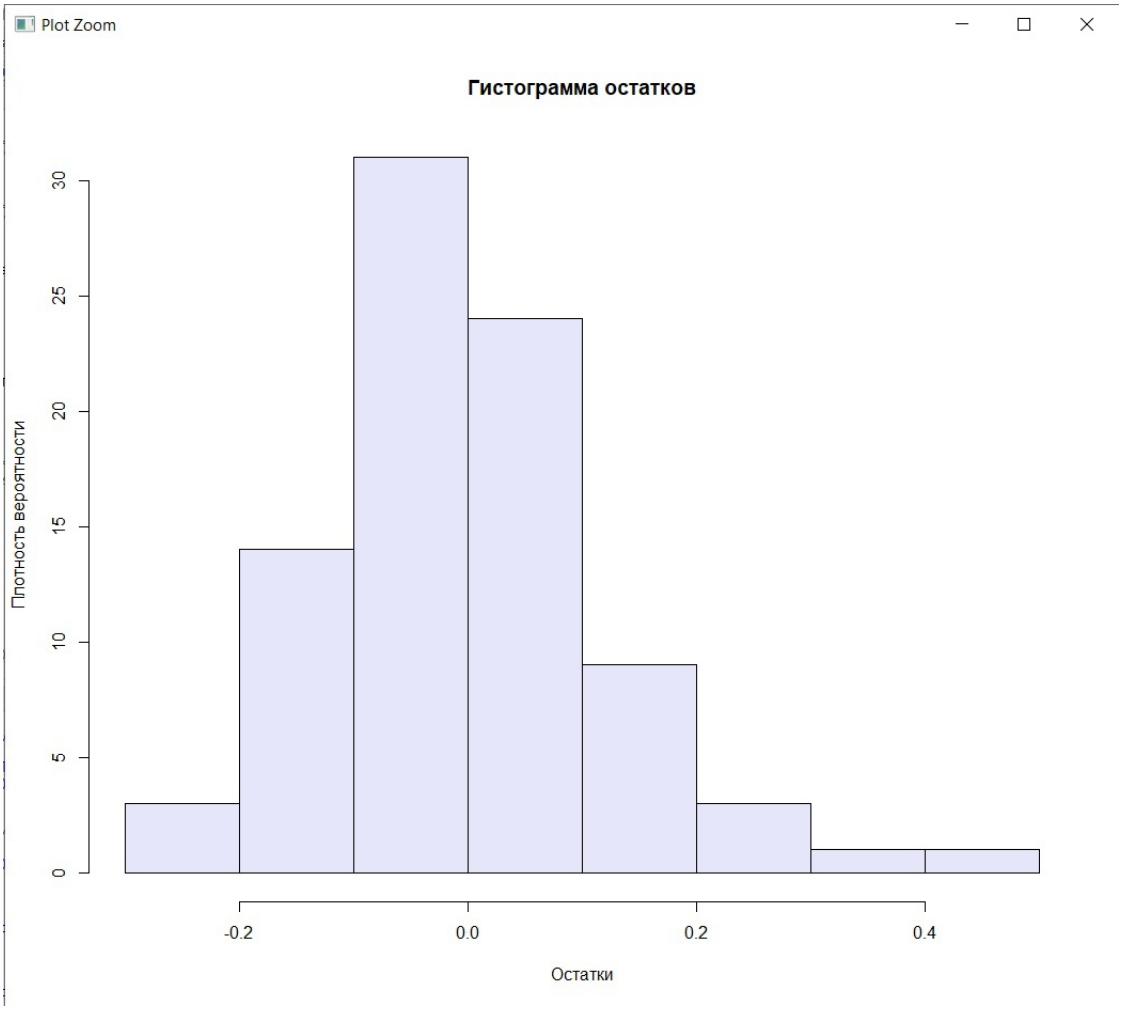

Рис. 2: Гистограмма остатков модели № 1.1

Тест Шапиро-Уилка: $W = 0.96733, p - value = 0.02822$.
На уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, т.е. остатки имеют нормальное распределение.

4. Равенство нулю математического ожидания.
Тест Стьюдента: $t = 1.0515e - 16, p - value = 1$.
На уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, т.е. мат-ожидание случайной переменной равно нулю.

5. Гомоскедастичность.

Тест Бройша-Пагана: $BP = 17.141, p - value = 0.1444$.
Тест Голдфелда-Квандта: $GQ = 2.0553, p - value = 0.02642$.
На уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, т.е. дисперсия случайной переменной постоянна.

6. Автокорреляция.
Тест Бройша-Годфри: $LM = 3.388, p - value = 0.06567$.
На уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, т.е. остатки не зависят друг от друга.

7. Мультиколлинеарность.
Коэффициент вздутия дисперсии VIF по факторам:
S — 10.219569
MR — 1.575589
TMR — 2.929324
HPI — 4.620030
P — 2.327428
LE — 3.826782
UP — 3.032979
MF — 3.732823
ER — 5.251388
MD — 1.797803
GRP — 3.006433
QLI — 5.271164

У признака S коэффициент вздутия дисперсии больше 10, значит, признак S мультиколлинеарен с другими признаками, что может искажать оценки модели, поэтому уберем его из модели № 1.2.

Прогноз по данным 2019 года.

На Рисунке 3 представлены результаты предсказания СКР по модели №1.1. Коэффициент корреляции между реальными и предсказанными значениями = 0.87, что по шкале Чеддока[14] означает сильную взаимосвязь.

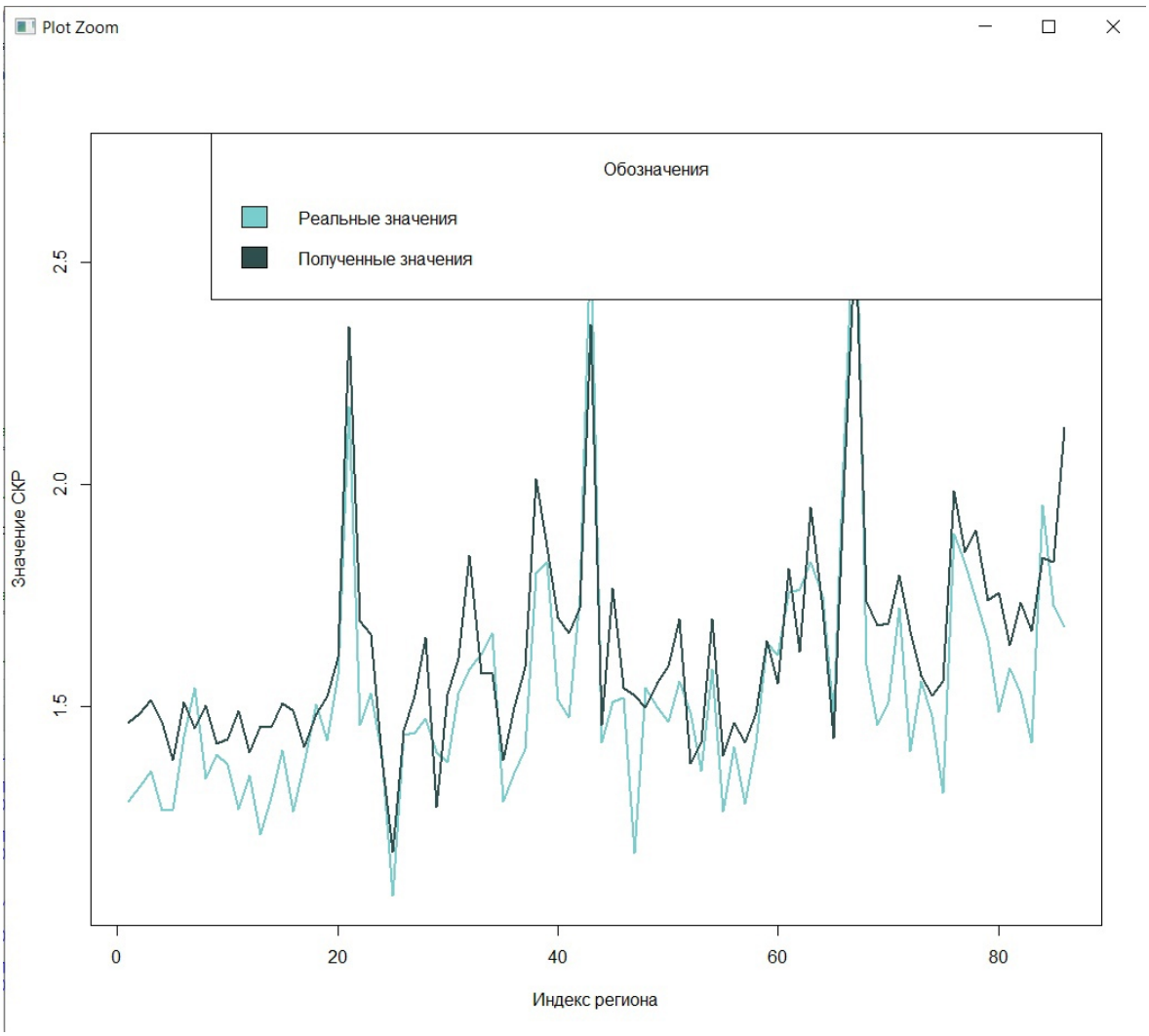

Рис. 3: Предсказание СКР по модели 1.1 на 2019 год и сравнение с реальными результатами

Модель 1.2: Построим модель на основе тех же данных, что и предыдущую, но исключим факторы, которые по результатам модели 1.1 являлись незначимыми.

$$\begin{aligned} TFR = 4.714 - 0.004878 * HPI + 0.0000189 * P \\ -0.001381 * MF - 0.001236 * MD \\ +0.00000008569 * GRP - 0.01122 * QLI \end{aligned}$$

Остается справедливым все, сказанное про предыдущую модель.

