Правительство Российской Федерации

**Федеральное государственное автономное учреждение  
высшего профессионального образования**

**«Национальный исследовательский университет   
«Высшая школа экономики»**

**Факультет бизнеса и менеджмента**

**Школа бизнес-информатики**

**Проектная работа**

**По теме:** Влияние социально-экономических факторов на число ДТП по регионам РФ

Подготовили:

Немченко Олег, 134 гр

Яковлева Дарья, 134 гр

Долин Василий, 134 гр

Проверили:

Иванова Мария Евгеньевна

Поляков Константин Львович

**Москва 2016**

Оглавление

[1. Постановка задачи 4](#_Toc453014955)

[1.1. Формулировка прикладной проблемы, которую планируется решить 4](#_Toc453014956)

[1.2. Потенциальные потребители решения, задачи, которые они смогут решать, используя полученные результаты 4](#_Toc453014957)

[1.3. Основные гипотезы, которые планируется проверить в рамках решения задачи 4](#_Toc453014958)

[1.4. Основные источники данных 5](#_Toc453014959)

[1.5. Публикации на тему 5](#_Toc453014960)

[2. Предварительный анализ собранных данных 6](#_Toc453014961)

[2.1. Анализ особенностей данных 6](#_Toc453014962)

[2.1.1. Число дорожно-транспортных происшествий 7](#_Toc453014963)

[2.1.2. Реальная начисленная заработная плата 10](#_Toc453014964)

[2.1.3. Уровень безработицы 14](#_Toc453014965)

[2.1.4. Категория региона по удельному весу расходов на оплату ЖКУ 18](#_Toc453014966)

[2.1.5. Рейтинговый балл региона по уровню жизни 25](#_Toc453014967)

[2.1.6. Население 32](#_Toc453014968)

[2.1.7. Средний возраст автомобилей 36](#_Toc453014969)

[2.1.8. Количество автомобилей на 1000 человек 40](#_Toc453014970)

[2.1.9. Пассажирооборот 44](#_Toc453014971)

[2.2. Анализ статистической связи 48](#_Toc453014972)

[2.2.1 Диаграмма рассеивания Реальная заработная плата и число ДТП 48](#_Toc453014973)

[2.2.2. Диаграмма рассеивания Население и число ДТП 48](#_Toc453014974)

[2.2.3. Диаграмма рассеивания Рейтинговый балл по уровню жизни и число ДТП 49](#_Toc453014975)

[2.2.4. Диаграмма рассеивания Доля безработных и число ДТП 50](#_Toc453014976)

[2.2.5. Диаграмма рассеивания Средний возраст автомобиля и число ДТП 51](#_Toc453014977)

[2.2.6. Диаграмма рассеивания Пассажиропоток и число ДТП 51](#_Toc453014978)

[2.2.7. Диаграмма рассеивания Количество автомобилей на 1000 человек и число ДТП 52](#_Toc453014979)

[2.2.8. Диаграмма Box-Whiskers для Коммунальных расходов и числа ДТП 52](#_Toc453014980)

[2.2.9. Корреляционные матрицы и анализ ANOVA 53](#_Toc453014981)

[3. Линейная регрессия 58](#_Toc453014982)

[4. Эссе по найденным публикациям 59](#_Toc453014983)

[5. Спецификация базовой модели 60](#_Toc453014984)

[5.1. Связь структуры базовой модели с основными гипотезами исследования. 60](#_Toc453014985)

[5.2. Связь структуры базовой модели с результатами предварительного анализа данных 61](#_Toc453014986)

[6. Оценивание базовой модели 62](#_Toc453014987)

[6.1. Проверка мультиколлинеарности 62](#_Toc453014988)

[6.1.1 VIF-анализ 62](#_Toc453014989)

[6.1.2 Индекс обусловленности 64](#_Toc453014990)

[6.2. Проверка гетероскедастичности 64](#_Toc453014991)

[6.2.1 Проверка гетероскедастичности произвольной формы. Тест Breusch-Pagan 64](#_Toc453014992)

[6.2.2 Проверка гетероскедастичности по конкретной переменной. Тест Park 64](#_Toc453014993)

[6.2.3 Тест Goldfeld-Quandt 66](#_Toc453014994)

[6.3. Проверка гипотез 68](#_Toc453014995)

[7. Оптимизация базовой модели 68](#_Toc453014996)

[7.1 Модифицированный коэффициент детерминации 69](#_Toc453014997)

[7.2 Информационные критерии 70](#_Toc453014998)

[7.3 Проверка гипотез для оптимальной модели 71](#_Toc453014999)

[7.4 Проверка мультиколлинеарности и гетероскедастичности в рамках оптимальной модели 73](#_Toc453015000)

[7.4.1 Проверка мультиколлинеарности с помощью VIF-анализа 73](#_Toc453015001)

[7.4.2 Проверка мультиколлинеарности с помощью индекса обусловленности 77](#_Toc453015002)

[7.4.3 Проверка гетероскедастичности произвольной формы. Тест Breusch-Pagan 77](#_Toc453015003)

[7.4.4 Проверка гетероскедастичности по конкретной переменной. Тест Park 78](#_Toc453015004)

[7.4.5 Проверка гетероскедастичности по конкретной переменной. Тест Goldfeld-Quandt 80](#_Toc453015005)

[8. Построение доверительных интервалов 81](#_Toc453015006)

[8.1 Построение доверительных интервалов для истинного значения регрессии 81](#_Toc453015007)

[8.2 Построение доверительных интервалов для параметров модели 82](#_Toc453015008)

[9. Обнаружение выбросов и влиятельных наблюдений в модели 83](#_Toc453015009)

[9.1. Расстояние Махаланобиса 83](#_Toc453015010)

[9.2. Расстояние Кука 84](#_Toc453015011)

[9.3. Ковариационное отношение 84](#_Toc453015012)

[9.4. Меры DFFIT и DFFTIS 84](#_Toc453015013)

[9.5. Меры DFBETA и DFBETAS 85](#_Toc453015014)

# Постановка задачи

## Формулировка прикладной проблемы, которую планируется решить

В основе прикладной задачи лежит ежегодная тенденция к снижению числа ДТП с одновременным износом дорог и недостаточностью мероприятий по их ремонту и строительству. В связи с этим несколько противоречивым фактом было обращено внимание на социально-экономические показатели, которые теоретически могут оказывать влияние на положение дел в ситуации с дорожно-транспортными происшествиями.  
Так, в рамках нашего исследования необходимо проанализировать, как рост реальной заработной платы на 2% и доля безработицы 5% повлияют на количество ДТП в регионе.

Чтобы достичь поставленную цель, был исследован механизм зависимости числа ДТП всех категорий от различных социально-экономических показателей субъектов Российской Федерации. Для построения модели были выбраны следующие факторы:

* Прирост реальной заработной платы
* Численность населения
* Рейтинговый балл региона по уровню жизни
* Уровень безработицы
* Удельный вес расходов на оплату жилищно-коммунальных услуг
* Средний возраст автомобилей
* Число автомобилей на тысячу человек

Все данные взяты за 2014 год по регионам РФ в связи с их доступностью и полнотой.

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

Основными потребителями исследования могут являться страховые компании, имеющие филиалы в разных регионах страны, и правительство РФ. Результаты исследования могут быть использованы для прогнозирования страховых случаев и влияния на количество ДТП за счет индексирования.

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

* Число ДТП в регионах прямо пропорционально зависит от численности населения и количества автомобилей на тысячу человек;
* Число ДТП в регионах находится в зависимости от рейтингового балла качества жизни. А именно, чем выше балл, тем меньше количество ДТП, что может объясняться тем, что при более высоком уровне благосостояния семей в регионе жители покупают более дорогие автомобили и более трепетно к ним относятся, что отражается на их стиле вождения;
* Число ДТП в регионах имеет прямую зависимость от таких экономических показателей региона, как прирост реальной заработной платы и уровень безработицы. Чем выше уровень безработицы и меньше прирост реальной заработной платы, тем меньше средств у жителей для поддержания автомобиля на должно уровне и, как результат, тем больше количество ДТП.
* Уровень безработицы имеет прямое влияние на средний возраст автомобилей, и в совокупности переменные при своем увеличении также увеличивают число ДТП.

## Основные источники данных

Данные о приросте заработной платы, численности населения уровня безработицы, пассажирообороте взяты на сайте Федеральной службы государственной статистики (<http://www.gks.ru/bgd/regl/B14_14p/Main.htm)>.

Статистика по числу ДТП взята на сайте информационного интернет-агентства о ДТП, автомобилях и дорожном движении (http://www.1gai.ru/513826-statistika-dtp-v-rossii-za-yanvar-noyabr-2014-goda-1-str.html).

