**Национальный исследовательский университет**

**«Высшая школа экономики»**

Факультет бизнеса и менеджмента

Отчет по дисциплине Эконометрика

**Выполнили:**

студенты 3 курса 131 группы

Полушина Елизавета

Семина Мария

**Проверили:**

Поляков К.Л.

Иванова М.Е.

2016

**Оглавление**

[1. Общая постановка задачи 4](#_Toc453016492)

[1.1 Формулировка прикладной проблемы 4](#_Toc453016493)

[1.2 Потенциальные потребители решения; задачи, которые они смогут решать, используя полученные результаты 4](#_Toc453016494)

[1.3. Основные гипотезы, которые планируется проверить в рамках решения задачи 4](#_Toc453016495)

[1.4. Основные источники данных 4](#_Toc453016496)

[1.5. Публикации по данной тематике 5](#_Toc453016497)

[2. Предварительный анализ собранных данных 6](#_Toc453016498)

[2.1. Анализ особенностей данных: потенциальные ошибки и пропущенные значения, группы и выбросы 6](#_Toc453016499)

[2.1.1. Графический анализ количественных переменных. Гистограммы. 6](#_Toc453016500)

[2.1.2. Числовой анализ 12](#_Toc453016501)

[2.2. Анализ статистической связи. 13](#_Toc453016502)

[2.2.1. Графический анализ пары «числовая зависимая переменная – качественная независимая переменная». Диаграмма Бокса-Уискера (Box-Whisker). 13](#_Toc453016503)

[2.2.2. Графический анализ пары «числовая зависимая переменная – числовая независимая переменная». Диаграмма рассеивания (Scatter plot). 16](#_Toc453016504)

[3. Спецификация базовой модели. 20](#_Toc453016505)

[3.1. Связь структуры базовой модели с основными гипотезами исследования. 20](#_Toc453016506)

[3.2. Связь структуры базовой модели с результатами предварительного анализа данных. 21](#_Toc453016507)

[1.2.1. Состав переменных 21](#_Toc453016508)

[4. Оценивание базовой модели. 24](#_Toc453016509)

[4.1. Анализ наличия мультиколлинеарности 24](#_Toc453016510)

[4.2. Анализ наличия гетероскедастичности 26](#_Toc453016511)

[4.2.1. Тест Парка 26](#_Toc453016512)

[2.2.3. Тест Глейзера 27](#_Toc453016513)

[4.2.2. Тест Голдфелда-Квандта 28](#_Toc453016514)

[4.3. Проверка гипотез 28](#_Toc453016515)

[5. Оптимизация базовой модели 33](#_Toc453016516)

[5.1. Анализ наличия мультиколлиниарности 35](#_Toc453016517)

[5.2. Анализ наличия гетероскедастичности 36](#_Toc453016518)

[5.3. Проверка гипотез 37](#_Toc453016519)

[6. Доверительные интервалы 40](#_Toc453016520)

[7. Обнаружение выбросов и влиятельных наблюдений в регрессионном анализе 42](#_Toc453016521)

[7.1. Расстояние Махаланобиса 42](#_Toc453016522)

[7.2. Расстояние Кука 42](#_Toc453016523)

[7.3. Ковариационное отношение 43](#_Toc453016524)

[7.4. Мера стандартизованная *DFFIT* (*DFFITS*) 43](#_Toc453016525)

[7.5. Мера стандартизованная *DFBETA* (*DFBETAS*) 43](#_Toc453016526)

[8. Эссе по найденным публикациям. 44](#_Toc453016527)

[8.1. Эссе по статье Б.Б. Прохорова, И.В. Горшковой, Е.В. Тарасовой «Условия жизни населения и общественное здоровье» 44](#_Toc453016528)

[8.2. Эссе по статье Б.Б. Прохорова «Социально-экономические особенности федеральных округов России и здоровье населения» 45](#_Toc453016529)

[8.3. Использование статей в настоящей работе. 46](#_Toc453016530)

# 1. Общая постановка задачи

## 1.1 Формулировка прикладной проблемы

Улучшение здоровья населения не может быть достигнуто только с помощью преобразований в системе медицинского обслуживания населения. Поэтому необходимо провести мероприятия на уровне всей страны, улучшающие значимые социально-экономические показатели и нацеленные на повышение качества здоровья населения. Какие социально-экономические факторы влияют на состояние здоровья населения на уровне всей страны? Какие мероприятия следует провести для улучшения показателей выявленных факторов?

## 1.2 Потенциальные потребители решения; задачи, которые они смогут решать, используя полученные результаты

Проблему поддержания и улучшения качества здоровья населения в рамках всей страны нужно решать комплексно, т.е. на всех организационных уровнях. Поэтому потенциальными потребителями нашего исследования могут быть:

-государственные учреждения,

-службы здравоохранения.

Используя полученные результаты исследования, потенциальные потребители смогут составить план проведения определенных мероприятий, направленных на повышение качества здоровья населения, выделить надлежащую сумму в бюджете страны и распределить ее по регионам России.

## 1.3. Основные гипотезы, которые планируется проверить в рамках решения задачи

- Чем больше численность населения региона с денежными доходами ниже прожиточного минимума, тем выше показатель заболеваемости.

- Подверженность заболеваниям населения с низким уровнем дохода возрастает с уменьшением объема использования бытовых услуг.

- Заболеваемость зависит от качества медицинского обслуживания.

## 1.4. Основные источники данных

Основным источником статистической информации о российских регионах послужили материалы, приводимые в сборниках Госкомстата (http://www.gks.ru/). Также были использованы данные Портала правовой статистики (http://crimestat.ru/). На основе презумпции зависимости уровня заболеваемости населения в регионах России от социально-экономических факторов был составлен предварительный список показателей, которые влияют на здоровье граждан.

Количественная зависимая переменная – заболеваемость на 1000 человек населения.

Количественные независимые переменные:

* Численность населения с денежными доходами ниже прожиточного минимума (в процентах от общей численности населения субъекта),
* Численность среднего медицинского персонала на 10 000 человек населения (человек),
* Удельный вес ветхого и аварийного жилищного фонда в общей площади всего жилищного фонда (в процентах),
* Объем бытовых услуг на душу населения (рублей).

Качественная независимая переменная – уровень преступности в регионах.

## 1.5. Публикации по данной тематике

В процессе исследования нами были изучены следующие публикации по выбранной теме:

- Б.Б. Прохоров, И.В. Горшкова, Е.В. Тарасова “Условия жизни населения и общественное здоровье” (URL: http://www.ecfor.ru/pdf.php?id=2003/5/10)

- Б.Б. Прохоров “Социально-экономические особенности федеральных округов

России и здоровье населения” (URL: http://cyberleninka.ru/article/n/sotsialno-ekonomicheskie-osobennosti-federalnyh-okrugov-rossii-i-zdorovie-naseleniya)

- Б.Б. Прохоров “Общественное здоровье и экономика” (URL: http://demoscope.ru/weekly/2007/0293/biblio01.php)

-И. Назарова “Динамика здоровья населения и роль реформ в его ухудшении” (URL: http://www.socpolitika.ru/rus/social\_policy\_research/applied\_research/document86.shtml)

-А.С. Киселев, М.Г. Шестаков, А.Ю. Михайлов “Зависимость здоровья населения от динамики уровня жизни” (URL: http://www.socpolitika.ru/rus/social\_policy\_research/349/document657.shtml)

# 2. Предварительный анализ собранных данных

## 2.1. Анализ особенностей данных: потенциальные ошибки и пропущенные значения, группы и выбросы

### 2.1.1. Графический анализ количественных переменных. Гистограммы[[1]](#footnote-1).

**1. Численность населения с денежными доходами ниже прожиточного минимума (в процентах от общей численности населения субъекта)**

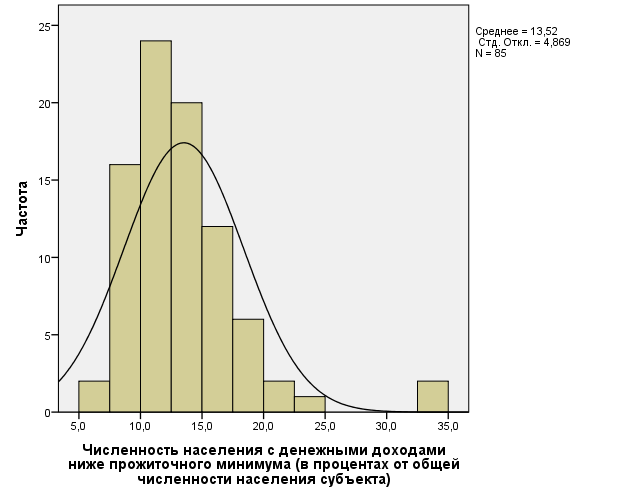


Рисунок 1

* **Форма гистограммы**

Полученная форма гистограммы схожа с формой нормальной кривой.

С помощью инструментов IBM SPSS Statistics произведем подсчет критерия эксцесса, позволяющего проверить «плосковершинность» или «узковершинность» эмпирического распределения, выраженную в числовой форме, и коэффициента асимметрии, меры отклонения распределения частоты от симметричного распределения. Распределение является нормальным, при условии, что показатели эксцесса и асимметрии находятся в диапазоне [-1;1]. Если показатели или асимметрии, или эксцесса не находятся в данном диапазоне, распределение нельзя назвать нормальным.