1. Коэффициент детерминации: $R^2 = 0.7349$.

2. Значимость коэффициентов.

В данной модели на уровне значимости $\alpha = 0.05$ все коэффициенты были признаны значимыми, а именно:
HPI— индекс качества здравоохранения
P — количество населения
MF — соотношение количества женщин и мужчин
MD — соотношение браков и разводов
GRP — валовый региональный продукт
QLI — индекс качества жизни
β_0 — свободный член

3. Нормальность распределения остатков.

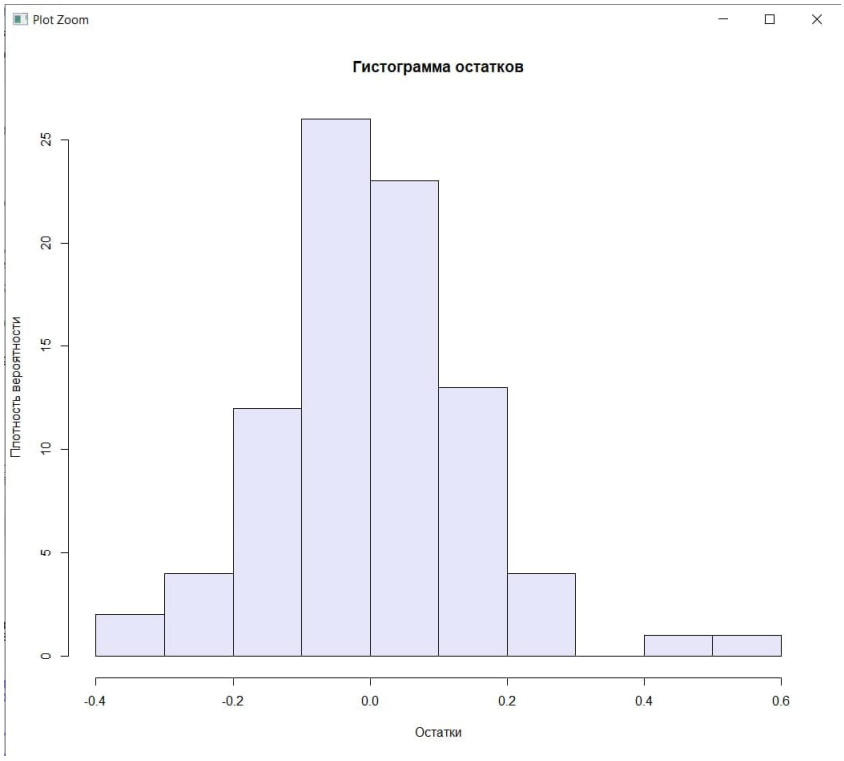

Рис. 4: Гистограмма остатков модели № 1.2

Тест Шапиро-Уилка: $W = 0.96524, p - value = 0.02046$
На уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, т.е. остатки имеют нормальное распределение.

4. Равенство нулю математического ожидания.
Тест Стьюдента: $t = 2.2363e - 16, p - value = 1$.

На уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, т.е. математическое ожидание случайной переменной равно нулю.

5. Гомоскедастичность.
   Тест Бройша-Пагана: $BP = 26.667, p - value = 0.0001671$.
   Тест Голдфелда-Квандта: $GQ = 2.4051, p - value = 0.005017$.
   На уровне значимости $\alpha = 0.01$ гипотеза H_0 отвергается, т.е. присутствует гетероскедастичность.

6. Автокорреляция.
   Тест Бройша-Годфри: $LMtest = 7.0409, p - value = 0.007967$.
   На уровне значимости $\alpha = 0.01$ гипотеза H_0 отвергается, т.е. присутствует автокорреляция остатков.

7. Мультиколлинеарность.
   Коэффициент вздутия дисперсии VIF по факторам:
   HPI — 1.272614
   P — 1.932466
   MF — 1.520990
   MD — 1.446972
   GRP — 1.512514
   QLI — 2.011908

   У всех признаков коэффициент вздутия дисперсии меньше 10, значит мультиколлинеарность между признаками отсутствует.

В данной модели присутствует гетероскедастичность и автокорреляция, значит, оценки, полученные по методу наименьших квадратов, могут быть смещенными.

Прогноз по данным 2019 года.

Коэффициент корреляции между реальными и предсказанными значениями = 0.86, что по шкале Чеддока означает сильную взаимосвязь. Но из-за наличия гетероскедастичности и автокорреляции данную модель стоит скорректировать.

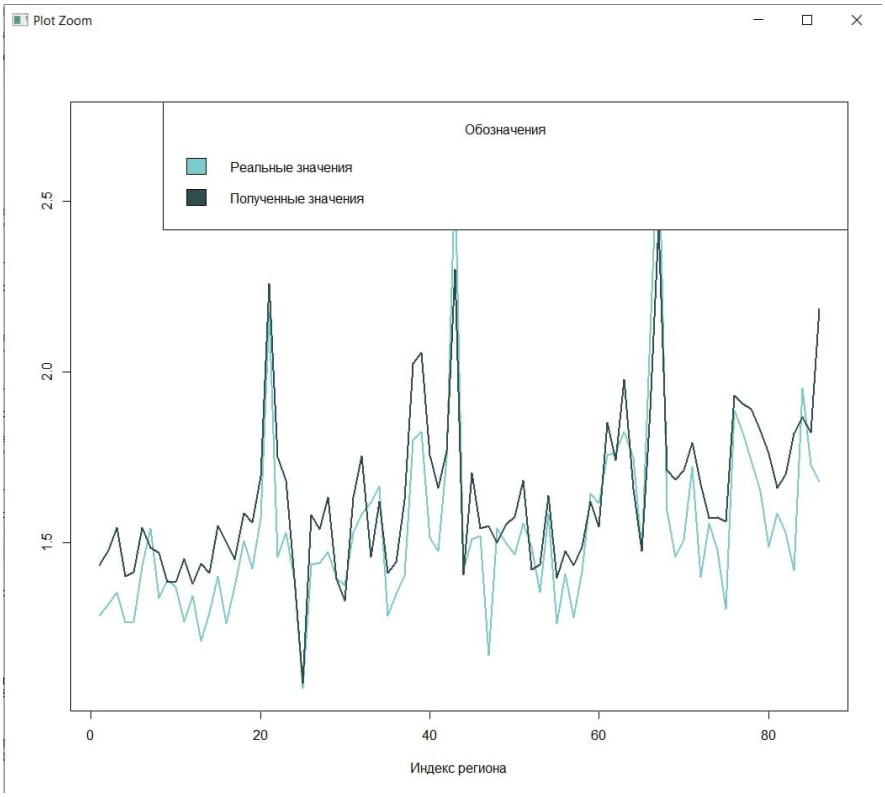

Рис. 5: Предсказание СКР по модели 1.2 на 2019 год и сравнение с реальными результатами

Модель 1.3: Отбор значимых факторов проводился с помощью дисперсионного однофакторного анализа. Чтобы определить влияние факторов, показания по каждому из них делим на несколько групп. Выдвигаем гипотезу факторного анализа $H_0 : a_1 = ... = a_k, i = 1...k$ где i — количество срезов, а a_i — среднее значение зависимого фактора в группе i. Нашу нулевую гипотезу можно переформулировать так: фактор не оказывает никакого влияния на зависимый показатель — суммарный коэффициент рождаемости. Альтернативная гипотеза утверждает, что различия в разных группах существенны. Чтобы принять или отклонить нулевую гипотезу — о равенстве k средних, будем применять классический метод дисперсионного анализа — критерий Фишера. Выполним проверку гипотезы с помощью функции anova, уровень значимости возьмем $\alpha = 0.01$. Также для проведения факторного анализа требуется нормальность зависимой переменной, поэтому применяем логарифмическую трансформацию.

После проведения однофакторного дисперсионного анализа относительно каждого фактора были отобраны следующие показатели:

HPI — индекс качества здравоохранения

P — количество населения

LE — ожидаемая продолжительность жизни

UP — удельный вес городского населения

MF — соотношение количества женщин и мужчин

MD — соотношение браков и разводов

GRP — валовый региональный продукт

QLI — индекс качества жизни

Далее построена модель множественной регрессии на основе этих факторов:

$$\begin{aligned} TFR = 6.306 - 0.003627 * HPI + 0.00002389 * P \\ -0.0252 * LE - 0.004845 * UP \\ -0.001076 * MF - 0.001267 * MD \\ +0.0000001095 * GRP - 0.00786 * QLI \end{aligned}$$

Для данной модели:

1. Коэффициент детерминации: $R^2 = 0.7751$.

2. Значимость коэффициентов.
   В данной модели на уровне значимости $\alpha = 0.05$ все коэффициенты были признаны значимыми.