Рейтинг регионов по качеству жизни семей взят из исследования российского агентства международной информации «РИА Новости» (<http://vid1.rian.ru/ig/ratings/life_2014.pdf>).

Информация о количестве автомобилей на 1000 человек была взята с портала Knoema с атласом данных и инфографикой ([http://knoema.ru/pewjcue/регионы-россии-число-собственных-легковых-автомобилей-1970-2014?tsId=1000080&action=download#](http://knoema.ru/pewjcue/регионы-россии-число-собственных-легковых-автомобилей-1970-2014?tsId=1000080&action=download) )

Данные о среднем возрасте автомобилей позаимствованы на сайте аналитического агентства «Автостат» (<http://www.autostat.ru/news/21297/> )

## Публикации на тему

Поиск по публикациям на заявленную тему не дал результатов, так как все исследования числа ДТП фокусируются лишь на технических характеристиках автомобилей, погодных условия и состоянии дорог.

# Предварительный анализ собранных данных

## Анализ особенностей данных

Для исследования было выбрано 10 переменных:

* Число дорожно-транспортных происшествий (числовая, зависимая, измеряется в шт. и представляет собой натуральное число)
* Пострадавшие (раненые и погибшие) в ДТП (номинальная, зависимая, допустимые значения: 0, 1)
* Прироста реальной заработной платы (числовая, независимая, измеряется в процентном соотношении к предыдущему периоду, действительное число)
* Уровень безработицы (числовая, независимая, измеряется в процентном соотношении к общей численности населения)
* Категория региона по удельному весу расходов на оплату жилищно-коммунальных услуг (номинальная, независимая, допустимые значения: 1,2,3,4,5,6,7)
* Население (числовая, независимая, измеряется в тыс. человек, натуральное число)
* Рейтинговый балл региона по уровню жизни (числовая, независимая, представляет собой действительное число)
* Средний возраст автомобилей (числовая, независимая, измеряется к количестве лет, действительное число)
* Количество автомобилей на 1000 человек (числовая, независимая, измеряется в шт. на тыс. человек, действительное число)
* Пассажирооборот (числовая, независимая, измеряется в млн. пассажирокилометрах, натуральное число)

По оценкам авторов исследования, данная выборка может быть не особо репрезентативной, так как за основу взяты субъекты РФ в количестве 84 штук и предкризисный год. При учете большего количества городов, возможно, в том числе зарубежных, можно было бы выявить более точные зависимости и построить более точные регрессии. Выбор регионов РФ и года был продиктован полнотой набора данных и их доступностью. Так, взятая для анализа информация не имеет пропущенных значений и ошибок.

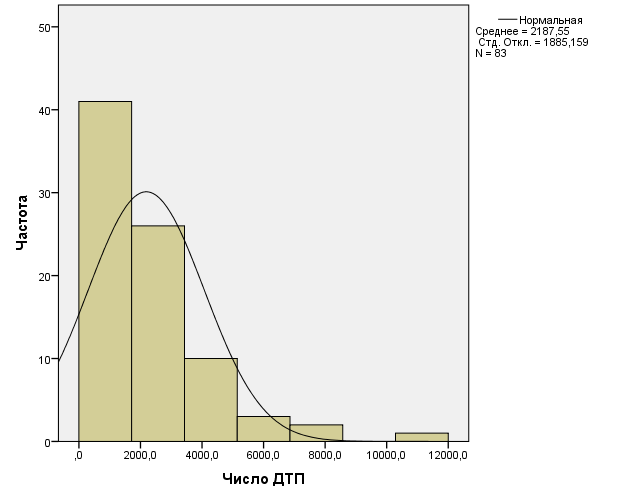
Кроме того, важно отметить, что изначально существует естественная группировка данных по округам (Центральный, Южный и т.д.), однако был произведен намеренный отказ от данного разделения, так как данные в рамках каждой такой группы слишком разнообразны и имеют большой разброс.

### Число дорожно-транспортных происшествий

Зависимая переменная"число дорожно-транспортных происшествий"представляет собой общее количество ДТП всех категорий, произошедших на территории Российской Федерации в 2014 году.

Данные представлены по всем 83 регионам Российской Федерации и не имеют пропусков.

Для проведения графического анализа была построенная гистограмма. Количество столбцов на гистограмме частот рассчитывалось по формуле Стерджесса и получилось равным 7.



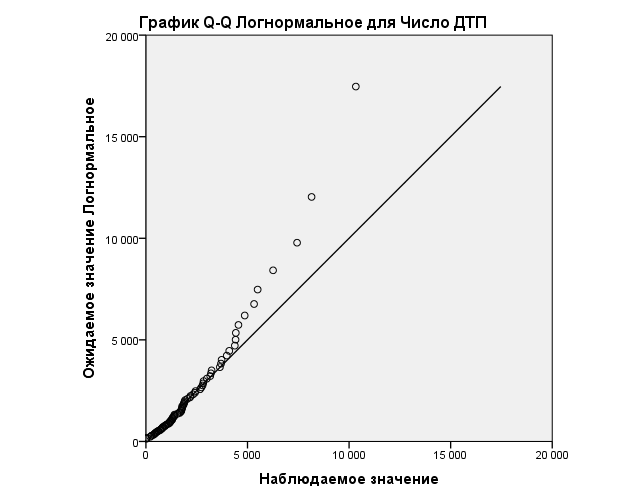
По построенной гистограмме можно судить о наличии асимметрии данных вправо и наличии высокой вершины. Чтобы выяснить, подтвердить выявленные на графике асимметрию и положительный эксцесс, а также убедиться, что распределение числа ДТП не является нормальным, были вычислены коэффициенты эксцесса и асимметрии. Кроме того, был применен тест Колмогорова-Смирнова (таблица 1 и 2)

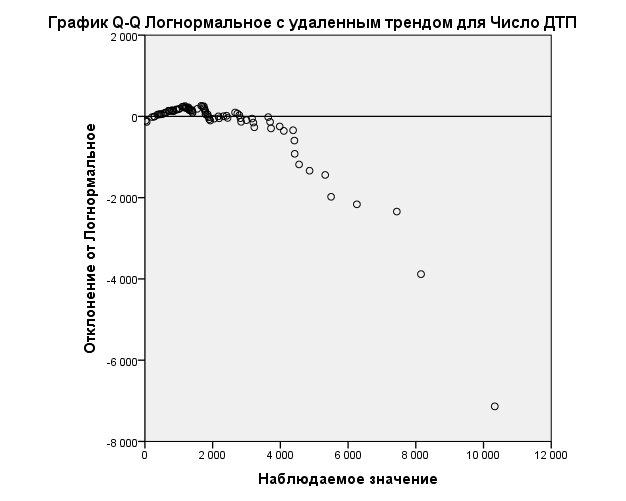
|  |  |  |
| --- | --- | --- |
| **Статистика** | | |
| Число ДТП | | |
| N | Допустимо | 83 |
| Пропущенные | 1 |
| Асимметрия | | 1,873 |
| Стандартная Ошибка асимметрии | | ,264 |
| Эксцесс | | 4,489 |
| Стандартная ошибка эксцесса | | ,523 |

**Таблица 1 «Коэффициенты асимметрии и эксцесса для числа ДТП»**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | **Одновыборочный критерий Колмогорова-Смирнова** | | | |  | | Число ДТП | | N | | 83 | | Параметры нормального распределенияa,b | Среднее значение | 2187,554 | | Среднеквадратичная отклонения | 1885,1588 | | Наибольшие экстремальные расхождения | Абсолютная | ,181 | | Положительные | ,181 | | Отрицательные | -,125 | | Статистика критерия | | ,181 | | Асимптотическая значимость (2-сторонняя) | | ,000c | | a. Проверяемое распределение является нормальным. | | | | b. Вычислено из данных. | | | | c. Коррекция значимости Лильефорса. | | |   **Таблица 2 «Тест Колмогорова-Смирнова для числа ДТП»** |

Следовательно, асимметрия данных и островершинность распределения действительно присутствуют, а само по себе распределение не является нормальным. С помощью графиков Q-Q было определено, что данное распределение скорее логнормально. Таким образом, расчетная переменная скорее всего будет под знаком логарифма.