В таблице 1 указаны значения, выходящие за рамки диапазона (Ex = 7,034; As = 2,1). Следовательно, принимая за основную гипотезу, что данные взяты из нормально распределенной генеральной совокупности, а за альтернативную, соответственно, что данные взяты не из нормально распределенной генеральной совокупности, мы отвергаем основную гипотезу. В данном случае перед нами «узковершинность».

|  |  |  |
| --- | --- | --- |
| **Статистика** | | |
| Численность населения с денежными доходами ниже прожиточного минимума (в процентах от общей численности населения субъекта) | | |
| N | Допустимо | 85 |
| Пропущенные | 0 |
| Асимметрия | | 2,100 |
| Стандартная Ошибка асимметрии | | ,261 |
| Эксцесс | | 7,034 |
| Стандартная ошибка эксцесса | | ,517 |

Таблица 1

* **Выбросы**

Очевиден выброс. Выбросу соответствует значение 34,7, относящееся к региону Российской Федерации – Республика Калмыкия и Республика Тыва. Данный выброс можно объяснить очень высокими показателями безработицы в данных регионах (12.5% и 19.3% соответственно). Как следствие, в субъектах наблюдается самый низкий объем ВРП по стране, который в 16-20 раз меньше среднего значения по всем регионам РФ.

С помощью правила трех сигм определим, на самом ли деле перед нами выброс:

(P|X - \mu\,| < 3\sigma) = 2Ф(3)

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

34,7 – 13,52 < 3 \* 4,869

21,18 < 14,607

Неравенство не выполняется, значит, данное значение является выбросом. Можно изъять данные выбросы из анализа. Однако обнаруженные выбросы были не удалены из модели, потому как мы посчитали важным оставить имеющиеся маргинальные значения для дальнейшего хода анализа.

**2. Численность среднего медицинского персонала на 10 000 человек населения, человек**

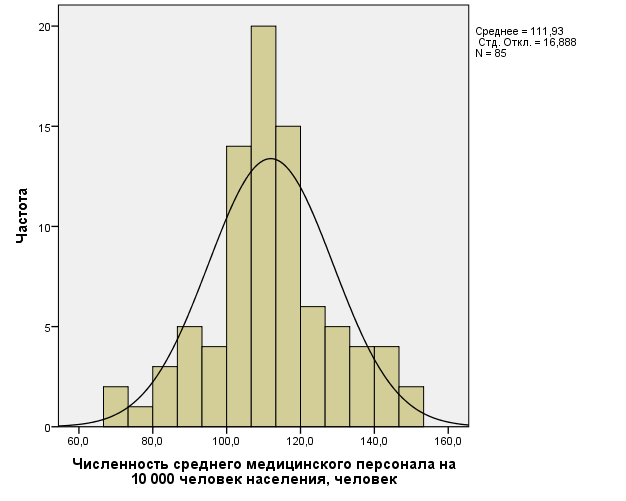


Рисунок 2

* **Форма гистограммы**

Полученная форма гистограммы схожа с формой нормальной кривой.

С помощью инструментов IBM SPSS Statistics произведем подсчет критерия эксцесса и асимметрии. В таблице 2 указаны значения как эксцесса, так и асимметрии, входящие в диапазон [-1;1] (Ex = 0,322; As = 0,089). Так, принимая за основную гипотезу, что данные взяты из нормально распределенной генеральной совокупности, а за альтернативную, соответственно, что данные взяты не из нормально распределенной генеральной совокупности, мы принимаем основную гипотезу.

|  |  |  |
| --- | --- | --- |
| **Статистика** | | |
| Численность среднего медицинского персонала на 10 000 человек населения, человек | | |
| N | Допустимо | 85 |
| Пропущенные | 0 |
| Асимметрия | | ,089 |
| Стандартная Ошибка асимметрии | | ,261 |
| Эксцесс | | ,322 |
| Стандартная ошибка эксцесса | | ,517 |

Таблица 2

* **Выбросы**

Выбросов не обнаружено.

**3. Удельный вес ветхого и аварийного жилищного фонда в общей площади всего жилищного фонда (в процентах)**

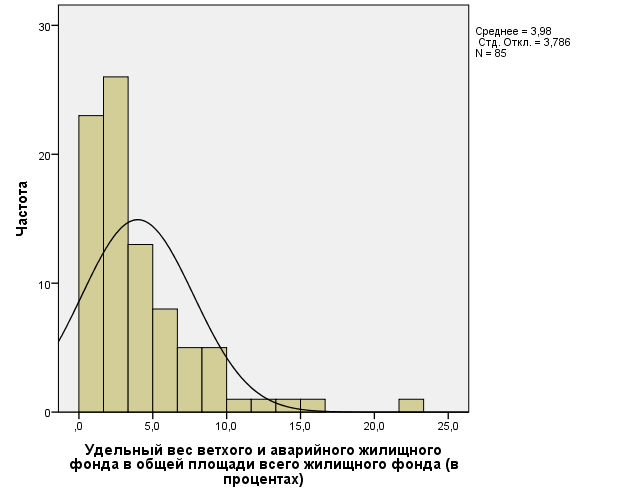


Рисунок 3

* **Форма гистограммы**

Правый хвост, очевидно, тяжелее, чем левый.

С помощью инструментов IBM SPSS Statistics высчитаем значение коэффициента асимметрии и эксцесса. Данные представлены на таблице 3. Коэффициенты значительно отличаются от нуля, не входят в промежуток[-1;1] (Ex = 7, 664; As = 2,376). Так, принимая за основную гипотезу, что данные взяты из нормально распределенной генеральной совокупности, а за альтернативную, соответственно, что данные взяты не из нормально распределенной генеральной совокупности, мы отвергаем основную гипотезу.

|  |  |  |
| --- | --- | --- |
| **Статистика** | | |
| Удельный вес ветхого и аварийного жилищного фонда в общей площади всего жилищного фонда (в процентах) | | |
| N | Допустимо | 85 |
| Пропущенные | 0 |
| Асимметрия | | 2,376 |
| Стандартная Ошибка асимметрии | | ,261 |
| Эксцесс | | 7,664 |
| Стандартная ошибка эксцесса | | ,517 |

Таблица 3

* **Выбросы**

Очевиден выброс. Выбросу соответствует значение 22,8, относящееся к региону Российской Федерации – Республика Ингушетия. Данный выброс можно объяснить рядом грубых нарушений со стороны правительства при использовании бюджетных средств в рамках федеральной целевой программы (ФЦП) «Социально-экономическое развитие Республики Ингушетия на 2010-2016 годы». Значительная часть расходов приходилась на капитальное строительство, включающее как промышленные объекты, так и социальные, однако ни один проект не был завершен согласно плану. Как следствие, в Ингушетии самый высокий показатель по аварийному и ветхому жилью, в частности, практически пятая часть всего жилья относится к этой категории.

С помощью правила трех сигм определим, на самом ли деле перед нами выброс:

22,8 – 3,98 < 3 \* 3,786

18,82 < 11,358

Неравенство не выполняется, значит, данное значение является выбросом. Можно изъять данный выброс из анализа. Однако обнаруженный выброс не был удален из модели, потому как мы посчитали важным оставить имеющиеся маргинальные значения для дальнейшего хода анализа.

**4. Объем бытовых услуг на душу населения**

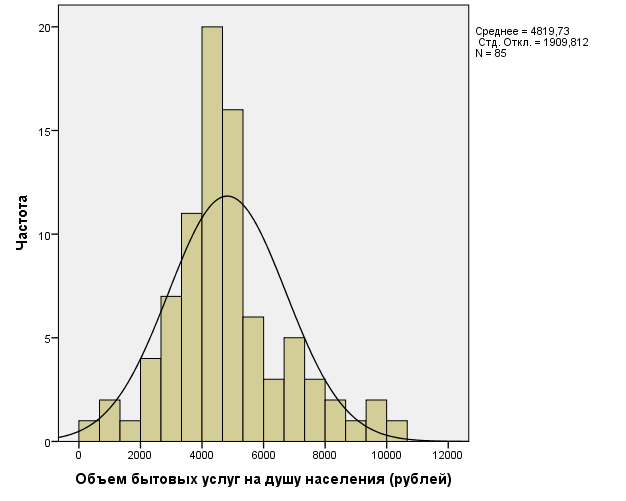


Рисунок 4

* **Форма гистограммы**

Полученная форма гистограммы схожа с формой нормальной кривой.

С помощью инструментов IBM SPSS Statistics произведем подсчет критерия эксцесса и асимметрии. В таблице 4 указаны значения, входящиее в диапазон [-1;1] (Ex = 0,656; As = 0,609). Так, принимая за основную гипотезу, что данные взяты из нормально распределенной генеральной совокупности, а за альтернативную, соответственно, что данные взяты не из нормально распределенной генеральной совокупности, мы принимаем основную гипотезу.

|  |  |  |
| --- | --- | --- |
| **Статистика** | | |
| Объем бытовых услуг на душу населения (рублей) | | |
| N | Допустимо | 85 |
| Пропущенные | 0 |
| Асимметрия | | ,609 |
| Стандартная Ошибка асимметрии | | ,261 |
| Эксцесс | | ,656 |
| Стандартная ошибка эксцесса | | ,517 |

Таблица 4

* **Выбросы**

Выбросов не обнаружено.

**5. Заболеваемость на 1000 человек населения**

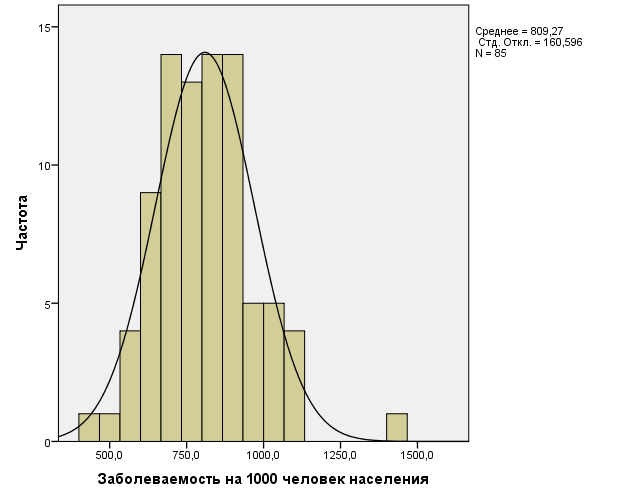


Рисунок 5

* **Форма гистограммы**

Полученная форма гистограммы схожа с формой нормальной кривой.