3. Нормальность распределения остатков.

   Тест Шапиро-Уилка: $W = 0.98128, p - value = 0.2473$.
   На уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, т.е. остатки имеют нормальное распределение.

4. Равенство нулю математического ожидания.
   Тест Стьюдента: $t = 4.8001e - 17, p - value = 1$.
   На уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, т.е. мат-ожидание случайной переменной равно нулю.

5. Гомоскедастичность.
   Тест Бройша-Пагана: $BP = 22.691, p - value = 0.003785$.

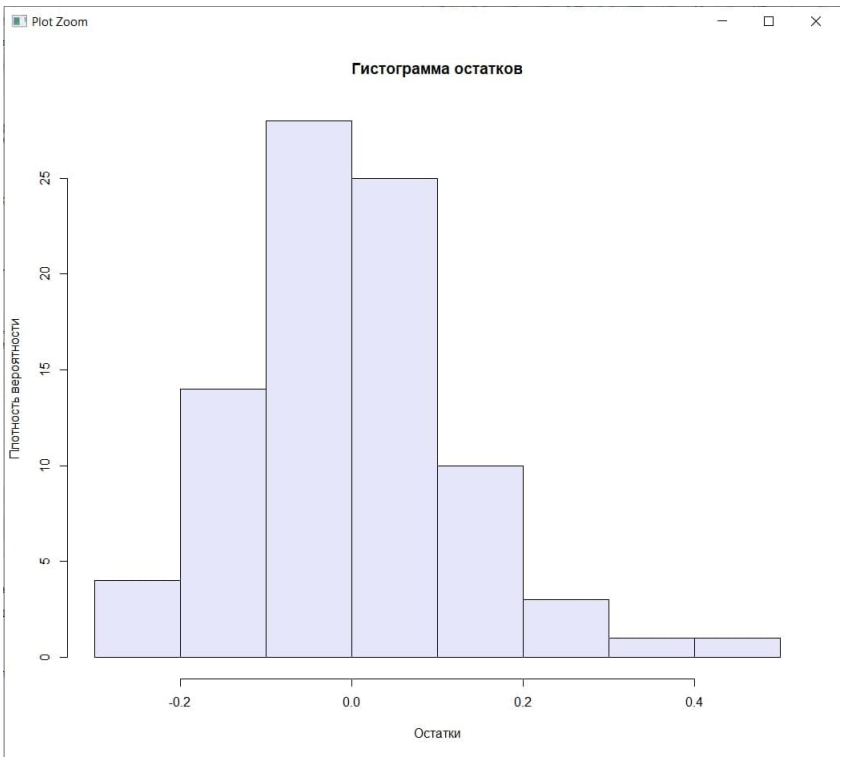

Рис. 6: Гистограмма остатков модели № 1.3

Тест Голдфелда-Квандта: $GQ = 2.4648, p - value = 0.005103$.
На уровне значимости $\alpha = 0.01$ гипотеза H_0 отвергается, т.е. присутствует гетероскедастичность.

6. Автокорреляция.
   Тест Бройша-Годфри: $LMtest = 2.7044, p - value = 0.1001$.
   На уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, т.е. отсутствует автокорреляция остатков.

7. Мультиколлинеарность.
   Коэффициент вздутия дисперсии VIF по факторам:
   HPI — 3.279363
   P — 1.963464
   LE — 3.567226
   UP — 1.917999
   MF — 1.604916
   MD — 1.582706
   GRP — 1.719025

QLI — 2.456071

У всех признаков коэффициент вздутия дисперсии меньше 10, значит мультиколлинеарность между признаками отсутствует.

Прогноз по данным 2019 года

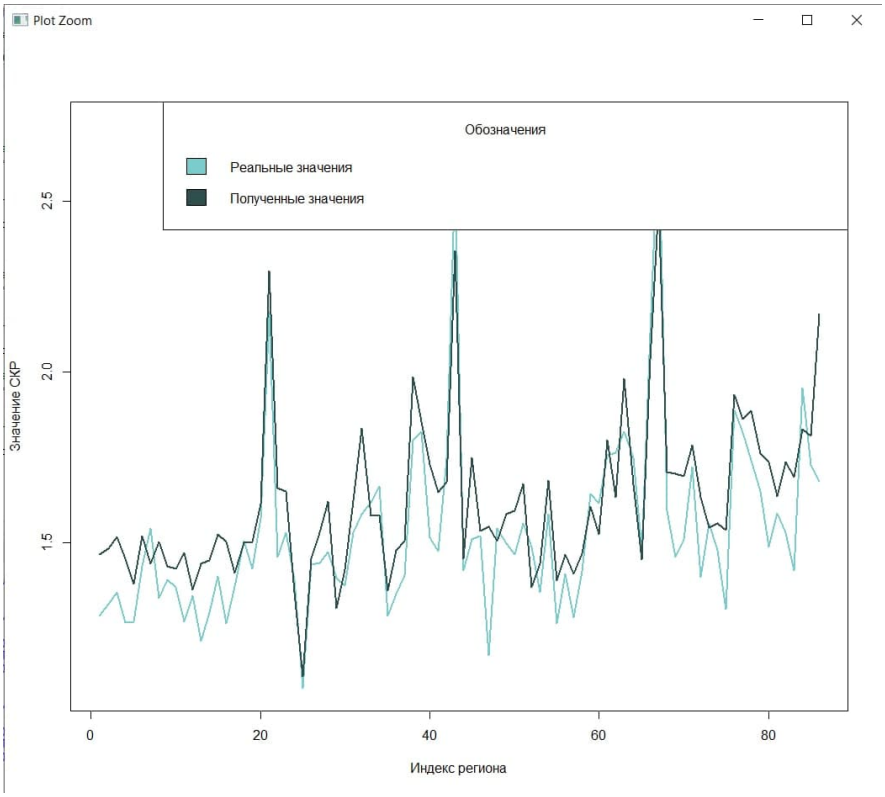

Рис. 7: Предсказание СКР по модели 1.3 на 2019 год и сравнение с реальными результатами

Коэффициент корреляции между реальными и предсказанными значениями = 0.88, что по шкале Чеддока означает сильную взаимосвязь.

3.2 Построение моделей по второй выборке

Модель 2.1: Построим модель на основе всех имеющихся факторов, но из данных уберем Москву и Санкт-Петербург.

$$\begin{aligned} TFR = 8.162 - 0.000008336 * S + 0.0004084 * MR \\ -0.02777 * TMR - 0.00638 * HPI \\ +0.00003398 * P - 0.02423 * LE \\ -0.001618 * UP - 0.002475 * MF \\ +0.001626 * ER - 0.001298 * MD \\ +0.0000001616 * GRP - 0.008941 * QLI \end{aligned}$$

Для данной модели:

1. Коэффициент детерминации: $R^2 = 0.81$.

2. Значимость коэффициентов.

   В данной модели на уровне значимости $\alpha = 0.05$ были признаны значимыми коэффициенты при факторах:
   S — заработная плата
   LE — ожидаемая продолжительность жизни
   HPI — индекс качества здравоохранения
   P — количество населения
   MF — соотношение количества женщин и мужчин
   MD — соотношение браков и разводов
   GRP — валовый региональный продукт
   QLI — индекс качества жизни
   β_0 — свободный член

   Так как мы используем метод исключения, то для построения модели № 2.2 все факторы с незначимыми коэффициентами будут исключены.

3. Нормальность распределения остатков.

   Тест Шапиро-Уилка: $W = 0.97008, p - value = 0.04756$.

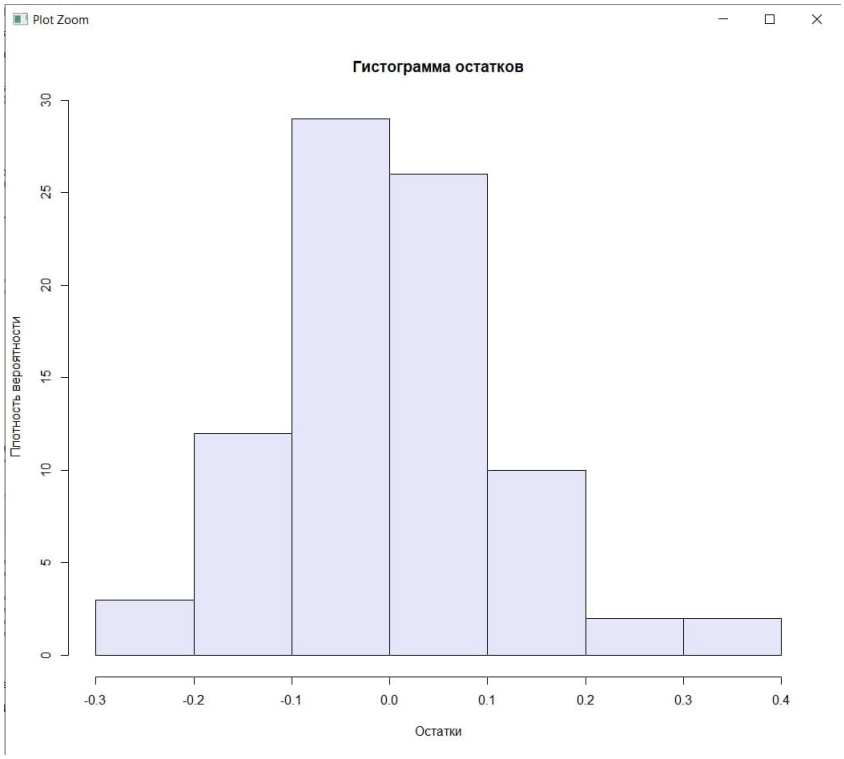

Рис. 8: Гистограмма остатков модели № 2.1

На уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, т.е. остатки имеют нормальное распределение.