Описательные статистики переменной представлены в таблице 3. С помощью разницы среднего и медианы (2188 и 1740 соответственно) можно подтвердить наличие упомянутой выше асимметрии вправо. Кроме того, имеет место разброс данных относительно среднего значения, равный 3553823,860 (выборочная дисперсия). Стандартная отклонения равна 1885,1588.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Описательные статистики** | | | | |
|  | | | Статистика | Стандартная Ошибка |
| Число ДТП | Среднее значение | | 2187,554 | 206,9231 |
| 95% Доверительный интервал для среднего | Нижняя граница | 1775,918 |  |
| Верхняя граница | 2599,190 |  |
| Среднее по выборке, усеченной на 5% | | 1980,653 |  |
| Медиана | | 1740,000 |  |
| Дисперсия | | 3553823,860 |  |
| Стандартная отклонения | | 1885,1588 |  |
| Минимум | | 23,0 |  |
| Максимум | | 10332,0 |  |
| Диапазон | | 10309,0 |  |
| Межквартильный диапазон | | 1920,0 |  |
| Асимметрия | | 1,873 | ,264 |
| Эксцесс | | 4,489 | ,523 |

**Таблица 3 «Описательные статистики переменной «число ДТП»»**

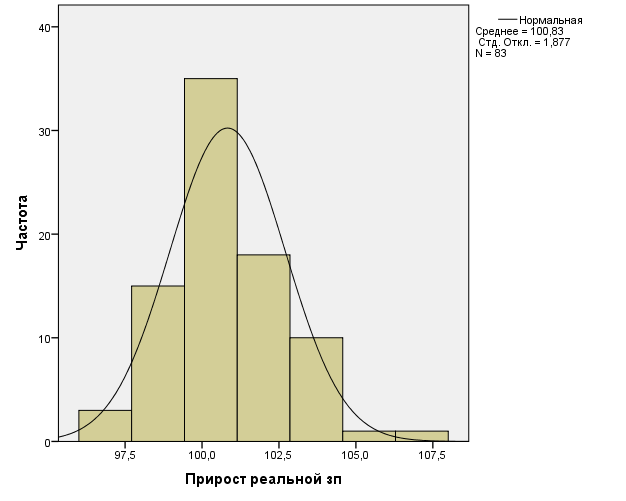
По критерию Райта были выявлены два выброса: Москва и Московская область, имеющие значительно большее количество ДТП, чем остальные регионы. Это, прежде всего, может связано с тем, что в этих местах населения большего количества жителей по сравнению с другими имеется больше автомобилей. Эти выбросы не исключаются из исходных данных, так как Москва и Московская область репрезентируют большой процент населения РФ (суммарно дают около 15% от общего числа населения России).

### Реальная начисленная заработная плата

Независимая переменная"реальная начисленная заработная плата" характеризует покупательную способность заработной платы в отчетном периоде в связи с изменением цен на потребительские товары и услуги по сравнению с базисным периодом.

Для этого рассчитывается индекс реальной начисленной заработной платы путем деления индекса номинальной начисленной заработной платы на индекс потребительских цен за один и тот же временной период. Информация по этому фактору представлена для всех регионов и не имеет пропусков.

Гистограмма, построенная на основе переменной, как и предыдущая, имеет 7 категорий. Опираясь на нее, можно сделать предположение об отсутствии асимметрии и эксцесса.



На основании коэффициентов асимметрии, эксцесса и теста Колмогорова-Смирнова было доказано, что распределение является нормальным, но имеют место незначительная асимметрия и положительный эксцесс (островершинность распределения) (таблицы 4,5).

|  |  |  |
| --- | --- | --- |
| **Статистика** | | |
| Прирост реальной зп | | |
| N | Допустимо | 83 |
| Пропущенные | 30 |
| Асимметрия | | ,466 |
| Стандартная Ошибка асимметрии | | ,264 |
| Эксцесс | | ,284 |
| Стандартная ошибка эксцесса | | ,523 |

**Таблица 4 «Коэффициенты асимметрии и эксцесса для переменной «прирост реальной заработной платы»»**

|  |  |  |
| --- | --- | --- |
| **Одновыборочный критерий Колмогорова-Смирнова** | | |
|  | | Прирост реальной зп |
| N | | 83 |
| Параметры нормального распределенияa,b | Среднее значение | 100,828 |
| Среднеквадратичная отклонения | 1,8773 |
| Наибольшие экстремальные расхождения | Абсолютная | ,084 |
| Положительные | ,084 |
| Отрицательные | -,042 |
| Статистика критерия | | ,084 |
| Асимптотическая значимость (2-сторонняя) | | ,200c,d |
| a. Проверяемое распределение является нормальным. | | |
| b. Вычислено из данных. | | |
| c. Коррекция значимости Лильефорса. | | |
| d. Это нижняя граница истинной значимости. | | |

**Таблица 5 «Тест Колмогорова-Смирнова для переменной «прирост реальной заработной платы»»**

Сравнение нормального распределения и распределения реальной было проделано с помощью графика Q-Q, чем была доказана их близость.





Описательные статистики переменной представлены в таблице 6. Так, среднее значение больше медианы на 0,145, что говорит о незначительной асимметрии вправо. Дисперсия и стандартная отклонения равны 3,524 и 1,8773 соответственно.

|  |  |  |  |
| --- | --- | --- | --- |
| **Описательные статистики** | | | |
|  | | | Статистика |
| Прирост реальной зп | Среднее значение | | 100,828 |
| 95% Доверительный интервал для среднего | Нижняя граница | 100,418 |
| Верхняя граница | 101,238 |
| Среднее по выборке, усеченной на 5% | | 100,784 |
| Медиана | | 100,683 |
| Дисперсия | | 3,524 |
| Стандартная отклонения | | 1,8773 |
| Минимум | | 97,0 |
| Максимум | | 106,7 |
| Диапазон | | 9,7 |
| Межквартильный диапазон | | 2,6 |
| Асимметрия | | ,466 |
| Эксцесс | | ,284 |

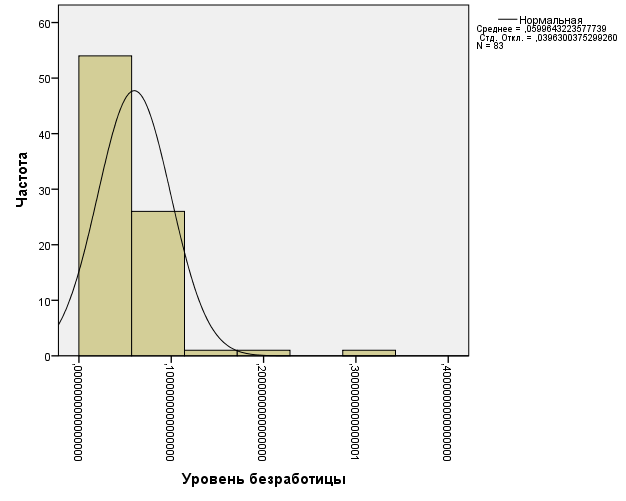
**Таблица 6 «Описательные статистики переменной «Прирост реальной заработной платы»»**

По критерию Райта выбросом является Чукотский автономный округ, имеющий максимальный прирост относительно других регионов. Кроме того, это регион известен своей максимальной по России средней заработной платой, что может связано с особенностями его местоположения и достаточной закрытости. Этот выброс исключается их исходных данных, так как Чукотский автономный округ не является критичным для исследования.

### Уровень безработицы

Независимая переменная «уровень безработицы»представляет собой количество безработных в регионе, деленое на количество трудоспособных. Информация по данному фактору есть по всем 83 регионам РФ.

С помощью графического представления данных было обнаружено, что существует асимметрия данных вправо и незначительный положительный эксцесс.



Для подтверждения гипотез, сделанных на основе гистограммы, были рассчитаны коэффициенты асимметрии и эксцесса, которые показали, что действительно присутствует асимметрия вправо и небольшая островершинность распределения. Кроме того, тест Колмогорова-Смирнова показал, что распределение не является нормальным (таблицы 7 и 8).