С помощью инструментов IBM SPSS Statistics произведем подсчет критерия эксцесса и асимметрии. В таблице 5 указано значение коэффициента эксцесса, выходящего за рамки диапазона [-1;1] (Ex = 1,737). Несмотря на то, что значение коэффициента асимметрии входит в диапазон [-1;1] (As = 0,721), если показатели или асимметрии, или эксцесса не находятся в данном диапазоне, распределение нельзя назвать нормальным. Следовательно, принимая за основную гипотезу, что данные взяты из нормально распределенной генеральной совокупности, а за альтернативную, соответственно, что данные взяты не из нормально распределенной генеральной совокупности, мы отвергаем основную гипотезу.

|  |  |  |
| --- | --- | --- |
| **Статистика** | | |
| Заболеваемость на 1000 человек населения | | |
| N | Допустимо | 85 |
| Пропущенные | 0 |
| Асимметрия | | ,721 |
| Стандартная Ошибка асимметрии | | ,261 |
| Эксцесс | | 1,737 |
| Стандартная ошибка эксцесса | | ,517 |

Таблица 5

* **Выбросы**

Очевиден выброс. Выбросу соответствует значение 1436,8, относящееся к региону Российской Федерации – Ненецкий автономный округ. Данный выброс можно объяснить проникновением ВИЧ-инфекции в популяцию народов субъекта с неуклонным ростом количества зараженных.

С помощью правила трех сигм определим, на самом ли деле перед нами выброс:

1436,8 – 809,27 < 3 \* 160,596

627,53 < 481,788

Неравенство не выполняется, значит, данное значение является выбросом. Можно изъять данный выброс из анализа. Однако обнаруженный выброс не был удален из модели, потому как мы посчитали важным оставить имеющиеся маргинальные значения для дальнейшего хода анализа.

### 2.1.2. Числовой анализ

В таблице 6 указаны минимум, максимум, среднее значение, стандартное отклонение и размах числовых величин. В таблицу включены как зависимая переменная (Заболеваемость на 1000 человек населения), так и независимые числовые переменные.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Описательные статистики** | | | | | |
|  | Мин | Макс | Среднее значение | Стандартное отклонение | Размах |
| Заболеваемость на 1000 человек населения | 434,7 | 1436,8 | 809,268 | 160,5963 | 1002,1 |
| Численность населения с денежными доходами ниже прожиточного минимума (в процентах от общей численности населения субъекта) | 6,9 | 34,7 | 13,515 | 4,8691 | 27,8 |
| Численность среднего медицинского персонала на 10 000 человек населения, человек | 71,1 | 152,6 | 111,933 | 16,8883 | 81,5 |
| Удельный вес ветхого и аварийного жилищного фонда в общей площади всего жилищного фонда (в процентах) | ,0 | 22,8 | 3,981 | 3,7858 | 22,8 |
| Объем бытовых услуг на душу населения (рублей) | 417 | 10102 | 4819,73 | 1909,812 | 9685 |

Таблица 6

Итак, числовой анализ выявил адекватность представленных данных.

## 2.2. Анализ статистической связи.

### 2.2.1. Графический анализ пары «числовая зависимая переменная – качественная независимая переменная». Диаграмма Бокса-Уискера (Box-Whisker).

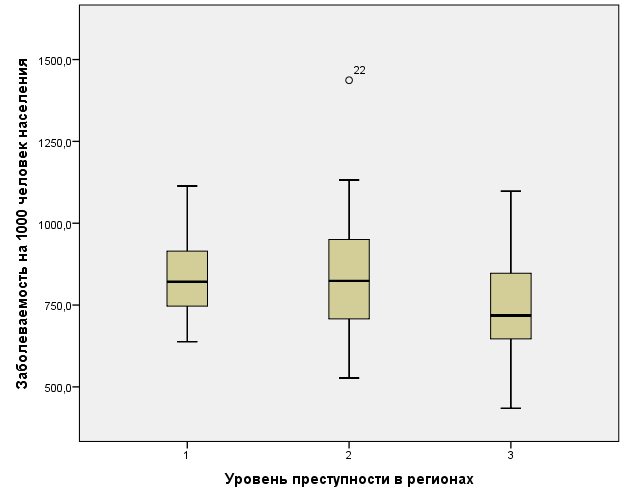


Рисунок 6

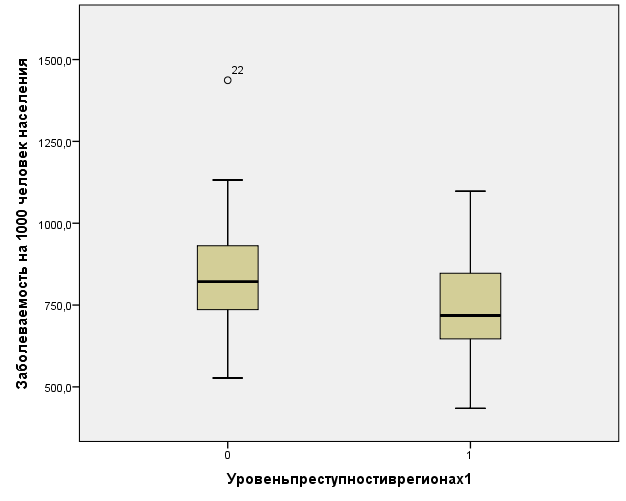
Проанализируем влияние уровня преступности на заболеваемость в регионах. Для этого разделим все субъекты РФ на 3 группы по уровню преступности: 1-наиболее высокий, 2 – средний, 3 – наиболее низкий.

Построив диаграммы Бокса-Вискера для данных групп переменных, можем отметить, что наименьшая средняя заболеваемость населения наблюдается в 3 группе(низкий уровень преступности). Наивысший показатель уровня заболеваемости почти совпадает у всех групп несмотря на то, что наибольший разброс значений заметен в группе с наименьшим показателем преступности. Т.е. в группе с меньшим количеством совершаемых преступлений существуют более низкие показатели заболеваемости, тогда как в остальных группах наименьший показатель заболеваемости сравнительно выше. Это может быть связано со стабильностью местного населения, а также уровнем жизни в регионе. Чем больше в данной местности людей, семи которых здесь проживают на протяжении поколений, чем больше домовладельцев (не арендаторов), а также чем выше уровень жизни населения, тем ниже уровень преступности. Именно в таких регионах с преобладанием патриархальных семей уровень заболеваемости значительно ниже.

Так как медианы 1-ой и 2-ой групп почти совпадают со средними значениями, имеет место симметрия существующих данных, когда как в 3-ей группе можем утверждать о наличии правосторонней асимметрии.

Наивысший показатель заболеваемости можно отметить во 2-ой группе. Кроме того, во 2-ой группе присутствует субъект с необычно высоким уровнем заболеваемости, так называемый выброс. Он может быть обусловлен проникновением ВИЧ-инфекции в популяцию народов Ямало-Ненецкого АО, которая продолжает свое распространение и, более того, на период 2014 года количество заболеваемых неуклонно росло.

Таким образом, нами было принято решение объединить 1-ую и 2-ую группу в связи с тем, что в 1-ой группе не существует показателей уровня заболеваемости, существенно отличающихся от показателей 2-ой группы, что помогает более наглядно продемонстрировать различия показателей двух новых образованных групп.



Проанализируем влияние уровня преступности на заболеваемость в регионах. Для этого разделим все субъекты РФ на 2 группы по уровню преступности: 1- высокий, 2 - низкий. Напомним, что при анализе 3-х групп нами были объединены первая и вторая в связи со схожестью показателей (первая входила во вторую группу).

Построив диаграммы Бокса-Вискера для данных групп переменных, можем отметить, что наименьшая средняя заболеваемость населения наблюдается во 2-ой группе(низкий уровень преступности). Наивысший показатель уровня заболеваемости наблюдается в 1-ой группе с высоким уровнем преступности. Наибольший разброс значений заметен в группе с наименьшим показателем преступности, т.е. в группе с меньшим количеством совершаемых преступлений существуют более низкие показатели заболеваемости. Это может быть интерпретировано через призму уровня жизни населения в том или ином регионе. Уровень преступности значительно выше в регионах с низким уровнем жизни равно как и уровень заболеваемости, в связи с чем прослеживается зависимость уровня заболеваемости от уровня преступности.

Так как медианы 1-ой группы почти совпадает со средним значением, имеет место симметрия существующих данных, когда как во 2-ой группе можем утверждать о наличии правосторонней асимметрии (медиана смещена вниз).

Наивысший показатель заболеваемости можно отметить в 1-ой группе. Кроме того, в 1-ой группе присутствует субъект с необычно высоким уровнем заболеваемости, так называемый выброс. Он может быть обусловлен проникновением ВИЧ-инфекции в популяцию народов Ямало-Ненецкого АО, которая продолжает свое распространение и, более того, на период 2014 года количество заболеваемых неуклонно росло.

### 2.2.2. Графический анализ пары «числовая зависимая переменная – числовая независимая переменная». Диаграмма рассеивания (Scatter plot).