4. Равенство нулю математического ожидания.
   Тест Стьюдента: $t = 1.7405e-16, p-value = 1$.
   На уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, т.е. мат-ожидание случайной переменной равно нулю.

5. Гомоскедастичность.
   Тест Бройша-Пагана: $BP = 17.587, p-value = 0.1288$.
   Тест Голдфелда-Квандта: $GQ = 2.1986, p-value = 0.0189$.
   На уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, т.е. дисперсия случайной переменной постоянна.

6. Автокорреляция.
   Тест Бройша-Годфри: $LM = 2.1231, p-value = 0.1451$.
   На уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, т.е. остатки не зависят друг от друга.

7. Мультиколлинеарность.

   Коэффициент вздутия дисперсии VIF по факторам:

   S — 9.267883

   MR — 1.514091

   TMR — 3.020611

   HPI — 4.748508

   P — 1.742423

   LE — 3.598274

   UP — 2.695073

   MF — 4.055871

   ER — 4.959437

   MD — 1.779443

   GRP — 3.226013

   QLI — 4.343165

   У всех признаков коэффициент вздутия дисперсии меньше 10, значит мультиколлинеарность между признаками отсутствует.

Прогноз по данным 2019 года.

На рисунке 9 представлены результаты предсказания СКР по модели № 2.1. Коэффициент корреляции между реальными и предсказанными значениями = 0.86, что по шкале Чеддока означает сильную взаимосвязь.

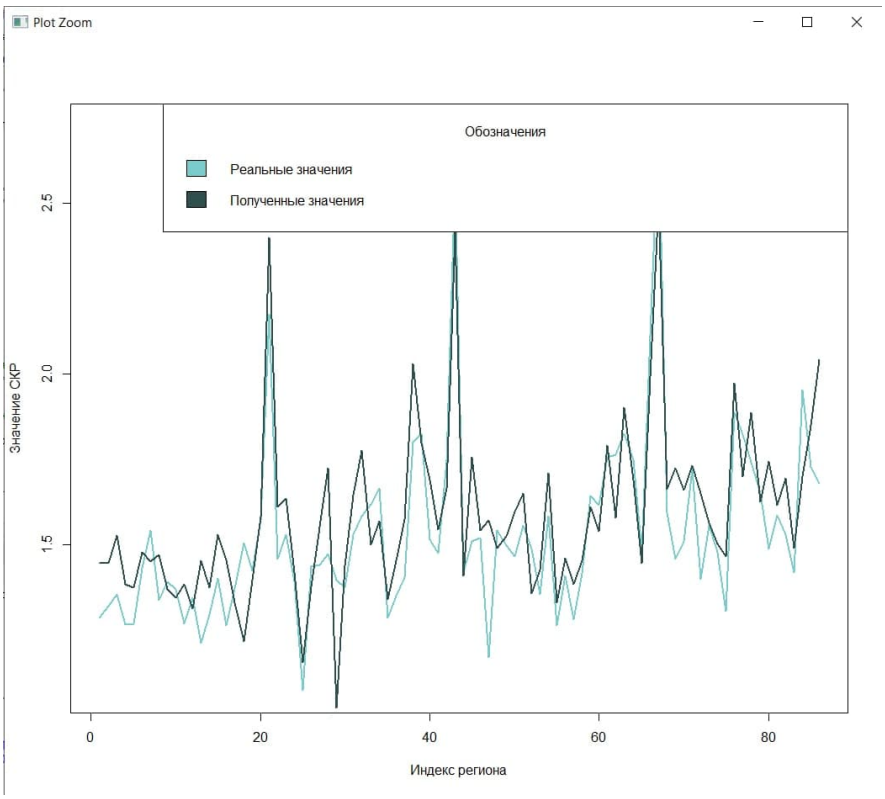

Рис. 9: Предсказание СКР по модели 2.1 на 2019 год и сравнение с реальными результатами

Модель 2.2: Данная модель построена на данных с исключением Москвы и Санкт-Петербурга и включала в себя факторы признанные значимыми после построения модели 2.1.

$$\begin{aligned} TFR = 7.939 - 0.00000933 * S - 0.005852 * HPI \\ +0.00003486 * P - 0.02023 * LE \\ -0.00269 * MF - 0.001295 * MD \\ +0.0000001639 * GRP - 0.008987 * QLI \end{aligned}$$

Для данной модели:

1. Коэффициент детерминации: $R^2 = 0.8013$.

2. Значимость коэффициентов.
   В данной модели на уровне значимости $\alpha = 0.05$ все коэффициенты были признаны значимыми.

3. Нормальность распределения остатков.

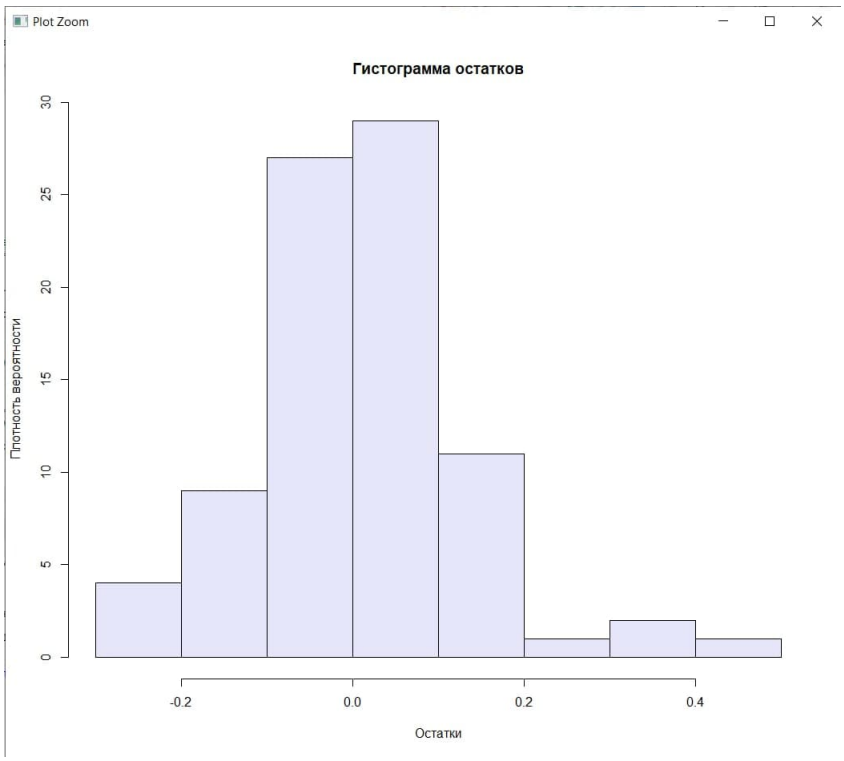

Рис. 10: Гистограмма остатков модели № 2.2

Тест Шапиро-Уилка: $W = 0.97445, p - value = 0.09308$.
На уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, т.е. остатки имеют нормальное распределение.

4. Равенство нулю математического ожидания.
   Тест Стьюдента: $t = 1.5053e - 16, p - value = 1$.
   На уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, т.е. мат-ожидание случайной переменной равно нулю.

5. Гомоскедастичность.
   Тест Бройша-Пагана: $BP = 19.079, p - value = 0.01445$.
   Тест Голдфелда-Квандта: $GQ = 1.9761, p - value = 0.02721$.
   На уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, т.е. присутствует гомоскедастичность.

6. Автокорреляция.
   Тест Бройша-Годфри: $LM = 3.3691, p - value = 0.06643$.
   На уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, т.е. отсутствует автокорреляция остатков.

7. Мультиколлинеарность.

   Коэффициент вздутия дисперсии VIF по факторам:

   S — 6.400884

   HPI — 3.995307

   P — 1.704643

   LE — 3.426514

   MF — 2.916055

   MD — 1.576184

   GRP — 2.824604

   QLI — 2.065157

   У всех признаков коэффициент вздутия дисперсии меньше 10, значит мультиколлинеарность между признаками отсутствует.

   Прогноз по данным 2019 года.