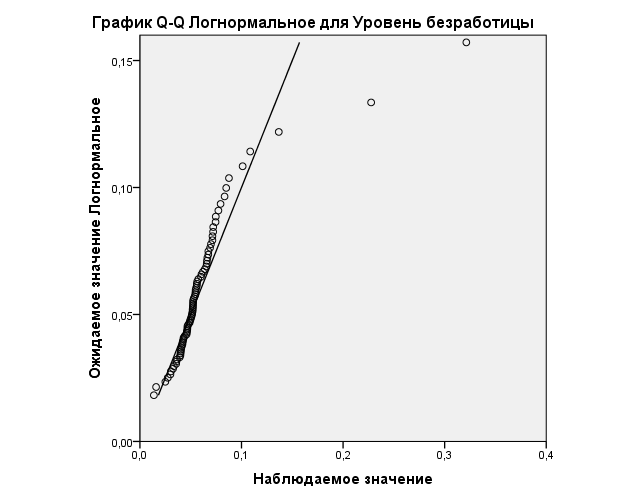
|  |  |  |
| --- | --- | --- |
| **Статистика** | | |
| Уровень безработицы | | |
| N | Допустимо | 83 |
| Пропущенные | 0 |
| Асимметрия | | 4,570 |
| Стандартная Ошибка асимметрии | | ,264 |
| Эксцесс | | 26,220 |
| Стандартная ошибка эксцесса | | ,523 |

**Таблица 7 «Коэффициенты асимметрии и эксцесса для переменной «уровень безработицы»»**

|  |  |  |
| --- | --- | --- |
| **Одновыборочный критерий Колмогорова-Смирнова** | | |
|  | | Уровень безработицы |
| N | | 83 |
| Параметры нормального распределенияa,b | Среднее значение | ,059964322357774 |
| Среднеквадратичная отклонения | ,039630037529926 |
| Наибольшие экстремальные расхождения | Абсолютная | ,235 |
| Положительные | ,235 |
| Отрицательные | -,176 |
| Статистика критерия | | ,235 |
| Асимптотическая значимость (2-сторонняя) | | ,000c |
| a. Проверяемое распределение является нормальным. | | |
| b. Вычислено из данных. | | |
| c. Коррекция значимости Лильефорса. | | |

**Таблица 8 «Тест Колмогорова-Смирнова для переменной «уровень безработицы»»**

По графику Q-Q было определено, что распределение скорее логнормально. Таким образом, если данная переменная будет входить в формулу расчета числа ДТП, то под математической операцией логарифма.





Описательные статистики переменной доказывают асимметрию вправо: среднее значение больше медианы на 0,07745 (что весомо при таких значениях данной переменной). Дисперсия равна 0,002, а стандартная отклонения - 0,039630037529926.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Описательные статистики** | | | | |
|  | | | Статистика | Стандартная Ошибка |
| Уровень безработицы | Среднее значение | | ,059964322357774 | ,004349961742836 |
| 95% Доверительный интервал для среднего | Нижняя граница | ,051310862739918 |  |
| Верхняя граница | ,068617781975630 |  |
| Среднее по выборке, усеченной на 5% | | ,054635225079965 |  |
| Медиана | | ,052219321148825 |  |
| Дисперсия | | ,002 |  |
| Стандартная отклонения | | ,039630037529926 |  |
| Минимум | | ,0137388970541249 |  |
| Максимум | | ,3212851405622491 |  |
| Диапазон | | ,3075462435081242 |  |
| Межквартильный диапазон | | ,0240741353184530 |  |
| Асимметрия | | 4,570 | ,264 |
| Эксцесс | | 26,220 | ,523 |

**Таблица 9 «Описательная статистика для переменной «уровень безработицы»**

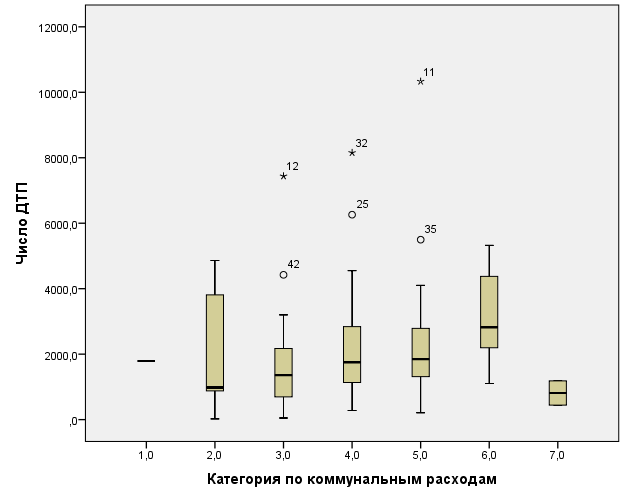
Выбросами по критерию Райта являются Ингушетия и Дагестан, где крайне большое число безработных при достаточно низком по сравнению с другими регионами количестве населения. Эти выбросы были исключены из исходных данных, так как регионы не являются критическими для исследования.

### Категория региона по удельному весу расходов на оплату ЖКУ

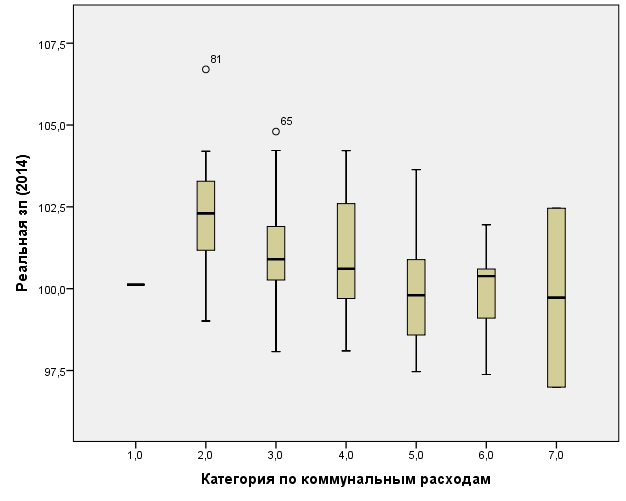
Независимая переменная«категория региона по удельному весу расходов на оплату жилищно-коммунальных услуг» является номинальной. Всего существует 7 категорий, которые пронумерованы от 1 до 7. При этом в категорию 1 входит только один регион – Москва.

Для анализа переменной были построены диаграммы Бокса-Вискера в сочетании с другими данными, используемыми в рамках исследования.

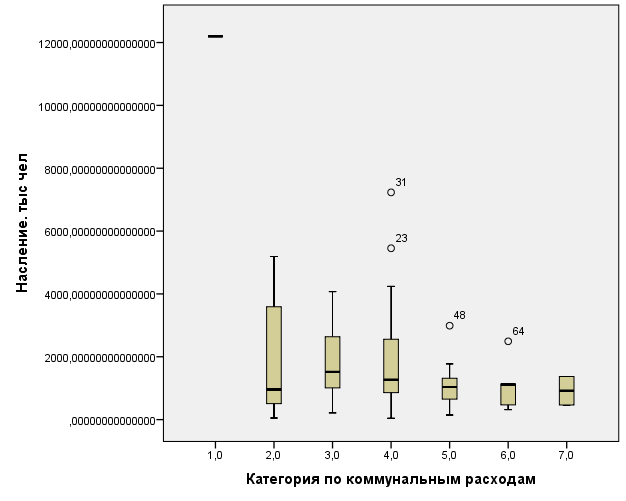
В пересечении числа ДТП и категорий по коммунальным расходам получается три критических выброса: Еврейская автономная область, Забайкальский край и Мурманская область. Всем выбросам свойственны примерно средние категории по доли расходов на коммунальные платежи, при это количество ДТП выше среднего показателя. Ящикам категорий 1, 2 и 7 свойственна относительная симметричность данных, тогда как ящики категории 4 и 6 не совсем симметричны, а ящик 2 имеет крайнюю степень асимметричности – медиана практически в два с половиной раза меньше второго квартиля.



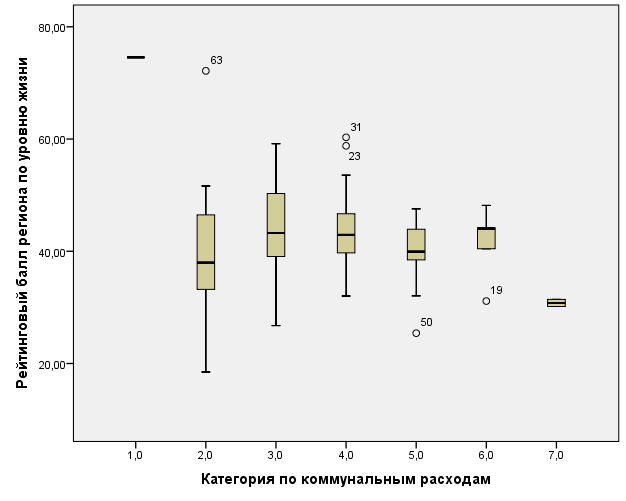
В пересечении прироста заработной платы и категорий по коммунальным расходам нет критических выбросов. При этом наблюдается крайне широкий разброс данных в 7 категории и сильная асимметричность в 6. Остальные категории относительно симметричны.