**Заболеваемость на 1000 человек – Численность населения с денежными доходами ниже прожиточного минимума (в процентах от общей численности населения субъекта)**

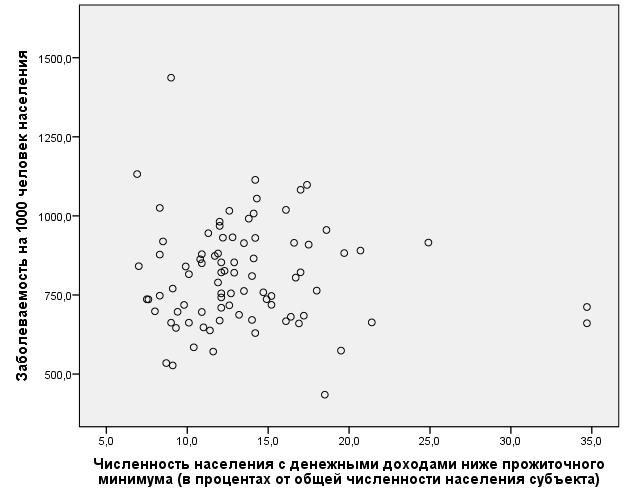


Рисунок 7

Зависимость схожа с линейной. Очевидны выбросы, выделенные ранее. Для того чтобы учесть данные выбросы можно использовать функцию .

**Заболеваемость на 1000 человек – Численность среднего медицинского персонала на 10 000 человек населения, человек**

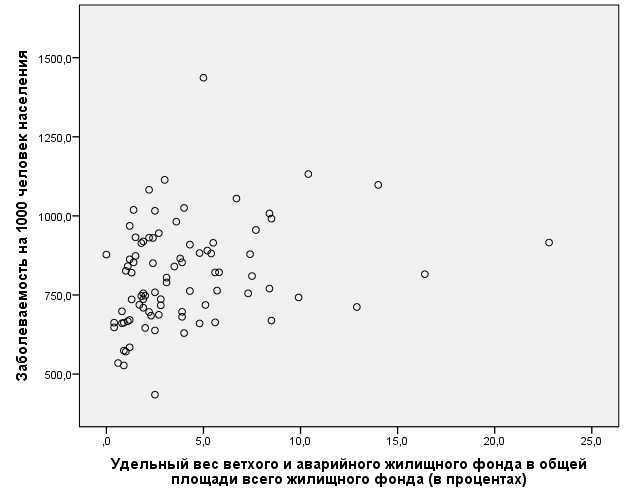


Рисунок 8

Судя по графику, зависимость схожа с линейной. Однако расположение точек также схоже с формой графика .

**Заболеваемость на 1000 человек – Удельный вес ветхого и аварийного жилищного фонда в общей площади всего жилищного фонда (в процентах)**

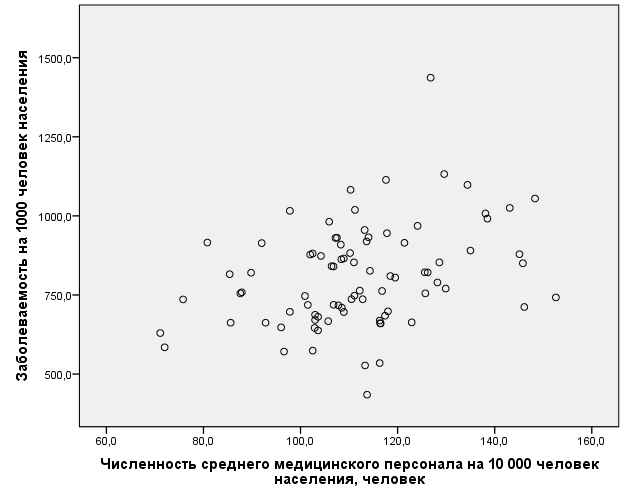


Рисунок 9

Зависимость схожа с линейной.

**Заболеваемость на 1000 человек – Объем бытовых услуг на душу населения (рублей)**

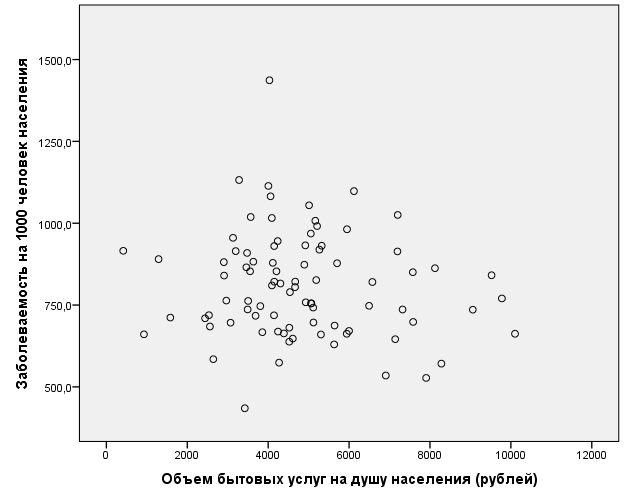


Рисунок 10

Судя по графику, зависимость схожа с линейной. Также зависимость на диаграмме схожа с формой гиперболы (1/x), сдвинутой вправо и вверх.

Таким образом, было выявлено, что лучшими функциональными формами для вхождения в настоящую регрессионную модель для переменных являются:

* X1 или X1;
* X2;
* X3;
* X4;
* X5.

Y = β0 + β1 X1 + β2 X2 + β3 X3 + β4 X4 + dX5 или

Y = β0 + β1 X1 + β2 X2 + β3 X3 + β4 X4 + dX5

**Дисперсионный анализ ANOVA**

Проведем дисперсионный анализ ANOVA.

За нулевую гипотезу примем утверждение о равенстве средних значений:

H0: \mu\,1=\mu\,2=\mu\,3

Альтернативная гипотеза следующая: не все средние равны, то есть имеются, по крайней мере, две группы, отличающиеся средними значениями:

HА: \mu\,1≠\mu\,2≠\mu\,3

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ANOVA** | | | | | |
| Заболеваемость на 1000 человек населения | | | | | |
|  | Сумма квадратов | ст.св. | Средний квадрат | F | Знач. |
| Между группами | 185301,010 | 2 | 92650,505 | 3,835 | ,026 |
| Внутри групп | 1981156,795 | 82 | 24160,449 |  |  |
| Всего | 2166457,804 | 84 |  |  |  |

Дисперсионный анализ дает значимые результаты (p < 0,05). Чем больше значение F, тем больше различаются средние значения между группами (F = 3,835). Мы можем заключить, что средняя заболеваемость на 1000 человек населения значительно отлична хотя бы в одной группе, обозначающей уровень преступности.

# 3. Спецификация базовой модели.

## 3.1. Связь структуры базовой модели с основными гипотезами исследования.

Гипотеза 1: Чем больше жилья в аварийном состоянии в регионе, тем выше показатель заболеваемости.

Данная гипотеза предполагает положительную функциональную связь между независимой переменной X3 (численность населения с денежными доходами ниже прожиточного минимума) и зависимой переменной Y (заболеваемость на 1000 человек населения), что предполагает положительное значение коэффициента корреляции и положительное значение оценки коэффициента при переменной X3.

Однако прежде, чем проверить связь переменных, следует проверить гипотезу о значимости найденных оценок коэффициентов.

Гипотеза 2: Подверженность заболеваниям населения с низким уровнем дохода возрастает с уменьшением объема использования бытовых услуг.

Данная гипотеза подразумевает положительную связь между независимой переменной X1 (численность населения с денежными доходами ниже прожиточного минимума) и зависимой переменной Y(заболеваемость на 1000 человек населения), что предполагает положительное значение коэффициента корреляции. Кроме того, выше указанные переменные должны быть отрицательно связаны с переменной X4 (объем бытовых услуг на душу населения), что предполагает отрицательное значение соответствующих коэффициентов корреляции.

Однако прежде, чем проверить связь переменных, следует проверить гипотезу о значимости найденных оценок коэффициентов.

Гипотеза 3: Чем качественнее медицинское обслуживание в регионе, там меньше уровень заболеваемости.

Данная гипотеза предполагает отрицательную функциональную связь между независимой переменной X2 (численность среднего медицинского персонала на 10 000 человек населения) и зависимой переменной Y (заболеваемость на 1000 человек населения), что предполагает отрицательное значение коэффициента корреляции и отрицательное значение оценки коэффициента при переменной X2.

Однако прежде, чем проверить связь переменных, следует проверить гипотезу о значимости найденных оценок коэффициентов.

## 3.2. Связь структуры базовой модели с результатами предварительного анализа данных.

### 1.2.1. Состав переменных

Количественная зависимая переменная – заболеваемость на 1000 человек населения.

Количественные независимые переменные:

* Численность населения с денежными доходами ниже прожиточного минимума (в процентах от общей численности населения субъекта),
* Численность среднего медицинского персонала на 10 000 человек населения (человек),
* Удельный вес ветхого и аварийного жилищного фонда в общей площади всего жилищного фонда (в процентах),
* Объем бытовых услуг на душу населения (рублей).

Качественная независимая переменная – уровень преступности в регионах. Все субъекты РФ были изначально разделены на 3 группы по уровню преступности: 1-наиболее высокий, 2 – средний, 3 – наиболее низкий. Далее предварительный анализ переменных показал целесообразность создания лишь двух групп, о чем будет указано в подразделе 1.2.3. Анализ пары «числовая зависимая переменная – качественная независимая переменная» с помощью диаграммы Бокса-Уискера (Box-Whisker).

Регрессионная модель в общем виде:

Y = β0 + β1 X1 + β2 X2 + β3 X3 + β4 X4 + dX5, где

Y – заболеваемость на 1000 человек населения;

X1 – численность населения с денежными доходами ниже прожиточного минимума;

X2 – численность среднего медицинского персонала на 10 000 человек населения;

X3 – удельный вес ветхого и аварийного жилищного фонда в общей площади всего жилищного фонда;

X4 – объем бытовых услуг на душу населения;

X5 – уровень преступности в регионах.