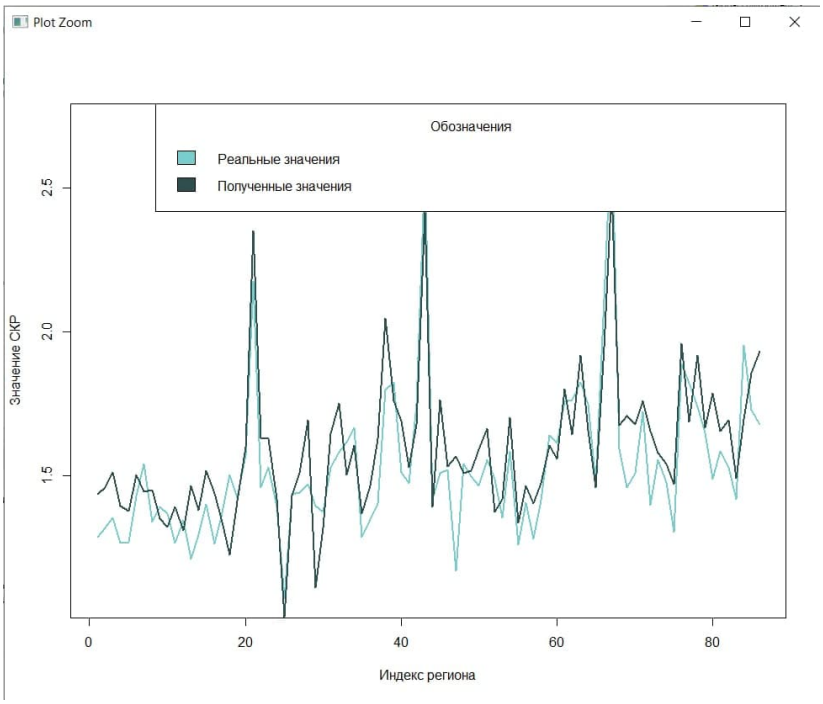

Рис. 11: Предсказание СКР по модели 2.2 на 2019 год и сравнение с реальными результатами

Коэффициент корреляции между реальными и предсказанными значениями = 0.86, что по шкале Чеддока означает сильную взаимосвязь. Оценки модели получены эффективными, несмещенными и состоятельными.

Модель 2.3: Отбор значимых факторов проводился с помощью дисперсионного однофакторного анализа, как и в примере 1.3.

После проведения однофакторного дисперсионного анализа относительно каждого фактора были отобраны следующие показатели:
HPI — индекс качества здравоохранения;
P — количество населения;
LE — ожидаемая продолжительность жизни;
UP — удельный вес городского населения;
MF — соотношение количества женщин и мужчин;
MD — соотношение браков и разводов;
GRP — валовый региональный продукт;
QLI — индекс качества жизни.
Далее построена модель множественной регрессии на основе этих факторов:

$$\begin{aligned} TFR = 6.313 - 0.003741 * HPI + 0.00003288 * P \\ -0.02408 * LE - 0.004991 * UP \\ -0.001138 * MF - 0.001227 * MD \\ +0.0000001105 * GRP - 0.008811 * QLI \end{aligned}$$

Для данной модели:

1. Коэффициент детерминации: $R^2 = 0.7813$.

2. Значимость коэффициентов.
   В данной модели на уровне $\alpha = 0.05$ все коэффициенты были признаны значимыми.

3. Нормальность распределения остатков.

   Тест Шапиро-Уилка: $W = 0.97856, p - value = 0.1743$.
   На уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, т.е. остатки имеют нормальное распределение.

4. Равенство нулю математического ожидания.
   Тест Стьюдента: $t = 2.019e - 16, p - value = 1$.

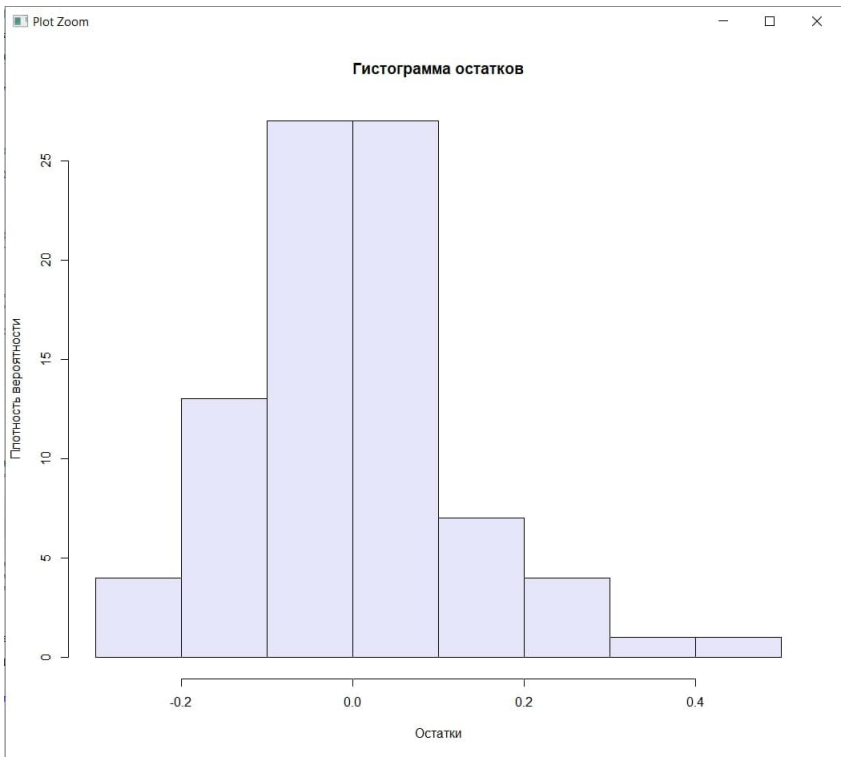

Рис. 12: Гистограмма остатков модели № 1.3

На уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, т.е. математическое ожидание случайной переменной равно нулю.

5. Гомоскедастичность.
   Тест Бройша-Пагана: $BP = 25.012, p - value = 0.001547$.
   Тест Голдфелда-Квандта: $GQ = 2.5807, p - value = 0.003969$.
   На уровне значимости $\alpha = 0.01$ гипотеза H_0 отвергается, т.е. присутствует гетероскедастичность.

6. Автокорреляция.
   Тест Бройша-Годфри: $LM = 1.8757, p - value = 0.1708$.
   На уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, т.е. отсутствует автокорреляция остатков.

7. Мультиколлинеарность.
   Коэффициент вздутия дисперсии VIF по факторам:
   HPI — 3.342168
   P — 1.708812
   LE — 3.448986

UP — 1.794516
MF — 1.588185
MD — 1.599150
GRP — 1.708339
QLI — 2.211987

У всех признаков коэффициент вздутия дисперсии меньше 10, значит мультиколлинеарность между признаками отсутствует.

Прогноз по данным 2019 года

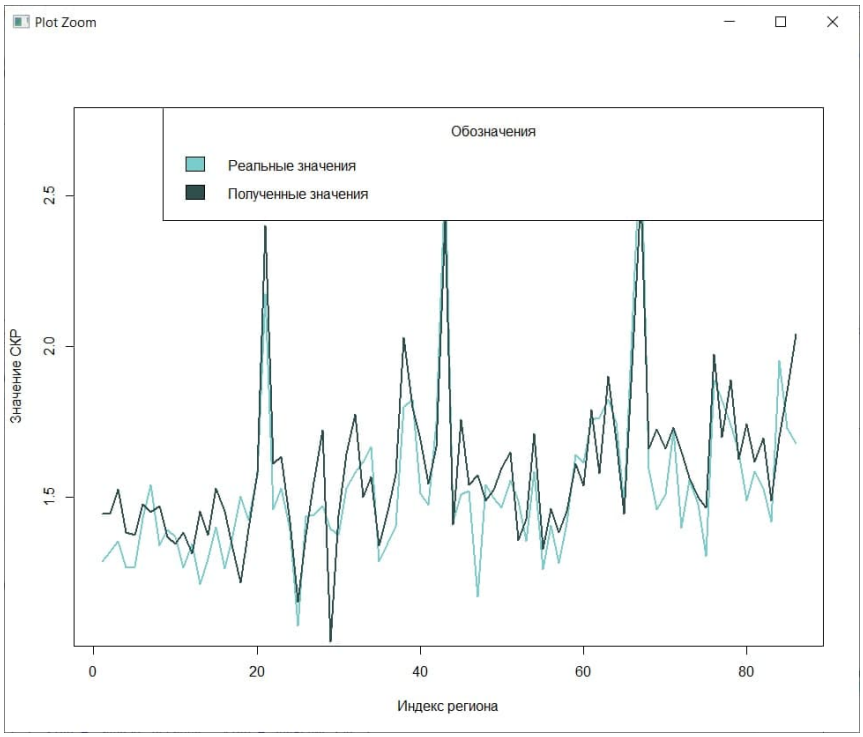

Рис. 13: Предсказание СКР по модели 2.3 на 2019 год и сравнение с реальными результатами

Коэффициент корреляции между реальными и предсказанными значениями = 0.85, что по шкале Чеддока означает сильную взаимосвязь.

3.3 Построение моделей с общим коэффициентом рождаемости

В этом пункте построим модели множественной регрессии, где зависимым фактором будет общий коэффициент рождаемости. После чего

сравним, есть ли разница между факторами, влияющими на СКР и на ОКР.

Модель 3.1: Модель 3.1 построена уже с исключением факторов с незначимыми коэффициентами на данных без Москвы и Санкт-Петербурга

$$\begin{aligned} B = 49.18 - 0.00004757 * S - +0.005262 * MR \\ -0.02908 * HPI + 0.0003738 * P \\ -0.002129 * MF - 0.01161 * MD \\ +0.000001009 * GRP - 0.08387 * QLI \end{aligned}$$

Для данной модели:

1. Коэффициент детерминации: $R^2 = 0.8249$.

2. Значимость коэффициентов.
   В данной модели на уровне значимости $\alpha = 0.05$ все коэффициенты были признаны значимыми.