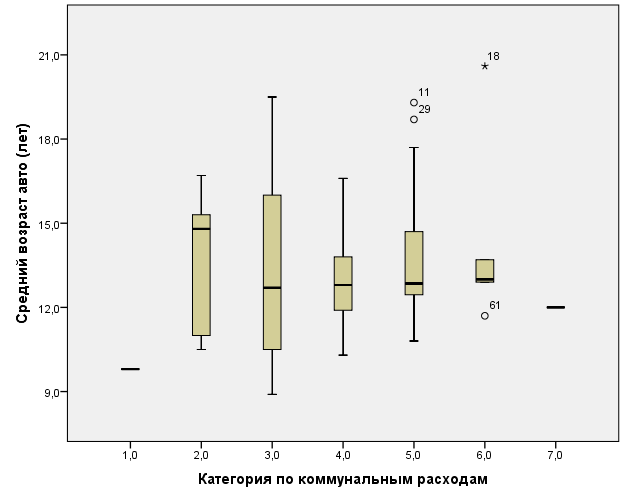
Критические выбросы в пересечении населения с данными категориями также отсутствуют. Примечательно также то, что с увеличением номера категории сокращается разброс данных в ящиках. Значительно отличающейся категорией является 6, так как медиана в ящике практически совпадает со вторым квартилем.



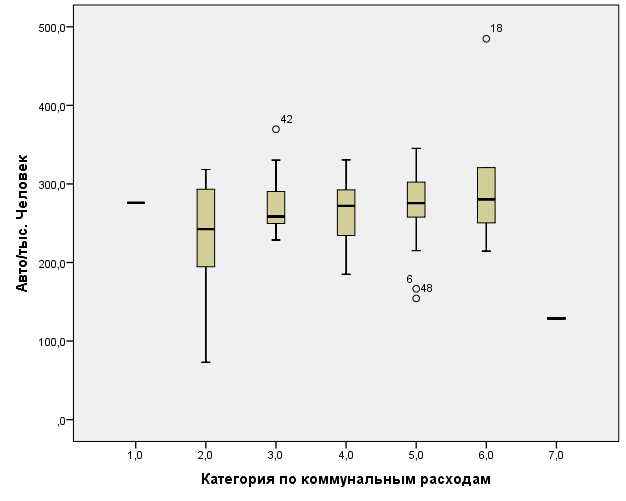
В рамках рассмотрения совокупности рейтингового балла по уровню жизни и категорий регионов по коммунальным платежам вновь не наблюдается критических выбросов. Кроме того, снова присутствует сокращение разброса данных с ростом номера категории. Еще один примечательный факт – категория 7 имеет минимальную медиану относительно других.



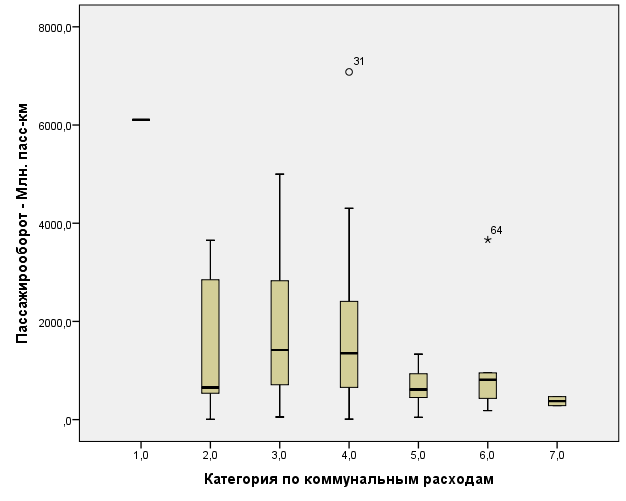
Камчатская область является выбросом в пересечении категорий со средним возрастом автомобилей в регионе. Область имеет 6 категорию по доле платежей за коммунальные услуги среди других расходов и максимальный срок эксплуатации автомобилей по всей РФ. Это может быть объяснено, например, крайне высокой ценой на новые автомобили. Кроме того, наблюдается большой разброс в категории 2: средний возраст начинается с 10,5 лет и заканчивается 16,7.



В пересечении количества автомобилей на 1000 человек и категорий регионов нет критических выбросов. Отличительной категорией здесь может считаться 2, так как она имеет больший разброс, чем остальные. При этом асимметричность ящиков наблюдается во всех категориях, кроме тех, где представлен один регион (1 категория) или значения переменных совпадают (7 категория).



Критический выброс в пересечении категорий регионов по коммунальным платежам с пассажирооборотом – Саратовская область. Ей свойственна 6 категория по доле расходов на коммунальные платежи и значительный пассажирооборот. В рамках данной диаграммы наблюдается также тенденция сокращения разброса данных о пассажиропотоке с увеличением номера категории.



Теоретически, число групп можно было бы сократить, однако при этом возникнет неоднородность между группами, заключающаяся в числе элементов, входящих в каждую категорию.

### Рейтинговый балл региона по уровню жизни

Независимая переменная «рейтинговый балл региона по уровню жизни»была получена в результате объединения 61 показателя, объединенного в 10 групп, которые характеризуют все основные аспекты условий проживания в регионе, от уровня экономического развития и объема доходов населения, до обеспеченности населения различными видами услуг и климатических условий в регионе проживания. Группы, в которые объединены показатели рейтинга:

* *Уровень доходов населения*, который рассчитывался на основании следующих показателей: объем вкладов (депозитов) физических лиц в банках на одного жителя, отношение денежных доходов населения к стоимости фиксированного набора потребительских товаров и услуг, доля населения с доходами ниже прожиточного минимума;
* *Жилищные условия населения*, которые оценивались на основании следующих показателей: общая площадь жилых помещений, приходящаяся в среднем на одного жителя за исключением ветхого и аварийного фонда; доля населения, проживающего в многоквартирных домах, признанных в установленном порядке аварийными; удельный вес общей площади, оборудованной водопроводом; удельный вес общей площади, оборудованной водоотведением (канализацией); удельный вес общей площади, оборудованной горячим водоснабжением; уровень износа коммунальной инфраструктуры;
* *Обеспеченность объектами социальной инфраструктуры*. В данную группу объединены показатели, которые характеризуют обеспеченность населения объектами по основным сферам жизни. Это обеспеченность объектами здравоохранения, обеспеченность объектами образования, обеспеченность объектами торговли и обеспеченность объектами досуга.

*Обеспеченность объектами здравоохранения* оценивалась по следующим критериям: обеспеченность врачами на 10 тыс. человек населения; мощность амбулаторно-поликлинических учреждений на 10 тыс. человек населения; обеспеченность больничными койками на 10 тыс. человек населения; средняя занятость койки в году; доля учреждений здравоохранения, здания которых находятся в аварийном состоянии или требуют капитального ремонта; доля выездов бригад скорой медпомощи доезжающих до места вызова в течение 20 минут; доля выездов бригад скорой медпомощи доезжающих до места ДТП в течение 20 минут.

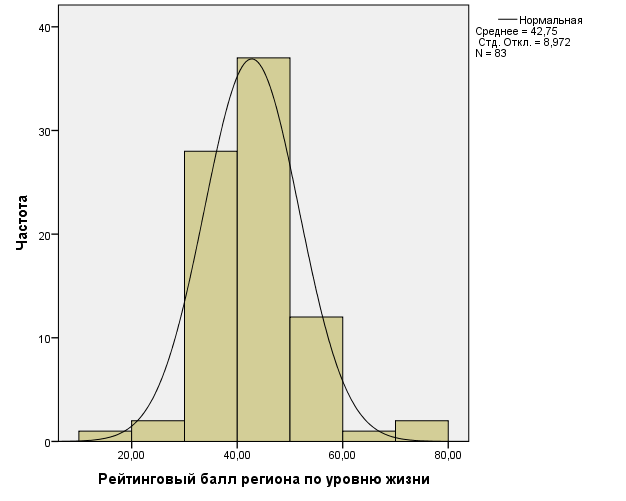
*Обеспеченность объектами образования* оценивалась по следующим критериям: обеспеченность детей дошкольного возраста местами в дошкольных образовательных учреждениях; численность детей, стоящих на учете для определения в дошкольные образовательные учреждения, на 1000 мест в дошкольных образовательных учреждениях; численность обучающихся государственных и муниципальных общеобразовательных учреждений (без вечерних (сменных)) общеобразовательных учреждений на 10 учителей; доля обучающихся в государственных и муниципальных общеобразовательных учреждениях (без вечерних (сменных)), занимающихся во вторую и третью смены; число студентов, принятых в образовательные учреждения высшего профессионального образования, на 1000 человек населения.

*Обеспеченность объектами торговли* оценивалась по следующим критериям: обеспеченность торговыми площадями на 1000 человек; оборот розничной торговли на одного жителя; обеспеченность автозаправочными станциями на 1000 км дорог.