Регрессионная модель при условии линейного вхождения всех переменных в модель:

Y = 769,861 – 10,493 X1 + 2,391 X2 + 9,719 X3 – 0,02X4 – 83,278X5

Отметим, что в данном случае все коэффициенты значимы с вероятностью ошибки 5%.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Коэффициентыa** | | | | | | |
| Модель | | Нестандартизованные коэффициенты | | Стандартизованные коэффициенты | т | Знач. |
| B | Стандартная Ошибка | Бета |
| 1 | (Константа) | 769,861 | 142,096 |  | 5,418 | ,000 |
| Численность населения с денежными доходами ниже прожиточного минимума (в процентах от общей численности населения субъекта) | -10,493 | 4,012 | -,318 | -2,615 | ,011 |
| Численность среднего медицинского персонала на 10 000 человек населения, человек | 2,391 | ,978 | ,251 | 2,445 | ,017 |
| Удельный вес ветхого и аварийного жилищного фонда в общей площади всего жилищного фонда (в процентах) | 9,719 | 4,468 | ,229 | 2,175 | ,033 |
| Уровень преступности в регионах | -83,278 | 34,016 | -,245 | -2,448 | ,017 |
| Объем бытовых услуг на душу населения (рублей) | -,020 | ,010 | -,241 | -1,973 | ,052 |
| a. Зависимая переменная: Заболеваемость на 1000 человек населения | | | | | | |

Обратим внимание на величину случайной составляющей (769,861). Как известно, случайная составляющая отражает, во-первых, влияние ошибок измерений (чем больше ошибок, тем больше значение случайной составляющей), во-вторых, влияние всех неучтенных факторов. В настоящей модели ошибки, безусловно, возможны, но они могут касаться только недостаточного количества данных, потому как, например, пропуски в данных отсутствуют.

В нашем случае значительное влияние оказывают неучтенные факторы. Изученные нами работы по схожей тематике отмечают большое количество независимых переменных, касающихся различных сфер жизни человека, например, культурного развития регионов, их инфраструктуры и многого другого.

В разделе 4 будут приведены и сравнены результаты оптимизации базовой модели на основе модифицированного коэффициента детерминации и информационных критериев (AIC, BIC).

# 4. Оценивание базовой модели.

## 4.1. Анализ наличия мультиколлинеарности

С целью предварительного анализа взаимосвязи показателей построена матрица R — таблица парных коэффициентов корреляции.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Корреляции** | | | | | | |
|  | | | Численность населения с денежными доходами ниже прожиточного минимума (в процентах от общей численности населения субъекта) | Численность среднего медицинского персонала на 10 000 человек населения, человек | Удельный вес ветхого и аварийного жилищного фонда в общей площади всего жилищного фонда (в процентах) | Объем бытовых услуг на душу населения (рублей) |
|  | Численность населения с денежными доходами ниже прожиточного минимума (в процентах от общей численности населения субъекта) | Коэффициент корреляции | 1,000 | ,125 | ,301\*\* | -,519\*\* |
| Знач. (2-х сторонняя) | . | ,254 | ,005 | ,000 |
| N | 85 | 85 | 85 | 85 |
| Численность среднего медицинского персонала на 10 000 человек населения, человек | Коэффициент корреляции | ,125 | 1,000 | ,442\*\* | -,092 |
| Знач. (2-х сторонняя) | ,254 | . | ,000 | ,405 |
| N | 85 | 85 | 85 | 85 |
| Удельный вес ветхого и аварийного жилищного фонда в общей площади всего жилищного фонда (в процентах) | Коэффициент корреляции | ,301\*\* | ,442\*\* | 1,000 | -,310\*\* |
| Знач. (2-х сторонняя) | ,005 | ,000 | . | ,004 |
| N | 85 | 85 | 85 | 85 |
| Объем бытовых услуг на душу населения (рублей) | Коэффициент корреляции | -,519\*\* | -,092 | -,310\*\* | 1,000 |
| Знач. (2-х сторонняя) | ,000 | ,405 | ,004 | . |
| N | 85 | 85 | 85 | 85 |
| \*\*. Корреляция значима на уровне 0,01 (двухсторонняя). | | | | | | |

Анализ матрицы парных коэффициентов корреляции показывает, что результативный признак наиболее тесно связан с показателем х3 — удельным весом ветхого и аварийного жилищного фонда в общей площади всего жилищного фонда (ryx3 = 0,301).

Между аргументами нет функциональной связи, мультиколлинеарность также отсутствует, о чем свидетельствуют низкие коэффициенты корреляции (например, rx1x2 = 0,125 и rx3x2 = 0,442).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Coefficientsa** | | | | | | | | |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Collinearity Statistics | |
| B | Std. Error | Beta | Tolerance | VIF |
| 1 | (Constant) | 653,130 | 137,973 |  | 4,734 | ,000 |  |  |
| Численность населения с денежными доходами ниже прожиточного минимума (в процентах от общей численности населения субъекта) | -9,992 | 4,130 | -,303 | -2,419 | ,018 | ,625 | 1,601 |
| Численность среднего медицинского персонала на 10 000 человек населения, человек | 2,970 | ,978 | ,312 | 3,038 | ,003 | ,927 | 1,079 |
| Удельный вес ветхого и аварийного жилищного фонда в общей площади всего жилищного фонда (в процентах) | 9,857 | 4,605 | ,232 | 2,141 | ,035 | ,831 | 1,203 |
| Объем бытовых услуг на душу населения (рублей) | -,017 | ,010 | -,199 | -1,594 | ,115 | ,631 | 1,585 |
| a. Dependent Variable: Заболеваемость на 1000 человек населения | | | | | | | | |

Данное предположение доказывают и показатели VIF, каждый из которых менее 10, что свидетельствует об отсутствии мультиколлинеарности. Коэффициенты толерантности превышают 0,1, что также говорит об отсутствии эффекта мультиколлинеарности независимых переменных.

## 4.2. Анализ наличия гетероскедастичности

### 4.2.1. Тест Парка

Для выявления гетероскедастичности проведем тест Парка[[2]](#footnote-2). Ниже кратко приведен алгоритм:

* ⌃Y = ⌃β0 + ⌃β1 X1 + ⌃β2 X2 + ⌃β3 X3 + ⌃β4 X4  – ⌃dX5

⌃Y = 769,861 – 10,493 X1 + 2,391 X2 + 9,719 X3 – 0,02X4 – 83,278X5

* Yi = ⌃Yi + ei

ei = Yi - ⌃Yi

На наличие гетероскедастичности подозреваются две переменные (на основании Диаграммы рассеивания – Scatter plot): Z1 = X2, Z2 = X4.

* ln(ei)2 = ⌃ ln(Zi) + vi
* H0 :⌃ = 0 (гомоскедастичность)

H1 :⌃ ≠ 0 (гетероскедастичность)

Однако, несмотря на то, что гетероскедастичность предполагается только у двух переменных, были проверены все 4 количественные независимые переменные.

1. **Численность населения с денежными доходами ниже прожиточного минимума**

* p-value = 0,997
* α = 0,05
* 0,997 > 0,05
* Нет оснований отвергать H0  (нет гетероскедастичности)

1. **Численность среднего медицинского персонала на 10 000 человек населения**

* p-value = 0,047
* α = 0,05
* 0,47 < 0,005
* Нет оснований принять H0 (есть гетероскедастичность)

1. **Удельный вес ветхого и аварийного жилищного фонда в общей площади всего жилищного фонда**

* p-value = 0,824
* α = 0,05
* 0,824 > 0,05
* Нет оснований отвергать H0  (нет гетероскедастичности)

1. **Объем бытовых услуг на душу населения**

* P-value = 0,7
* α = 0,05
* 0,7 > 0,05
* Нет оснований отвергать H0  (нет гетероскедастичности)

Итак, проведенный тест подтвердил наличие гетероскедастичности у переменной X2. Однако тест Парка следует дополнить и другими тестами.

### 2.2.3. Тест Глейзера

Проведем тест Глейзера.[[3]](#footnote-3)

В таблице ниже приведены значения p-value, превышающие значение 0,05, соответственно, нет оснований отвергать основную гипотезу об отсутствии гетероскедастичности.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Коэффициентыa** | | | | | | |
| Модель | | Нестандартизованные коэффициенты | | Стандартизованные коэффициенты | т | Знач. |
| B | Стандартная Ошибка | Бета |
| 1 | (Константа) | ,486 | 86,516 |  | ,006 | ,996 |
| Численность населения с денежными доходами ниже прожиточного минимума (в процентах от общей численности населения субъекта) | -,468 | 2,443 | -,027 | -,192 | ,849 |
| Численность среднего медицинского персонала на 10 000 человек населения, человек | ,950 | ,595 | ,188 | 1,597 | ,114 |
| Удельный вес ветхого и аварийного жилищного фонда в общей площади всего жилищного фонда (в процентах) | ,699 | 2,720 | ,031 | ,257 | ,798 |
| Уровень преступности в регионах | -2,026 | 20,711 | -,011 | -,098 | ,922 |
| Объем бытовых услуг на душу населения (рублей) | ,001 | ,006 | ,019 | ,136 | ,892 |
| a. Зависимая переменная: AbsUt | | | | | | |

### 4.2.2. Тест Голдфелда-Квандта

Для подтверждения или опровержения наличия гетероскедастичности проведем тест Голдфелда-Квандта[[4]](#footnote-4) (Goldfeld-Quandt test) только для переменной X2. Ниже кратко приведен алгоритм:

* Данные упорядочиваются по возрастанию переменной Z = X2.
* Выборка разбивается на три группы:

1 группа: (n-m)/2 первых наблюдений,

2 группа: (n-m)/2 последних наблюдений,

3 группа: m наблюдений.

m = n/4

* Строятся две регрессии для первой и второй группы.
* Рассчитывают F-статистика, которая равна отношению большей суммы квадратов остатков к меньшей.

Была рассчитана F-статистика: F = 0,098. Таким образом, на уровне значимости 5% принимается основания гипотеза, то есть гетероскедастичность не обнаружена.

## 4.3. Проверка гипотез

Гипотеза 1: Чем больше жилья в аварийном состоянии в регионе, тем выше показатель заболеваемости.