3. Нормальность распределения остатков.

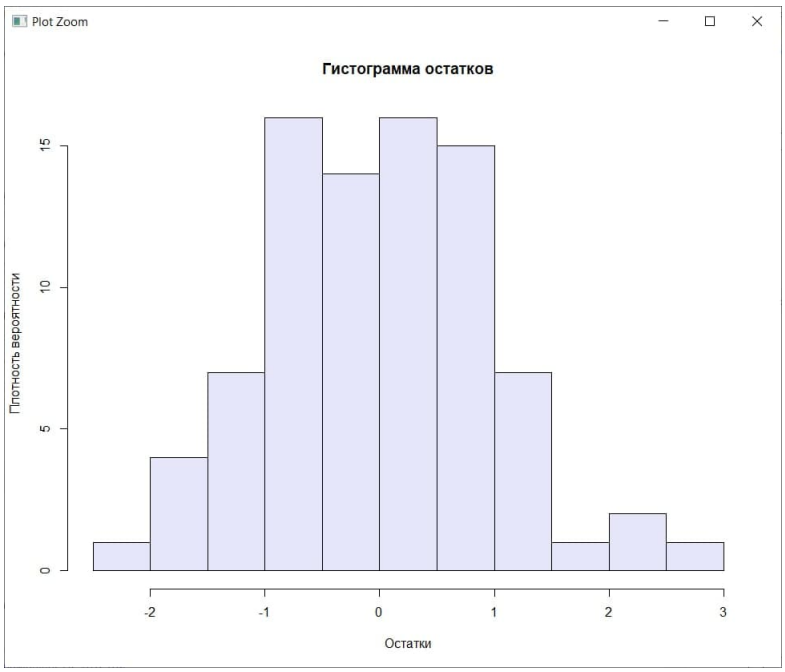

Рис. 14: Гистограмма остатков модели № 3.1

Тест Шапиро-Уилка: $W = 0.99072, p - value = 0.8164$.

На уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, т.е. остатки имеют нормальное распределение.

4. Равенство нулю математического ожидания.
   Тест Стьюдента: $t = -3.666e - 16, p - value = 1$.
   На уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, т.е. мат-ожидание случайной переменной равно нулю.

5. Гомоскедастичность.
   Тест Бройша-Пагана: $BP = 20.404, p - value = 0.008911$.
   Тест Голдфелда-Квандта: $GQ = 1.5567, p - value = 0.1044$.
   В одном тесте на уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, в другом — отвергается, значит будем считать, что гипотеза отвергается и остатки гетероскедастичны.

6. Автокорреляция.
   Тест Бройша-Годфри: $LMtest = 3.9087, p - value = 0.04804$.
   На уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, т.е. отсутствует автокорреляция остатков.

7. Мультиколлинеарность.
   Коэффициент вздутия дисперсии VIF по факторам:
   S — 6.400769
   MR — 1.412631
   HPI — 2.024780
   P — 1.699022
   MF — 2.890405
   MD — 1.471334
   GRP — 2.765700
   QLI — 2.362617
   У всех признаков коэффициент вздутия дисперсии меньше 10, значит, мультиколлинеарность между признаками отсутствует.

Прогноз по данным 2019 года.

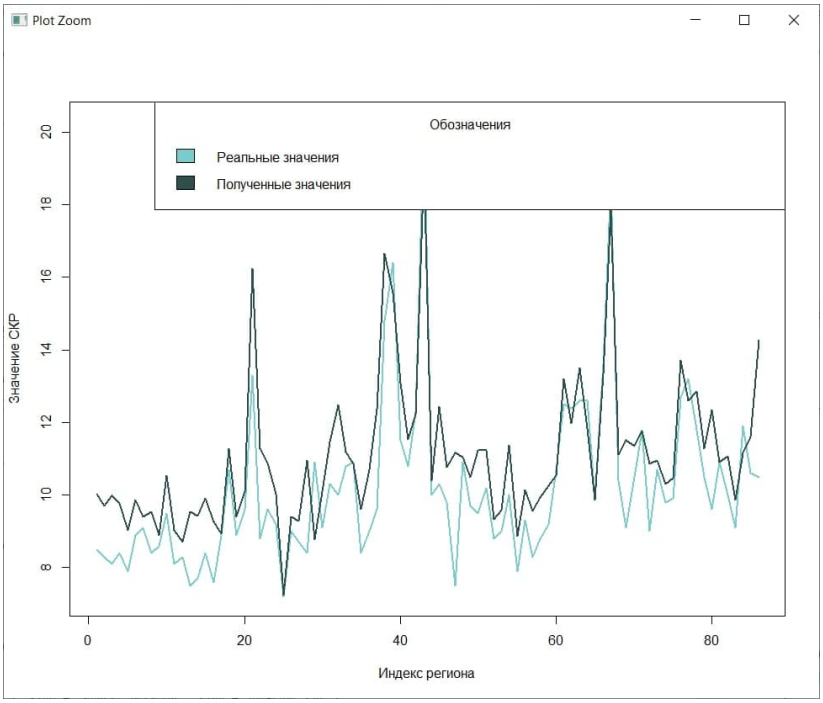

Рис. 15: Предсказание СКР по модели 3.2 на 2019 год и сравнение с реальными результатами

Коэффициент корреляции между реальными и предсказанными значениями = 0.89, что по шкале Чеддока означает сильную взаимосвязь.

Модель 3.2: Отбор значимых факторов проводился с помощью дисперсионного однофакторного анализа, как и в примере 1.3 и 2.3, а также были исключены факторы, признанные незначимыми после построения.

$$\begin{aligned} B = 39.36 - 0.01287 * HPI + 0.0003822 * P \\ -0.01495 * MF - 0.01176 * MD \\ +0.0000006183 * GRP - 0.07921 * QLI \end{aligned}$$

Для данной модели:

1. Коэффициент детерминации: $R^2 = 0.7931$.

2. Значимость коэффициентов.

   В данной модели на уровне $\alpha = 0.05$ все коэффициенты были признаны значимыми.

3. Нормальность распределения остатков.

   Тест Шапиро-Уилка: $W = 0.9779, p - value = 0.1578$.

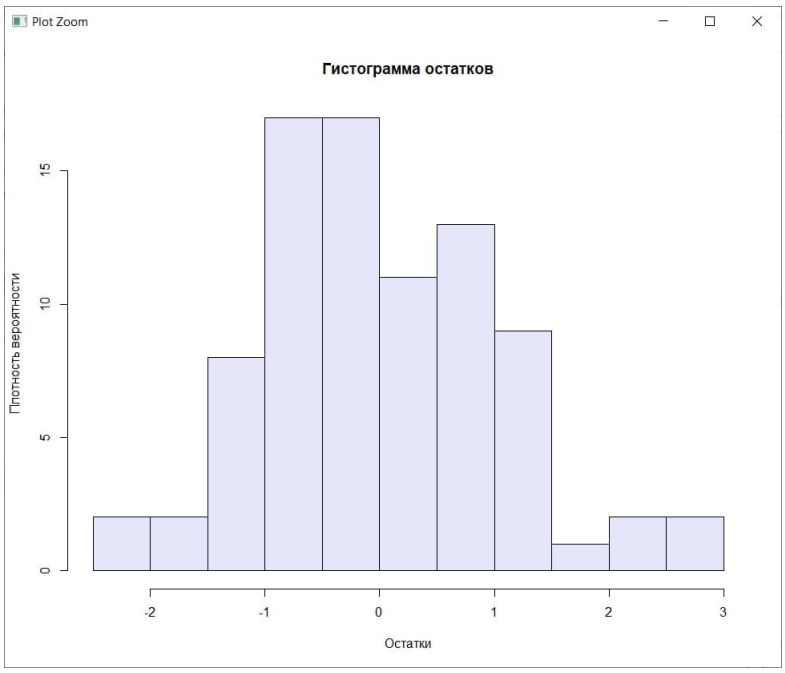

Рис. 16: Гистограмма остатков модели № 3.2

На уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, т.е. остатки имеют нормальное распределение.

4. Равенство нулю математического ожидания.
   Тест Стьюдента: $t = 1.2719e - 16, p - value = 1$.
   На уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, т.е. мат-ожидание случайной переменной равно нулю.

5. Гомоскедастичность.
   Тест Бройша-Пагана: $BP = 20.412, p - value = 0.002339$.
   Тест Голдфелда-Квандта: $GQ = 2.1155, p - value = 0.01478$.
   Как и в прошлой модели, в одном тесте гипотеза на уровне значимости принимается, а в другом — отвергается, следовательно, будем считать, что гипотеза отвергается и присутствует гетероскедастичность остатков.

6. Автокорреляция.
   Тест Бройша-Годфри: $LMtest = 3.6511, p - value = 0.05603$.
   На уровне значимости $\alpha = 0.01$ гипотеза H_0 принимается, т.е. отсутствует автокорреляция остатков.

7. Мультиколлинеарность.

   Коэффициент вздутия дисперсии VIF по факторам:

   HPI — 1.27092

   P — 1.697857

   MF — 1.505926

   MD — 1.458168

   GRP — 1.531534

   QLI — 1.788242

   У всех признаков коэффициент вздутия дисперсии меньше 10, значит мультиколлинеарность между признаками отсутствует.