*Обеспеченность объектами досуга* оценивалась по следующим критериям: позиции в рейтинге регионов по развитию спортивной инфраструктуры; обеспеченность ресторанами и столовыми, мест на 1000 жителей; численность зрителей театров на 1000 человек населения; число посещений музеев на 1000 человек населения;

* *Экологические и климатические условия* оценивались на основании: выбросов в атмосферу загрязняющих веществ от стационарных и передвижных источников на единицу площади населенных пунктов; оценки климата; обеспеченности населения питьевой водой, отвечающей требованиям безопасности; затратов на охрану окружающей среды на единицу площади; затратов на охрану окружающей среды на одного жителя;
* *Безопасность проживания* оценивалась по: оценка населением уровня криминогенности в субъекте РФ; смертность населения в ДТП в расчете на 100 тыс. человек населения; число травм, отравлений и некоторых других последствий воздействия внешних причин на 10000 человек населения;
* *Демографическая ситуация* оценивалась на основании: коэффициента естественного прироста (+), убыли (-) населения; миграционного прироста (убыль) населения (используется среднее значение за три года);
* *Здоровье населения и уровень образования* оценивались по: численности зарегистрированных больных с диагнозом, установленным впервые в жизни на 1 000 человек всего населения; уровню младенческой смертности; потерянным годам потенциальной жизни; выпуску из аспирантуры на 10000 человек населения; доле населения с высшим образованием в общей численности населения в возрасте 15 лет и более; численности жителей, не имеющих среднего образования на 1000 человек в возрасте 15 лет и более;
* *Освоенность территории и развитие транспортной инфраструктуры* оценивались по: плотности автомобильных дорог общего пользования с твердым покрытием, км дорог на 1000 кв км территории; плотности железнодорожных путей общего пользования, км путей на 10000 кв. км территории; доле городского населения; доле площади населенных пунктов в общей площади территории;
* *Уровень экономического развития* оценивался на основании: объема ВРП на душу населения; абсолютному объем ВРП; доли собственных доходов в общем объеме доходов консолидированных бюджетов; доли прибыльных предприятий; объема прямых иностранных инвестиций в расчете на 1 жителя абсолютного объем прямых иностранных инвестиций; уровеню безработицы; среднего время поиска работы;
* *Развитие малого бизнеса* оценивалось по: обороту малых и микропредприятий и индивидуальных предпринимателей в расчете на одного жителя; суммарному обороту малых и микропредприятий и индивидуальных предпринимателей; объему инвестиций в основной капитал малых и микропредприятий в расчете на одного жителя; объему инвестиций в основной капитал малых и микропредприятий; доле занятых в малом бизнесе и индивидуальном предпринимательстве в общей численности экономически активного населения.

Из гистограммы можно сделать вывод об отсутствии асимметрии и эксцесса, а распределение данных кажется нормальным.



Расчет коэффициентов говорит о присутствии незначительной асимметрии и положительного эксцесса. Тест Колмогорова-Смирнова показывает, что распределение нормально (таблица 10,11).

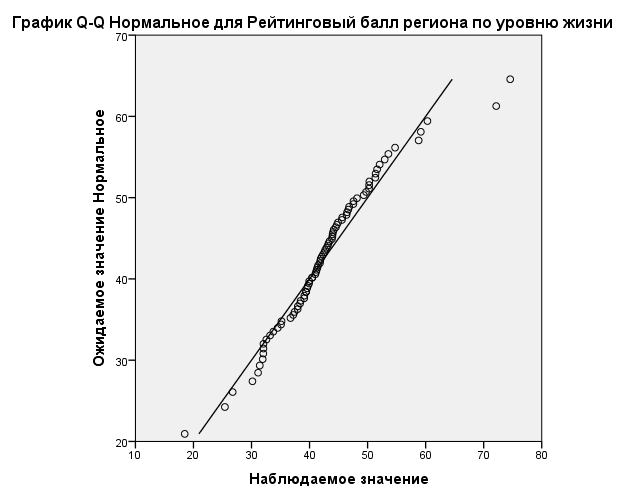
|  |  |  |
| --- | --- | --- |
| **Статистика** | | |
| Рейтинговый балл региона по уровню жизни | | |
| N | Допустимо | 83 |
| Пропущенные | 1 |
| Асимметрия | | ,728 |
| Стандартная Ошибка асимметрии | | ,264 |
| Эксцесс | | 2,456 |
| Стандартная ошибка эксцесса | | ,523 |

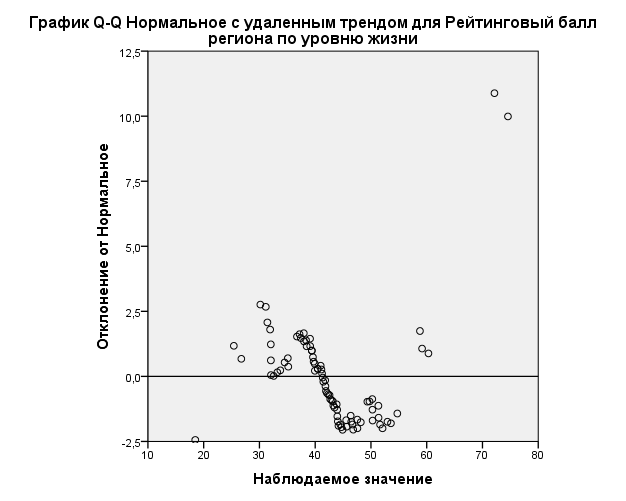
**Таблица 10 «Коэффициенты асимметрии и эксцесса для переменной «Рейтинговый балл региона по уровню жизни»»**

|  |  |  |
| --- | --- | --- |
| **Одновыборочный критерий Колмогорова-Смирнова** | | |
|  | | Рейтинговый балл региона по уровню жизни |
| N | | 83 |
| Параметры нормального распределенияa,b | Среднее значение | 42,7511 |
| Среднеквадратичная отклонения | 8,97170 |
| Наибольшие экстремальные расхождения | Абсолютная | ,092 |
| Положительные | ,092 |
| Отрицательные | -,067 |
| Статистика критерия | | ,092 |
| Асимптотическая значимость (2-сторонняя) | | ,079c |

|  |
| --- |
| a. Проверяемое распределение является нормальным. |
| b. Вычислено из данных. |
| c. Коррекция значимости Лильефорса.  **Таблица 11 «Тест Колмогорова-Смирнова для переменной «Рейтинговый балл региона по уровню жизни»»** |

Тот факт, что распределение действительно близко к нормальному, может быть дополнительно подтвержден графиком Q-Q. Таким образом, если данная переменная будет входить в формулу расчета числа ДТП, то под математической операцией логарифм.





По описательной статистике можно судить о наличии незначительно асимметрии вправо (среднее значение больше медианы лишь на 0,65). Дисперсия равна 80,491, стандартная отклонения - 8,97170.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Описательные статистики** | | | | |
|  | | | Статистика | Стандартная Ошибка |
| Рейтинговый балл региона по уровню жизни | Среднее значение | | 42,7511 | ,98477 |
| 95% Доверительный интервал для среднего | Нижняя граница | 40,7921 |  |
| Верхняя граница | 44,7101 |  |
| Среднее по выборке, усеченной на 5% | | 42,4077 |  |
| Медиана | | 42,1100 |  |
| Дисперсия | | 80,491 |  |
| Стандартная отклонения | | 8,97170 |  |
| Минимум | | 18,49 |  |
| Максимум | | 74,56 |  |
| Диапазон | | 56,07 |  |
| Межквартильный диапазон | | 8,84 |  |
| Асимметрия | | ,728 | ,264 |
| Эксцесс | | 2,456 | ,523 |

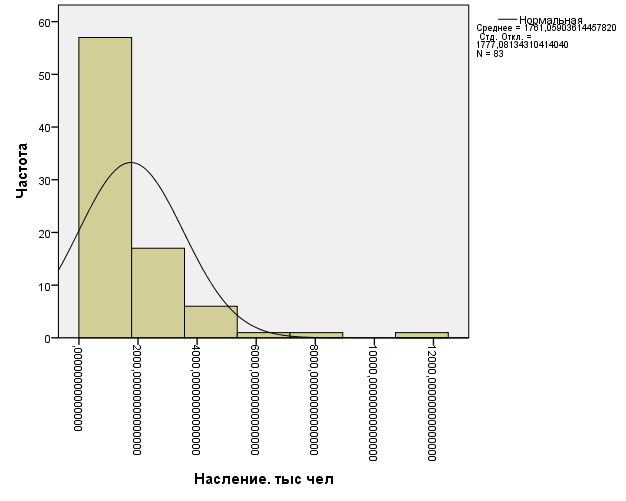
**Таблица 12 «Описательные статистики для переменной «рейтинговый балл региона по уровню жизни»»**

По критерию Райта выбросами являются Москва и Московская область, характеризующиеся крайне высоким уровнем благосостояния семей относительно жителей других регионов РФ. Эти выбросы не исключаются из исходных данных вследствие того, Москва и Московская область репрезентируют большой процент населения РФ.