Сначала проверим гипотезу о значимости коэффициента β3.

В уже приведенной ранее таблице коэффициентов были выделены значения p-value. Для оценки коэффициента β3 p-value = 0,033. 0,033 < 0,05, следовательно, нет оснований принять основную гипотезу с вероятностью ошибки 5%. Основная гипотеза отвергается в пользу альтернативной – коэффициент β3 значим.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Корреляции** | | | | |
|  | | | Заболеваемость на 1000 человек населения | Удельный вес ветхого и аварийного жилищного фонда в общей площади всего жилищного фонда (в процентах) |
|  | Заболеваемость на 1000 человек населения | Коэффициент корреляции | 1,000 | ,356\*\* |
| Знач. (односторонняя) | . | ,000 |
| N | 85 | 85 |
| Удельный вес ветхого и аварийного жилищного фонда в общей площади всего жилищного фонда (в процентах) | Коэффициент корреляции | ,356\*\* | 1,000 |
| Знач. (односторонняя) | ,000 | . |
| N | 85 | 85 |
| \*\*. Корреляция значима на уровне 0,01 (односторонняя). | | | | |

Полученный коэффициент корреляции – положительный, что, соответственно, подразумевает положительную связь переменных: чем выше значение одной переменной, тем выше значение другой. Само значение коэффициента корреляции (0,356) говорит о слабой коррелированности переменных Х3 и Y.

Полученное значение оценки коэффициента β2 = 9,719 с вероятностью ошибки 5% не дает оснований отвергнуть основную гипотезу (p-value = 0,033; 0,033 < 0,05).

Таким образом, можно сделать следующий вывод: чем больше удельный вес ветхого и аварийного жилищного фонда в общей площади всего жилищного фонда в регионе, тем выше уровень заболеваемости. Можно предположить, что улучшение качества жилья населения регионов (в случае нашей модели: уменьшение доли ветхого и аварийного жилья) может способствовать уменьшению уровня заболеваемости. Нет оснований отвергнуть основную гипотезу.

Гипотеза 2: Подверженность заболеваниям населения с низким уровнем дохода возрастает с уменьшением объема использования бытовых услуг.

Сначала проверим гипотезу о значимости коэффициента β1 и β4.

В уже приведенной ранее таблице коэффициентов были выделены значения p-value. Для оценок коэффициентов β1 и β4 p-value равно 0,011 и 0,052 (приблизительно 0,05), соответственно. 0,011 < 0,05, 0,05 = 0,05, следовательно, нет оснований принять основную гипотезу и в первом, и во втором случае с вероятностью ошибки 5%. Основная гипотеза отвергается в пользу альтернативной – коэффициенты β1 и β4 значимы.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Корреляции** | | | | | |
|  | | | Заболеваемость на 1000 человек населения | Численность населения с денежными доходами ниже прожиточного минимума (в процентах от общей численности населения субъекта) | Объем бытовых услуг на душу населения (рублей) |
|  | Заболеваемость на 1000 человек населения | Коэффициент корреляции | 1,000 | ,057 | -,106 |
| Знач. (односторонняя) | . | ,301 | ,167 |
| N | 85 | 85 | 85 |
| Численность населения с денежными доходами ниже прожиточного минимума (в процентах от общей численности населения субъекта) | Коэффициент корреляции | ,057 | 1,000 | -,519\*\* |
| Знач. (односторонняя) | ,301 | . | ,000 |
| N | 85 | 85 | 85 |
| Объем бытовых услуг на душу населения (рублей) | Коэффициент корреляции | -,106 | -,519\*\* | 1,000 |
| Знач. (односторонняя) | ,167 | ,000 | . |
| N | 85 | 85 | 85 |
| \*\*. Корреляция значима на уровне 0,01 (односторонняя). | | | | | |

Рассмотрим полученные коэффициенты корреляции. Так, из данной таблицы видно, что показатели заболеваемости и численности населения с денежными доходами ниже прожиточного минимума имеют положительную связь (коэффициент корреляции положителен и равен 0,057). Показатель объема бытовых услуг на душу населения связан отрицательно как с показателем заболеваемости, так и с показателем численности населения с денежными доходами ниже прожиточного минимума (коэффициенты корреляции равны -0,106 и -0,519, соответственно).

Таким образом, можно предположить, что при увеличении уровня заболеваемости и населения с денежными доходами ниже прожиточного минимума, объем бытовых услуг уменьшается. Нет оснований отвергнуть основную гипотезу.

Гипотеза 3: Чем качественнее медицинское обслуживание в регионе, там меньше уровень заболеваемости.

Сначала проверим гипотезу о значимости коэффициента β4.

В уже приведенной ранее таблице коэффициентов были выделены значения p-value. Для оценки коэффициента β2 p-value = 0,017. 0,017 < 0,05, следовательно, нет оснований принять основную гипотезу с вероятностью ошибки 5%. Основная гипотеза отвергается в пользу альтернативной – коэффициент β2 значим.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Корреляции** | | | | |
|  | | | Заболеваемость на 1000 человек населения | Численность среднего медицинского персонала на 10 000 человек населения, человек |
| Ро Спирмана | Заболеваемость на 1000 человек населения | Коэффициент корреляции | 1,000 | ,346\*\* |
| Знач. (односторонняя) | . | ,001 |
| N | 85 | 85 |
| Численность среднего медицинского персонала на 10 000 человек населения, человек | Коэффициент корреляции | ,346\*\* | 1,000 |
| Знач. (односторонняя) | ,001 | . |
| N | 85 | 85 |
| \*\*. Корреляция значима на уровне 0,01 (односторонняя). | | | | |

Полученный коэффициент корреляции – положительный, что, соответственно, подразумевает положительную связь переменных: чем выше значение одной переменной, тем выше значение другой. Само значение коэффициента корреляции (0,346) говорит о слабой коррелированности переменных Х2 и Y.

Полученное значение оценки коэффициента β2 = 2,393 с вероятностью ошибки 5% не дает оснований принять основную гипотезу (p-value = 0,017; 0,017 < 0,05).

Таким образом, можно сделать следующий вывод: повышение качества медицинского обслуживания, в нашей модели выраженного в численности среднего медицинского персонала на 10 000 человек населения, не влияет на снижение заболеваемости в регионах. Данный вывод может быть объяснен тем, что для обозначения качества медицинского обслуживания выбран некорректный показатель.

Используя показатель, представленный в настоящей модели, можно предположить, что наращивание медицинского персонала не принесет пользы в борьбе с заболеваемостью населения. Нет оснований принимать основную гипотезу.

# 5. Оптимизация базовой модели

В таблице ниже приведены значения оценок коэффициентов для регрессионной модели, в которой все независимые переменные входят линейно (Модель 1). Как уже указывалось ранее, все коэффициенты значимы с вероятностью ошибки 5%.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Коэффициентыa** | | | | | | |
| Модель | | Нестандартизованные коэффициенты | | Стандартизованные коэффициенты | т | Знач. |
| B | Стандартная Ошибка | Бета |
| 1 | (Константа) | 769,861 | 142,096 |  | 5,418 | ,000 |
| Численность населения с денежными доходами ниже прожиточного минимума (в процентах от общей численности населения субъекта) | -10,493 | 4,012 | -,318 | -2,615 | ,011 |
| Численность среднего медицинского персонала на 10 000 человек населения, человек | 2,391 | ,978 | ,251 | 2,445 | ,017 |
| Удельный вес ветхого и аварийного жилищного фонда в общей площади всего жилищного фонда (в процентах) | 9,719 | 4,468 | ,229 | 2,175 | ,033 |
| Уровень преступности в регионах | -83,278 | 34,016 | -,245 | -2,448 | ,017 |
| Объем бытовых услуг на душу населения (рублей) | -,020 | ,010 | -,241 | -1,973 | ,052 |
| a. Зависимая переменная: Заболеваемость на 1000 человек населения | | | | | | |

В следующей таблице приведены и выделены значения модифицированного коэффициента детерминации и информационных критериев (AIC, BIC).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Сводка для модели** | | | | | | | | |
| Модель | R | R-квадрат | Скорректированный R-квадрат | Стандартная ошибка оценки | Критерии выбора | | | |
| Информационный критерий Акаике | Критерий предсказания Амемия | Критерий прогноза Мэллоуза | Байесовский критерий Шварца |
| 1 | ,521a | ,272 | ,226 | 141,3268 | 847,460 | ,839 | 6,000 | 862,116 |
| a. Предикторы: (константа), Объем бытовых услуг на душу населения (рублей), Численность среднего медицинского персонала на 10 000 человек населения, человек, Уровень преступности в регионах, Удельный вес ветхого и аварийного жилищного фонда в общей площади всего жилищного фонда (в процентах), Численность населения с денежными доходами ниже прожиточного минимума (в процентах от общей численности населения субъекта) | | | | | | | | |

В таблице ниже приведены значения оценок коэффициентов для регрессионной модели, в которой все независимые переменные входят линейно, кроме переменной X1, для которой была применена функциональная форма (Модель 2). Все коэффициенты значимы с вероятностью ошибки 5%, кроме коэффициента при переменной «Объем бытовых услуг на душу населения (рублей)» (коэффициент значим с вероятностью ошибки 10%). Отметим, что при изменении функциональной формы переменной «Численность среднего медицинского персонала на 10 000 человек населения, человек», другая переменная потеряла свою значимость.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Коэффициентыa** | | | | | | |
| Модель | | Нестандартизованные коэффициенты | | Стандартизованные коэффициенты | т | Знач. |
| B | Стандартная Ошибка | Бета |
| 1 | (Константа) | 503,775 | 128,211 |  | 3,929 | ,000 |
| Численность среднего медицинского персонала на 10 000 человек населения, человек | 2,207 | ,986 | ,232 | 2,238 | ,028 |
| Удельный вес ветхого и аварийного жилищного фонда в общей площади всего жилищного фонда (в процентах) | 8,646 | 4,476 | ,204 | 1,931 | ,057 |
| Уровень преступности в регионах | -95,080 | 35,112 | -,280 | -2,708 | ,008 |
| Объем бытовых услуг на душу населения (рублей) | -,020 | ,011 | -,235 | -1,835 | ,070 |
| Численность населения с денежными доходами ниже прожиточного минимума (в процентах от общей численности населения субъекта)(2) | 1844,866 | 825,605 | ,280 | 2,235 | ,028 |
| a. Зависимая переменная: Заболеваемость на 1000 человек населения | | | | | | |