Прогноз по данным 2019 года.

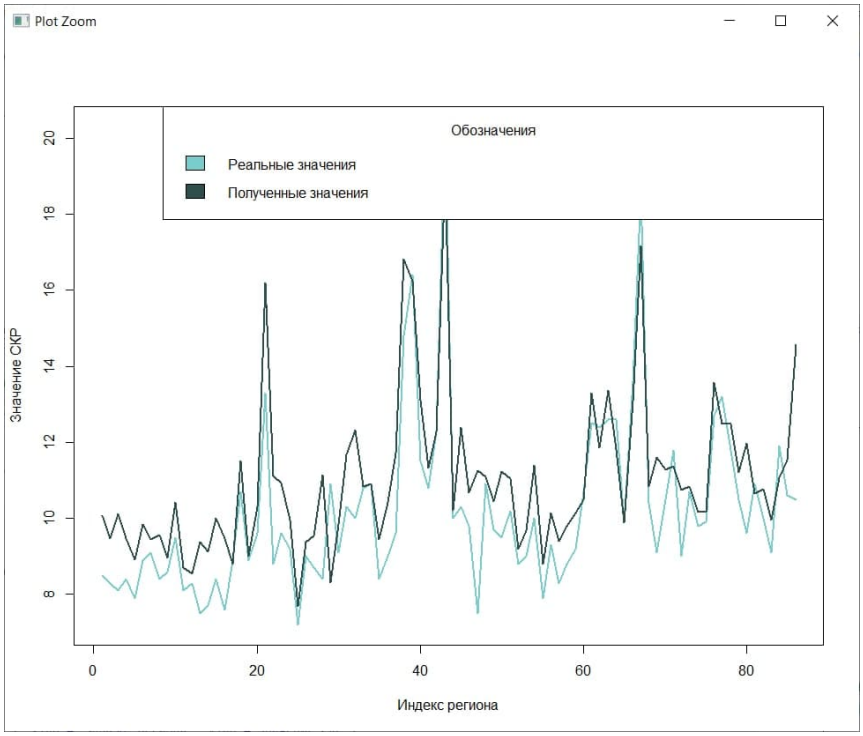

Рис. 17: Предсказание СКР по модели 3.2 на 2019 год и сравнение с реальными результатами

Коэффициент корреляции между реальными и предсказанными значениями = 0.88, что по шкале Чеддока означает сильную взаимосвязь.

Глава 4. Сравнение моделей

Сравнение моделей будем проводить, ориентируясь на несколько показателей: значение коэффициента детерминации, ошибку аппроксимации, критерий Акаике и коэффициент корреляции значения зависимой переменной за 2019 год с предсказанным значением этой переменной.

Критерий Акаике часто применяется для выбора наилучшей модели, он не только оценивает адекватность модели, но и штрафует за использование лишних параметров. Поэтому наилучшей считается та модель, у которой значение показателя Акаике наименьшее. На языке R сравнение моделей методом Акаике производится функцией AIC.

Ошибку аппроксимации применяют для выбора лучшей модели. Она показывает среднее отклонение расчетных значений от фактических. Соответственно, чем меньше значение, тем лучше. Ошибка аппроксимации рассчитывается по формуле:

$$A = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \tilde{y}_i}{y_i}\right| * 100\%$$

Итак, перейдем к сравнению моделей.

1. Сравнение моделей по первому набору данных.

Модель	R^2	Ошибка аппроксимации	AIC	cor
Модель 1.1	0.79	5,8%	-89.18113	0.87
Модель 1.2	0.73	6,3%	-77.06989	0.86
Модель 1.3	0.77	6,2%	-87.21929	0.88

Таблица 1: Сравнение моделей № 1.1, № 1.2, № 1.3

2. Сравнение моделей по второму набору данных.

3. Сравнение моделей с ОКР.

Теперь проведем сравнение как внутри каждой группы, так и между ними. На выборке, где были исключены Москва и Санкт-Петербург, модели

Модель	R^2	Ошибка аппроксимации	AIC	cor
Модель 2.1	0.81	5,5%	-91.62197	0.86
Модель 2.2	0.80	5,7%	-94.13108	0.86
Модель 2.3	0.78	5,9%	-86.10243	0.85

Таблица 2: Сравнение моделей № 2.1, № 2.2, № 2.3

Модель	R^2	Ошибка аппроксимации	AIC	cor
Модель 3.1	0.81	6,8%	250.0296	0.89
Модель 3.2	0.80	7,5%	260.0599	0.88

Таблица 3: Сравнение моделей № 3.1, № 3.2

получались качественнее. При почти равных предсказательных способностях (коэффициентах корреляции) ошибки аппроксимации в моделях второго набора данных меньше, а коэффициенты детерминации выше. Также среди моделей по второму набору данных только одна не удовлетворяет предпосылке о гомоскедастичности, в отличие от моделей первого набора, где нарушаются и другие предпосылки. Исключение Москвы и Санкт-Петербурга положительно повлияло на качество моделей с точки зрения соответствия требованиям регрессионного анализа, поэтому можно сделать вывод о том, что лучше строить модели по рождаемости для Москвы и Петербурга отдельно от других регионов. С учетом всего вышесказанного, наилучшей моделью, предсказывающей суммарный коэффициент рождаемости, выберем модель № 2.2: $TFR = 7.939 - 0.00000933 * S - 0.005852 * HPI + 0.00003486 * P - 0.02023 * LE - 0.00269 * MF - 0.001295 * MD + 0.0000001639 * GRP - 0.008987 * QLI$. Она удовлетворяет всем постулатам регрессионного анализа, считается лучшей по критерию Акаике среди моделей второго набора, имеет хорошую предсказательную способность, ненамного отличающуюся от других и не имеет незначимых факторов.

Теперь сравним модели, предсказывающие суммарный коэффициент рождаемости и общий коэффициент рождаемости. Будем сравнивать модели второго и третьего пунктов главы №3, так как и те и другие строились на наборе данных без Москвы и Санкт-Петербурга. Сравнивать качество моделей здесь не имеет смысла, так как влияние рассматривается на два разных фактора, поэтому сравним различие включаемых факторов в мо-

делях второй и третьей выборки. В общем факторы в моделях одинаковые, но можно заметить, что в моделях второй выборки присутствует фактор LE — ожидаемая продолжительность жизни при рождении, также он является значимым и в модели №1.3. Значит, можно сделать вывод о том, что при прогнозировании СКР следует учитывать этот фактор.

Если выбирать модель для предсказания общего коэффициента рождаемости, то из предложенных лучше выбрать модель №3.1, так как по всем характеристикам, представленным в таблице 3, модель будет наилучшей.

Лучшие характеристики наблюдаются по показателям: уровень здравоохранения (HPI), численность населения в регионе (P), соотношение мужчин и женщин (MF), соотношение браков и разводов (MD), качество жизни в регионе (QLI) и валовый региональный продукт на душу населения (GRP), так как эти показатели являются значимыми во всех моделях.

Очевидно, что с помощью политических мер сложно повлиять на соотношение мужчин и женщин, браков и разводов, а также на численность населения в регионе, в отличие от уровня здравоохранения и качества жизни.

Выводы

Таким образом, по результатам проведенного исследования:

- Собраны данные по 14 социально-экономическим показателям, имеющим влияние на рождаемость;
- Сформированы два набора данных для построения моделей и набор данных для проверки качества моделей;
- Построены два типа моделей с разными зависимыми факторами;
- Проведен сравнительный анализ моделей;
- Предложены модели для двух факторов рождаемости;
- Выявлены факторы, оказывающие наибольшее влияние на предлагаемые показатели.

К перспективам, в направлении которых возможно продолжение настоящего исследования, можно отнести следующее:

1. Включение в рассмотрение новые социально-экономических факторов, которые, возможно, имеют влияние на рождаемость;
2. При добавлении факторов расширение выборки посредством добавления других территориальных единиц;
3. Расширение границ задачи и переход к построению моделей на основе выборки по странам.

Заключение

Главной целью данной работы являлась разработка подхода к пониманию того, какие социально-экономические факторы влияют на рождаемость в регионах России. Исходя из полученных результатов, можно предположить, на какие сферы жизни может делать упор государство, чтобы поднять рождаемость в регионах, и, соответственно, в стране. Такого рода задача продиктована стремлением помочь найти ориентир при решении важнейших демографических проблем острых и актуальных для России.

Предложенный подход основан на построении многофакторной регрессионной модели методом исключения незначимых факторов на данных, находящихся в открытом доступе.

Результатом данной работы является построение 8 моделей множественной регрессии, на основании которых был сделан вывод о том, что рождаемость в большей степени зависит от таких показателей, как уровень здравоохранения, численность населения в регионе, соотношение мужчин и женщин, соотношение браков и разводов, качество жизни в регионе и валовый региональный продукт на душу населения.

Также в качестве выводов из проведенного исследования следует отметить, что рассчитывать показатели рождаемости по регионам следует отдельно от показателей рождаемости по Москве и Санкт-Петербургу.