### Население

Независимая переменная «население» представляет собой численность жителей в каждом регионе, выраженную в тысячах человек.

На основании представленной ниже диаграммы можно сделать предположение о наличии асимметрии вправо и положительного эксцесса.



Расчет коэффициентов асимметрии и эксцесса доказал наличие островершинного распределения и значительно асимметрии (таблица 13). Тест Колмогорова-Смирнова также показала, что распределение не является нормальным (таблица 14).

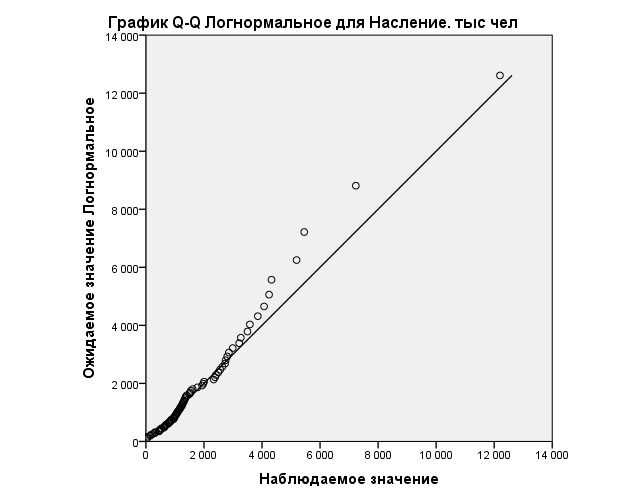
|  |  |  |
| --- | --- | --- |
| **Статистика** | | |
| Насление. тыс чел | | |
| N | Допустимо | 83 |
| Пропущенные | 30 |
| Асимметрия | | 3,133 |
| Стандартная Ошибка асимметрии | | ,264 |
| Эксцесс | | 14,431 |
| Стандартная ошибка эксцесса | | ,523 |

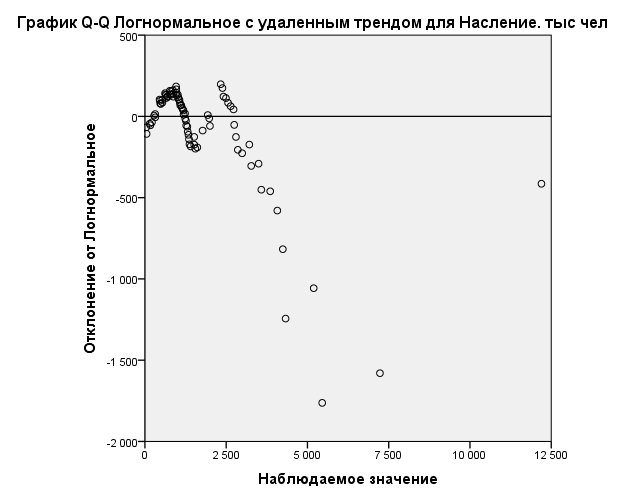
**Таблица 13 «Коэффициенты асимметрии и эксцесса для переменной «население»»**

|  |  |  |
| --- | --- | --- |
| **Одновыборочный критерий Колмогорова-Смирнова** | | |
|  | | Насление. тыс чел |
| N | | 83 |
| Параметры нормального распределенияa,b | Среднее значение | 1761,059036144578000 |
| Среднеквадратичная отклонения | 1777,081343104141000 |
| Наибольшие экстремальные расхождения | Абсолютная | ,210 |
| Положительные | ,210 |
| Отрицательные | -,167 |
| Статистика критерия | | ,210 |
| Асимптотическая значимость (2-сторонняя) | | ,000c |
| a. Проверяемое распределение является нормальным. | | |
| b. Вычислено из данных. | | |
| c. Коррекция значимости Лильефорса. | | |

**Таблица 14 «Тест Колмогорова-Смирнова для переменной «население»»**

Действительно, по графику Q-Q можно судить о том, что распределение скорее логнормальное. Таким образом, если данная переменная будет входить в формулу расчета числа ДТП, то под математической операцией логарифм.





С помощью сравнения среднего значения (1761) и медианы (1191) можно подтвердить значительную асимметрию вправо. Стандартная отклонение - 1777,081343104140600, дисперсия - 3158018,100.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Описательные статистики** | | | | |
|  | | | Статистика | Стандартная Ошибка |
| Насление. тыс чел | Среднее значение | | 1761,059036144578200 | 195,060018567311000 |
| 95% Доверительный интервал для среднего | Нижняя граница | 1373,022521453303600 |  |
| Верхняя граница | 2149,095550835853000 |  |
| Среднее по выборке, усеченной на 5% | | 1539,526572958500300 |  |
| Медиана | | 1191,000000000000000 |  |
| Дисперсия | | 3158018,100 |  |
| Стандартная отклонения | | 1777,081343104140600 |  |
| Минимум | | 43,40000000000000 |  |
| Максимум | | 12197,60000000000000 |  |
| Диапазон | | 12154,20000000000000 |  |
| Межквартильный диапазон | | 1649,70000000000000 |  |
| Асимметрия | | 3,133 | ,264 |
| Эксцесс | | 14,431 | ,523 |

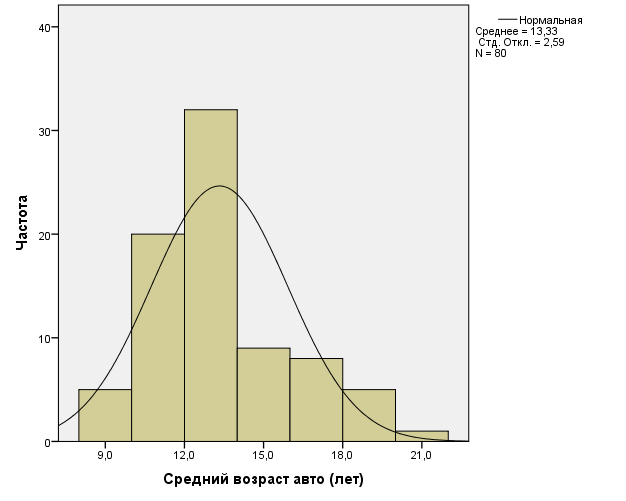
**Таблица 15 «Описательная статистика для переменной «население»»**

По критерию Райта выбросами являются Москва и Московская область – лидеры по количеству населения. Эти выбросы не исключаются из исходных данных, так как Москва и Московская область репрезентируют большой процент населения РФ.

### Средний возраст автомобилей

Независимая переменная «средний возраст автомобилей» имеет значения для всех регионов РФ и выражается в количестве лет.

Графическое изображение данных в виде гистограммы свидетельствует о незначительной асимметрии и эксцессе.



Проверка коэффициентов показала, что действительно существуют и асимметрия, и эксцесс, однако совсем незначительные (таблица 16). Тест Колмогорова-Смирнова говорит, что данные имеют нормальное распределение (таблица 17).

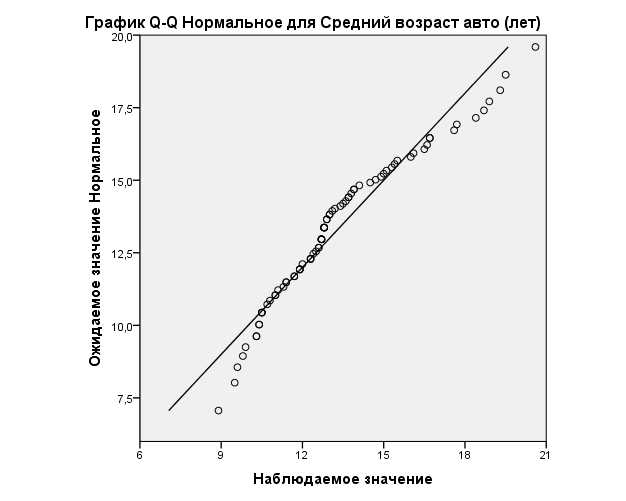
|  |  |  |
| --- | --- | --- |
| **Статистика** | | |
| Средний возраст авто (лет) | | |
| N | Допустимо | 80 |
| Пропущенные | 33 |
| Асимметрия | | ,829 |
| Стандартная Ошибка асимметрии | | ,269 |
| Эксцесс | | ,280 |
| Стандартная ошибка эксцесса | | ,532 |

**Таблица 16 «Коэффициенты асимметрии и эксцесса для переменной «средний возраст автомобилей»**

|  |  |  |
| --- | --- | --- |
| **Одновыборочный критерий Колмогорова-Смирнова** | | |
|  | | Средний возраст авто (лет) |
| N | | 80 |
| Параметры нормального распределенияa,b | Среднее значение | 13,329 |
| Среднеквадратичная отклонения | 2,5898 |
| Наибольшие экстремальные расхождения | Абсолютная | ,138 |
| Положительные | ,138 |
| Отрицательные | -,059 |
| Статистика критерия | | ,138 |
| Асимптотическая значимость (2-сторонняя) | | ,001c |
| a. Проверяемое распределение является нормальным. | | |
| b. Вычислено из данных. | | |
| c. Коррекция значимости Лильефорса. | | |

**Таблица 17 «Тест Колмогорова-Смирнова для переменной «средний возраст автомобилей»»**

Подтверждение нормальности распределения было получено также с помощью построения графика Q-Q.