В следующей таблице приведены и выделены значения модифицированного коэффициента детерминации и информационных критериев (AIC, BIC).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Сводка для модели** | | | | | | | | |
| Модель | R | R-квадрат | Скорректированный R-квадрат | Стандартная ошибка оценки | Критерии выбора | | | |
| Информационный критерий Акаике | Критерий предсказания Амемия | Критерий прогноза Мэллоуза | Байесовский критерий Шварца |
| 1 | ,506a | ,256 | ,209 | 142,8718 | 849,309 | ,857 | 6,000 | 863,965 |
| a. Предикторы: (константа), Численность населения с денежными доходами ниже прожиточного минимума (в процентах от общей численности населения субъекта)(2), Численность среднего медицинского персонала на 10 000 человек населения, человек, Уровень преступности в регионах, Удельный вес ветхого и аварийного жилищного фонда в общей площади всего жилищного фонда (в процентах), Объем бытовых услуг на душу населения (рублей) | | | | | | | | |

Анализ приведенных таблиц показал, что модель, в которой все независимые переменные входят линейно (Модель 1), лучше. (Скорректированный R-квадрат: 0,226 > 0,209 849,309; AIC: 847,460 < 849,309; BIC: 862,116 < 863,965)

Далее будет проведена проверка на мультиколлинеарность и гетероскедастичность Модели 2.

## 5.1. Анализ наличия мультиколлиниарности

Показатели VIF, каждый из которых менее 10, свидетельствует об отсутствии мультиколлинеарности. Коэффициенты толерантности превышают 0,1, что также говорит об отсутствии эффекта мультиколлинеарности независимых переменных.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Коэффициентыa** | | | | | | | | |
| Модель | | Нестандартизованные коэффициенты | | Стандартизованные коэффициенты | т | Знач. | Статистика коллинеарности | |
| B | Стандартная Ошибка | Бета | Допуск | VIF |
| 1 | (Константа) | 503,775 | 128,211 |  | 3,929 | ,000 |  |  |
| Численность населения с денежными доходами ниже прожиточного минимума (в процентах от общей численности населения субъекта)(2) | 1844,866 | 825,605 | ,280 | 2,235 | ,028 | ,600 | 1,666 |
| Численность среднего медицинского персонала на 10 000 человек населения, человек | 2,207 | ,986 | ,232 | 2,238 | ,028 | ,876 | 1,142 |
| Удельный вес ветхого и аварийного жилищного фонда в общей площади всего жилищного фонда (в процентах) | 8,646 | 4,476 | ,204 | 1,931 | ,057 | ,846 | 1,182 |
| Уровень преступности в регионах | -95,080 | 35,112 | -,280 | -2,708 | ,008 | ,882 | 1,134 |
| Объем бытовых услуг на душу населения (рублей) | -,020 | ,011 | -,235 | -1,835 | ,070 | ,574 | 1,742 |
| a. Зависимая переменная: Заболеваемость на 1000 человек населения | | | | | | | | |

## 

## 5.2. Анализ наличия гетероскедастичности

Тест Парка (проверка наличия гетероскедастичности X1)

* p-value = 0,334
* α = 0,05
* 0,334 > 0,05
* Нет оснований отвергать H0 (нет гетероскедастичности)

Тест Глейзера

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Коэффициентыa** | | | | | | |
| Модель | | Нестандартизованные коэффициенты | | Стандартизованные коэффициенты | т | Знач. |
| B | Стандартная Ошибка | Бета |
| 1 | (Константа) | -13,149 | 77,205 |  | -,170 | ,865 |
| Численность среднего медицинского персонала на 10 000 человек населения, человек | ,942 | ,594 | ,187 | 1,585 | ,117 |
| Удельный вес ветхого и аварийного жилищного фонда в общей площади всего жилищного фонда (в процентах) | ,667 | 2,696 | ,030 | ,247 | ,805 |
| Уровень преступности в регионах | -2,929 | 21,144 | -,016 | -,139 | ,890 |
| Объем бытовых услуг на душу населения (рублей) | ,001 | ,006 | ,012 | ,084 | ,933 |
| Численность населения с денежными доходами ниже прожиточного минимума (в процентах от общей численности населения субъекта)(2) | 124,865 | 497,155 | ,036 | ,251 | ,802 |
| a. Зависимая переменная: AbsUt | | | | | | |

Так как значение p-value для всех оценок коэффициентов превышает 0,05, нет оснований отвергать основную гипотезу (отсутствие гетероскедастичности) с вероятностью ошибки 5%.

## 5.3. Проверка гипотез

Гипотеза 1: Чем больше жилья в аварийном состоянии в регионе, тем выше показатель заболеваемости.

Сначала проверим гипотезу о значимости коэффициента β3.

В уже приведенной ранее таблице коэффициентов были выделены значения p-value. Для оценки коэффициента β3 p-value = 0,057. 0,057 < 0,05, следовательно, нет оснований отвергать основную гипотезу с вероятностью ошибки 5%. Основная гипотеза принимается – коэффициент β3 в данной модели значим только с вероятностью ошибки 10%, поэтому дальнейший анализ корреляции не производится.

Гипотеза 2: Подверженность заболеваниям населения с низким уровнем дохода возрастает с уменьшением объема использования бытовых услуг.

Сначала проверим гипотезу о значимости коэффициента β1 и β4.

В уже приведенной ранее таблице коэффициентов были выделены значения p-value. Для оценок коэффициентов β1 и β4 p-value равно 0,021 и 0,07, соответственно. 0,021 < 0,05, 0,07 > 0,05 следовательно, нет оснований принять основную гипотезу в первом случае и отвергнуть во втором случае с вероятностью ошибки 5%. Основная гипотеза отвергается в пользу альтернативной – коэффициент β1 значим. Основная гипотеза принимается – коэффициент β4 не значим. Дальнейший анализ корреляции не производится.

Гипотеза 3: Чем качественнее медицинское обслуживание в регионе, там меньше уровень заболеваемости.

Сначала проверим гипотезу о значимости коэффициента β4.

В уже приведенной ранее таблице коэффициентов были выделены значения p-value. Для оценки коэффициента β2 p-value = 0,028. 0,028 < 0,05, следовательно, нет оснований принять основную гипотезу с вероятностью ошибки 5%. Основная гипотеза отвергается в пользу альтернативной – коэффициент β2 значим.

Так как переменная X3 не меняла свою функциональную форму, последующий анализ аналогичен приведенному анализу в разделе 2.3. Проверка гипотез.

Таким образом, на данном этапе анализа регрессионной модели все предварительно заявленные независимые переменные входят в модель линейно: Y = 769,861 – 10,493 X1 + 2,391 X2 + 9,719 X3 – 0,02X4 – 83,278X5

# 6. Доверительные интервалы

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Коэффициентыa** | | | | | | | | |
| Модель | | Нестандартизованные коэффициенты | | Стандартизованные коэффициенты | т | Знач. | 95,0% Доверительный интервал для B | |
| B | Стандартная Ошибка | Бета | Нижняя граница | Верхняя граница |
| 1 | (Константа) | 653,130 | 137,973 |  | 4,734 | ,000 | 378,555 | 927,705 |
| Численность населения с денежными доходами ниже прожиточного минимума (в процентах от общей численности населения субъекта) | -9,992 | 4,130 | -,303 | -2,419 | ,018 | -18,211 | -1,773 |
| Численность среднего медицинского персонала на 10 000 человек населения, человек | 2,970 | ,978 | ,312 | 3,038 | ,003 | 1,024 | 4,916 |
| Удельный вес ветхого и аварийного жилищного фонда в общей площади всего жилищного фонда (в процентах) | 9,857 | 4,605 | ,232 | 2,141 | ,035 | ,693 | 19,021 |
| Объем бытовых услуг на душу населения (рублей) | -,017 | ,010 | -,199 | -1,594 | ,115 | -,038 | ,004 |
| a. Зависимая переменная: Заболеваемость на 1000 человек населения | | | | | | | | |

* Численность населения с денежными доходами ниже прожиточного минимума [-18,221; -1,773]. Довольно широкий интервал указывает на неточность оценки. В данный интервал входит ноль – ближе к правому краю.
* Численность среднего медицинского персонала на 10 000 человек населения [1,024; 4,916]. Узкий интервал указывает на точность оценки.
* Удельный вес ветхого и аварийного жилищного фонда в общей площади всего жил. фонда [0,693; 19,021]. Довольно широкий интервал указывает на неточность оценки. В данный интервал ноль не входит.
* Объем бытовых услуг на душу населения [-0,038; 0,004]. Узкий интервал указывает на точность оценки. В данный интервал входит ноль – ближе к центру.

# 7. Обнаружение выбросов и влиятельных наблюдений в регрессионном анализе

Проведенные ниже тесты позволили определить степень влияния показателей на регрессионную зависимость. Поскольку влиятельные наблюдения в большей мере определяют регрессионную зависимость, даже небольшие ошибки в их оценках могут привести к существенному искажению регрессии, что отрицательно скажется на ее прогностических свойствах.

Изученные ниже расстояния также позволят подтвердить или опровергнуть гипотезу о наличии выбросов в регрессионной модели, которые были выявлены на первом этапе анализа.