После сравнения были выделены две наилучших моделей для предсказания коэффициента рождаемости:

- Для предсказания суммарного коэффициента рождаемости:
  $TFR = 7.939 - 0.00000933 * S - 0.005852 * HPI + 0.00003486 * P - 0.02023 * LE - 0.00269 * MF - 0.001295 * MD + 0.0000001639 * GRP - 0.008987 * QLI$

- Для предсказания общего коэффициента рождаемости:
  $B = 49.18 - 0.00004757 * S - +0.005262 * MR - 0.02908 * HPI + 0.0003738 * P - 0.002129 * MF - 0.01161 * MD + 0.000001009 * GRP - 0.08387 * QLI$

Список литературы

[1] Капица С.П., Курдюмов С.П., Малинецкий Г.Г. Синергетика и прогнозы будущего. М // Едиториал УРСС, 2003. 288 с.

[2] Демографическая статистика. М.: КНОРУС, 2015. 480 с.

[3] Молчанова Е.В. Оценка влияния социально-экономического развития на региональные демографические процессы // Вестник Алтайской академии экономики и права. – 2019. – № 4 (часть 2) – С. 252-258

[4] Казенин К.И. Рождаемость в России в 2020 году: Региональная динамика // мониторинг экономической ситуации в России, №5 (137), март 2021 Г, 19 с.

[5] Регионы России. Социально-экономические показатели.2020 // статистический сборник, Москва, 2020: https://rosstat.gov.ru (дата обращения: 20.04.2021.)

[6] Мигунова О. В. Влияние социально-экономических факторов на демографические процессы в России и мире/ О. В. Мигунова // Международный научно-исследовательский журнал. — 2013. — № 7 (14) Часть 5. — С. 50—52. (дата обращения: 17.05.2021.).

[7] Брюшинкина А.А., Грозина А.В., Эконометрический анализ влияния социально- экономических факторов на уровень рождаемости в регионах российской федерации // Хроноэкономика. 2019. №2 (15).(дата обращения: 17.05.2021).

[8] Карасельникова М.В., Эконометрический анализ влияния социально-экономических факторов на уровень рождаемости в РФ // Хроноэкономика. 2019. №2 (15). (дата обращения: 17.05.2021).

[9] Булатов Р.А. Рождаемость как социально-демографический процесс // диссертация кандидата социологических наук, 22.00.03, 2005 г, 137с

[10] Население России 2017: двадцать пятый ежегодный демографический доклад // Издательский дом НИУ ВШЭ, 2019,480 с.

[11] Архангельский В.Н.,Иванова А.Е., Рыбаковский Л.Л., Рязанцев С.В. Практическая демография // Под редакцией Л. Л. Рыбаковского М., ЦСП, 2005, 280 с.

[12] R manual: https://cran.r-project.org/manuals.html (дата обращения: 17.05.2021.)

[13] Елисеева И. И., Курышева С. В., Костеева Т. В., Михайлов Б. А. Эконометрика // под ред. Елисеева И. И. М.: Финансы и статистика, 2003. 344 с.

[14] Каморников С. Ф. , Каморников С. С. Эконометрика // учеб. пособие. – М. // Интеграция, 2012, 262 с.

[15] Магнус Я. Р., Катышев П. К., Пересецкий А. А. Эконометрика. Начальный курс. Изд. 6-е. М.: Дело, 2004. 576 с.

[16] Буре В. М., Парилина Е. М., Седаков А. А. Методы прикладной статистики в R и Excel // 3 изд. Лань, 2018. 152 с.

[17] Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: классификация и снижение размерности. М. // Финансы и статистика, 1989. - 607 с.

[18] Домбровский, В.В. Эконометрика // Учебное пособие // В.В. Домбровский . – М : Издательский дом "Новый учебник 2004 . – 342 с.

[19] Буре В. М., Парилина Е. М.Теория вероятностей и математическая статистика // Издательство "Лань 2013. - 416 с.

[20] Шеффе Г. Дисперсионный анализ // Перевод с английского Севастьянова Б.А., Чистякова В.П. // Москва, 1980. - 512 с.

[21] Лакин Г.Ф. Биометрия // Учебное пособие, 4-е изд., 1990. - 352 с.

[22] Население России 2007: Пятнадцатый ежегодный демографический доклад // Отв. ред. А. Г. Вишневский. М.: ГУ-ВШЭ, 2009

[23] Anna Yurko, 2013. Assessing the Impact of the Maternity Capital Policy in Russia Using a Dynamic Model of Fertility and Employment // UCL SSEES Economics and Business working paper series 125, UCL School of Slavonic and East European Studies (SSEES).

[24] Kopeykina V. The maternity capital's impact on birth intervals in Russia: Survival analysis of the transition from the 1st to 2nd child // Stockholm University, Faculty of Social Sciences, Department of Sociology.

[25] Мигунова О. В. Влияние социально-экономических факторов на демографические процессы в России и мире // Международный научно-исследовательский журнал. — 2013. — № 7 (14) Часть 5. — С. 50—52. (дата обращения: 17.05.2021.).

[26] Мерков А. М., Поляков Л. Е. Санитарная статистика (пособие для врачей) // издательство "Медицина 1974 г., 384 с.

[27] Индекс качества жизни в регионах России: https://riarating.ru (дата обращения: 20.04.2021.)

[28] Индекс здравоохранения: https://roscongress.org (дата обращения: 20.04.2021.)

[29] RPubs: https://rpubs.com/ (дата обращения: 17.05.2021.)

Приложение

Приложение 1. Результаты построения графиков с помощью функции boxplot для определения наличия аномальных значений по каждому фактору.

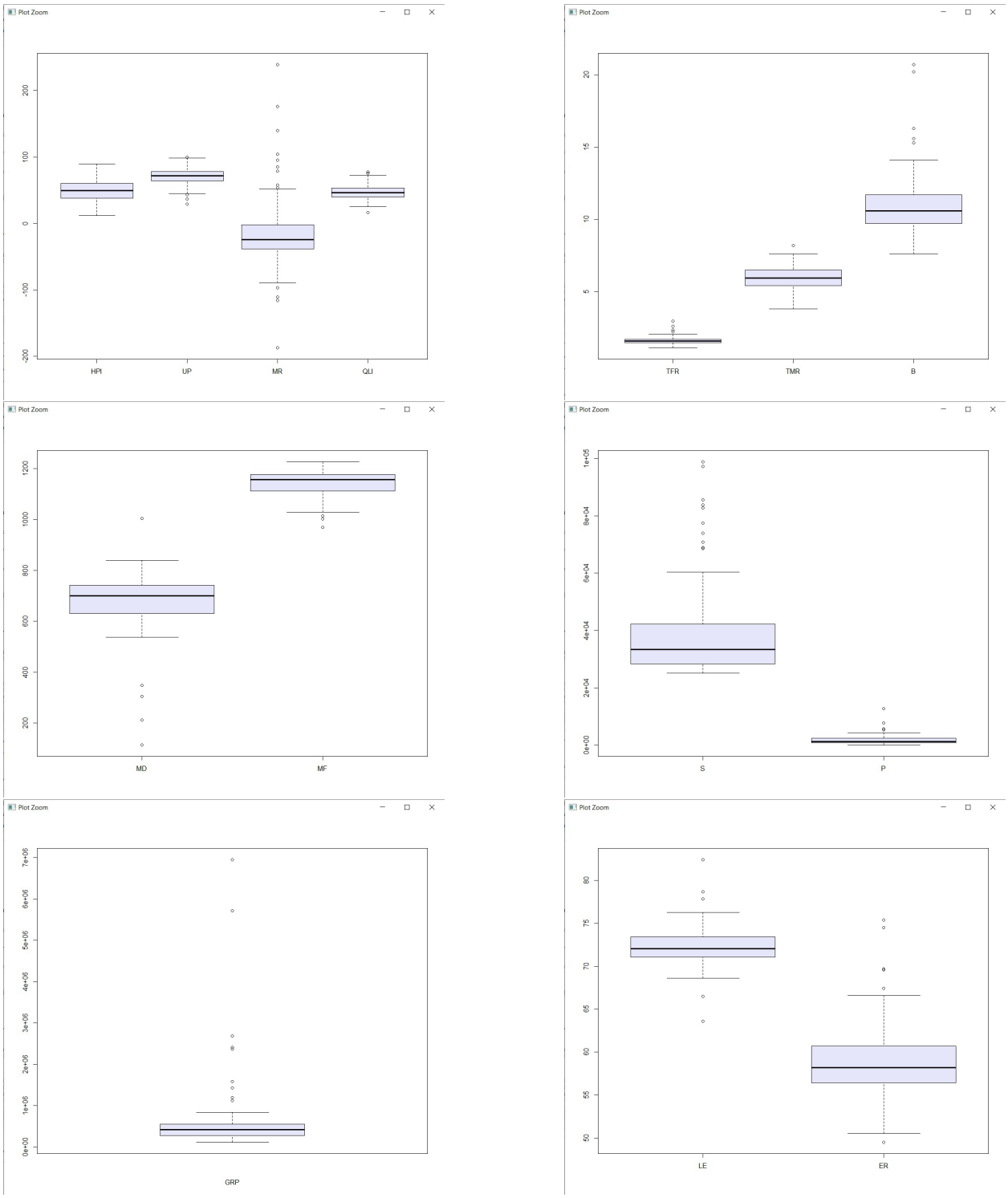

Рис. 18: Результаты построения графиков с помощью функции boxplot