## 7.1. Расстояние Махаланобиса

Расстояние Махаланобиса представляет собой расстояние от точки до центра системы, отнесенное к диаметру облака в этом направлении. Рассчитаем критическое значение *MD* при уровне значимости 5% (α=0,05):

После нахождения меры удаления наблюдения от центра системы в многопеременном анализе и сравнив каждый из показателей с критическим значением, приходим к выводу, что по величине расстояния Махаланобиса 5 значений на 5%-ном уровне значимости являются аномально удалёнными от центра системы, т.к. их *MDi* <*MD*[α=0,05]. Таким образом, нижеприведенные показатели можем считать подозрительными: 26, 32, 33, 37, 38, 62, 67, 77, 83.

## 7.2. Расстояние Кука

Расстояние Кука является обобщенной мерой, которая учитывает, как внутренний стьюдентизированный остаток, так и показатель влияния наблюдения. В нашей модели критическое значение для расстояния Кука находится как медиана *F*-распределения с числом степеней свободы  и *=n-k=81*, где *k* – количество параметров в модели. Критическое значение статистики Кука составляет 0.84624.

Таким образом, ни одно значение не может считаться значимо выделяющимся своим влиянием на регрессионную зависимость, поскольку ни одно из наблюдений не приближено к критическому значению.

## 7.3. Ковариационное отношение

Ковариационное отношение представляет собой отношение детерминанта ковариационной матрицы с удалённым *i*-тым наблюдением к детерминанту ковариационной матрицы для всего набора данных. Влиятельными считаются показатели, значимо отличающиеся от 1. Если , то наблюдение считается существенно влияющим на регрессионную зависимость.

В нашем случае влиятельными являются 42 из 85 наблюдений, в их число входят наблюдения № 26, 32, 33, 37, 38, 62, 67, 77, 83, как видно из таблицы с полученными результатами. Удаление любого из них исказит регрессионную зависимость и расширит доверительный интервал для предсказания. Поэтому нами было принято решение не исключать данные наблюдения из модели.

## 7.4. Мера стандартизованная *DFFIT* (*DFFITS*)

Как и расстояние Кука, *DFFITS* объединяет внешний стьюдентизированный остаток и показатель влияниянаблюдения. Для малых и средних объёмов выборок влиятельными наблюдениями признаются значения более 1.

То есть, согласно данному показателю, в нашей выборке ни одно наблюдение не является влиятельным.

## 7.5. Мера стандартизованная *DFBETA* (*DFBETAS*)

DFBETAS оценивают степень изменения отдельных параметров регрессионной модели при исключении из *i*-того наблюдения. Как и в случае *DFFITS* для малых и средних объёмов выборок влиятельными наблюдениями признаются *DFBETAS* со значениями более 1.

Таким образом, в нашей выборке ни один из показателей существенно не влияет на наклон линии регрессии.

В проведенных ниже тестах мы обнаружили возможные выбросы и определили влияющие наблюдения в предположении линейной зависимости Y. Однако удаление наблюдений №26, 32, 33, 37, 38, 62, 67, 77, 83 приведет к противоречию. Таким образом, нами было принято решение не исключать из списка наблюдений выявленные значимо выделяющиеся показатели.

# 8. Эссе по найденным публикациям.

## 8.1. Эссе по статье Б.Б. Прохорова, И.В. Горшковой, Е.В. Тарасовой «Условия жизни населения и общественное здоровье»

По мнению авторов статьи, интегральным выражением качества здоровья населения является ожидаемая продолжительность жизни (далее ОПЖ), что и является зависимой переменной в работе. Она в свою очередь подчиняется влиянию многих социально-экономических факторов: природных, эколого-гигиенических, производственных, но в первую очередь от социально-экономических. Авторы попытались сформировать верное представление о качестве общественного здоровья с помощью изучения корреляции между ОПЖ и различными социально-экономическими показателями.

Анализ проводился на рубеже начала 2000-х, что не может не влиять на результаты исследования. Выход страны из состояния глубоко экономического спада (1991-1999), а также переход к рыночной экономике сопровождался снижением уровня и качества жизни населения, усилением межрегиональной дифференциации. Именно она, по мнению авторов, усугубляет влияние социально-экономических факторов на уровень здоровья населения. «Прекратилось продолжавшееся более полутора столетий движение производства и капитала «вширь» и потенциально предполагавшее преодоление межрайонных диспропорций и освоение в конечном счете всей доступной территории. Теперь капитал и труд стекаются в крупные финансовые центры… Результатом этих перекосов в экономической и региональной политике стало усиление межрайонной дифференциации» (Любовный, Пчелинцев, 2006, с. 10-14).

В своей работе мы основывались на предыдущий опыт исследования автора, рассмотрели изученные им взаимосвязи и интерпретировали некоторые из них в соответствии с нынешней экономико-политической ситуацией в стране. Авторы исследовали взаимосвязи, дав им свое толкование с учетом произошедших накануне исторических событий, повлиявших на развитие страны и уровень ОПЖ. Кроме того, авторы отметили наиболее яркое воздействие комплекса негативных социальных характеристик региона на уровень продолжительности жизни. Мы проводим наше исследование спустя 15 лет, но тем не менее, наша страна также претерпевает изменения. С учетом нынешней экономической нестабильности в структуре управления и финансовой политике, мы решили исследовать влияние негативных факторов, таких как “Численность населения с денежными доходами ниже величины прожиточного минимума” и “Удельный вес ветхого и аварийного жилищного фонда в общей площади всего жилищного фонда” на уровень заболеваемости населения в том числе.

Авторы проследили прямую зависимость между ОПЖ и уровнем дифференциации доходов, жилищными условиями, уровнем преступности, дискомфортностью природных условий и социально значимыми заболеваниями. Он убежден, что “недостаточное внимание к этим проблемам со стороны руководства страны, регионов и муниципальных образований приводит к снижению качества общественного здоровья”, что является важной стратегической задачей государства.

## 8.2. Эссе по статье Б.Б. Прохорова «Социально-экономические особенности федеральных округов России и здоровье населения»

Б.Б. Прохоров подчеркивает важность здоровья населения в рамках государства: «Масштабы заботы государства о здоровье своих граждан, реальные усилия и достижения в этой области можно рассматривать как мерило социально-этической зрелости общества, уровня его гуманизма, совершенства всей государственной системы». Так, уменьшение уровня здоровья людей отражается на деятельности всего государства: увеличиваются расходы на здравоохранение, экономические потери, изменения в составе рабочей силы и т.д.

Состояние общественного здоровья автор приводит в качестве зависимой переменной, которое, в свою очередь, зависит от качества жизни населения, характеризующегося степенью удовлетворения материальных и культурных потребностей людей. Важно отметить, что общего повышения уровня здоровья населения не добиться преобразованиями в системе медицинского обслуживания. Для здорового населения необходимы качественное питание, достаточное количество образовательных учреждений, комфортные условия труда, транспортное обслуживание и многое другое.

Для рассмотрения качества общественного здоровья необходимо использовать различные показатели, наиболее важными и достоверными из которых являются: ОПЖ населения, мужчин и женщин отдельно, коэффициенты смертности (от всех причин и от каждой в отдельности), младенческая смертность, заболеваемость общая и некоторыми социально значимыми болезнями. Совокупность данных показателей позволила автору оценить уровень здоровья населения и на основе этой оценки сравнить между собой регионы России.

Качество жизни населения регионов, в свою очередь, оценивается такими факторами, как уровень индустриального развития, степень урбанизации, эколого-гигиеническая обстановка, комфортность природных условий.

Так, дальнейший анализ факторов производится отдельно по каждому федеральному округу, внутри которых регионы ранжируются по тому, насколько каждый из них благоприятен для жизни. Например, наиболее благоприятная экономическая ситуация, способствующая качеству общественного здоровья, на фоне других субъектов федерации наблюдается в Москве и автономных округах Тюменской области.

Также, оценивая медико-демографическую ситуацию в федеральных округах и анализируя социально-экономические показатели, автор проводит ранжирования федеральных округов по уровню необходимости осуществления мероприятий по улучшению условий жизни населения, повышению уровня общественного здоровья. По оценке Б.Б. Прохорова самого пристального внимания в этой связи требуют Сибирский и Дальневосточный округа и некоторые (депрессивные) районы Центра и Северо-Запада. Исследователем отмечено, что предупреждающие ухудшение уровня здоровья меры должны осуществляться и регулироваться на всех уровнях управления (федеральном, региональном и муниципальном), с обязательным учетом местных социально-экономических, природных, производственных и эколого-гигиенических особенностей.

## 8.3. Использование статей в настоящей работе.

Содержание данных статей послужило основой в принятии решения о том, какие социально-экономические показатели стоит выбрать для анализа влияния на заболеваемость населения России (независимые переменные). К тому же, благодаря изученным материалам, был принят во внимание следующий факт: социально-экономические показатели могут сильно разниться между регионами Российской Федерации, некоторые субъекты отличаются неблагоприятными условиями жизни населения. Так, при выделении и определении возможных причин возникновения выбросов были учтены особенности регионов, которым соответствовали конкретные выделяющиеся значения.

1. Для каждой гистограммы проведена нормальная кривая [↑](#footnote-ref-1)
2. Тест Парка проводился с помощью функций Excel. Линейная регрессия ln(ei)2 = ⌃γ ln(Zi) + vi была построена с помощью инструментов SPSS. [↑](#footnote-ref-2)
3. Тест Глейзера был проведен с помощью инструментов SPSS, используемый алгоритм представлен в следующей ссылке: <http://www.spsstests.com/2015/03/test-heteroskedasticity-glejser-using.html> [↑](#footnote-ref-3)
4. Тест Голдфелда-Квандта проводился с помощью функций Excel и инструментов SPSS. Расчёт F-статистики производился с помощью функции Excel. [↑](#footnote-ref-4)