Описательные статистики подтверждают наличие незначительной асимметрии вправо (среднее больше медианы на 529), в то время как дисперсия равна 6,707, а стандартная отклонения – 2,5898.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Описательные статистики** | | | | |
|  | | | Статистика | Стандартная Ошибка |
| Средний возраст авто (лет) | Среднее значение | | 13,329 | ,2895 |
| 95% Доверительный интервал для среднего | Нижняя граница | 12,752 |  |
| Верхняя граница | 13,905 |  |
| Среднее по выборке, усеченной на 5% | | 13,197 |  |
| Медиана | | 12,800 |  |
| Дисперсия | | 6,707 |  |
| Стандартная отклонения | | 2,5898 |  |
| Минимум | | 8,9 |  |
| Максимум | | 20,6 |  |
| Диапазон | | 11,7 |  |
| Межквартильный диапазон | | 3,4 |  |
| Асимметрия | | ,829 | ,269 |
| Эксцесс | | ,280 | ,532 |

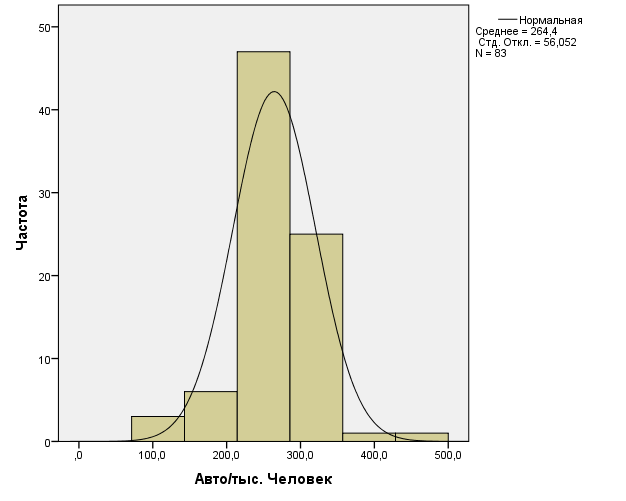
**Таблица 18 «Описательная статистика для переменной «средний возраст автомобилей»»**

По критерию Райта выбросом является Камчатская область, где максимальный средний возраст автомобилей (20,6). Это может быть связано с тем, что цены на новые автомобили в этом регионе очень высокие. Этот регион не исключен из исходных данных, так как он необходим для целостности картины в рамках исследования, чтобы отобразить ситуацию на Дальнем Востоке. При этом Камчатская область проходит по критерию протяженности дорожного полотна, тогда как другой регион Дальнего Востока, Чукотский автономный округ, был все же исключен, так как не имеет достаточной площади дорог.

### Количество автомобилей на 1000 человек

Независимая переменная «количество автомобилей на 1000 человек» рассчитана как количество автомобилей в регионах, деленное на 1000, и измеряется в шт/1000 человек.

Графическое изображение данных показывает, что существует небольшой положительный эксцесс и асимметрия влево.



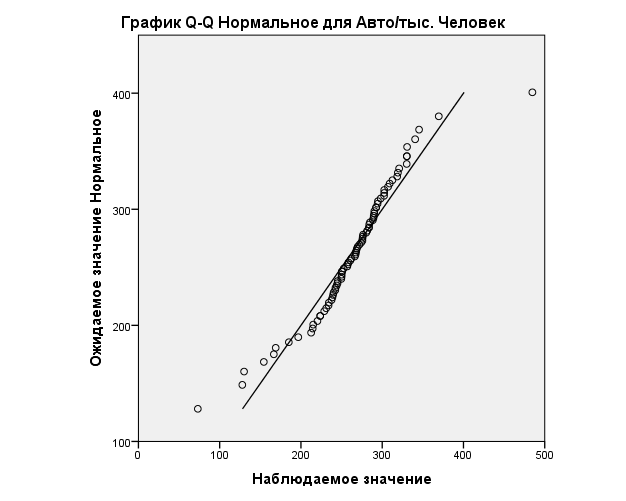
Расчет коэффициентов асимметрии и эксцесса и тест Колмогорова-Смирнова показали, что данные распределены нормально с небольшой островершинностью. Кроме того, есть такая же незначительная асимметрия влево (таблица 19 и 20).

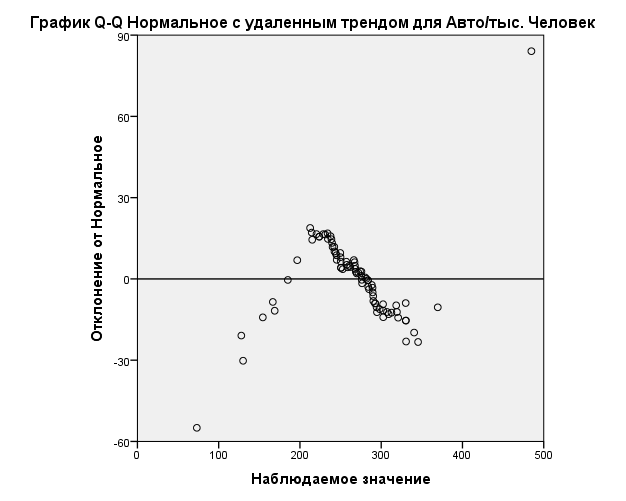
|  |  |  |
| --- | --- | --- |
| **Статистика** | | |
| Авто/тыс. Человек | | |
| N | Допустимо | 83 |
| Пропущенные | 30 |
| Асимметрия | | -,114 |
| Стандартная Ошибка асимметрии | | ,264 |
| Эксцесс | | 3,692 |
| Стандартная ошибка эксцесса | | ,523 |

**Таблица 19 «Коэффициенты асимметрии и эксцесса для переменной «количество автомобилей на 1000 человек»»**

|  |  |  |
| --- | --- | --- |
| **Одновыборочный критерий Колмогорова-Смирнова** | | |
|  | | Авто/тыс. Человек |
| N | | 83 |
| Параметры нормального распределенияa,b | Среднее значение | 264,404 |
| Среднеквадратичная отклонения | 56,0521 |
| Наибольшие экстремальные расхождения | Абсолютная | ,101 |
| Положительные | ,080 |
| Отрицательные | -,101 |
| Статистика критерия | | ,101 |
| Асимптотическая значимость (2-сторонняя) | | ,036c |
| a. Проверяемое распределение является нормальным. | | |
| b. Вычислено из данных. | | |
| c. Коррекция значимости Лильефорса. | | |

**Таблица 20 «Тест Колмогорова-Смирнова для переменной «количество автомобилей на 1000 человек»»**

График Q-Q также свидетельствует о том, что распределение нормально. 



Предположение о незначительной асимметрии влево подтверждается разностью среднего и медианы (медиана на 3,696 больше среднего). Данные о разбросе следующие: дисперсия равна 3141,835, а стандартная отклонения – 56,0521.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Описательные статистики** | | | | |
|  | | | Статистика | Стандартная Ошибка |
| Авто/тыс. Человек | Среднее значение | | 264,404 | 6,1525 |
| 95% Доверительный интервал для среднего | Нижняя граница | 252,164 |  |
| Верхняя граница | 276,643 |  |
| Среднее по выборке, усеченной на 5% | | 265,670 |  |
| Медиана | | 268,100 |  |
| Дисперсия | | 3141,835 |  |
| Стандартная отклонения | | 56,0521 |  |
| Минимум | | 73,1 |  |
| Максимум | | 484,8 |  |
| Диапазон | | 411,7 |  |
| Межквартильный диапазон | | 53,0 |  |
| Асимметрия | | -,114 | ,264 |
| Эксцесс | | 3,692 | ,523 |

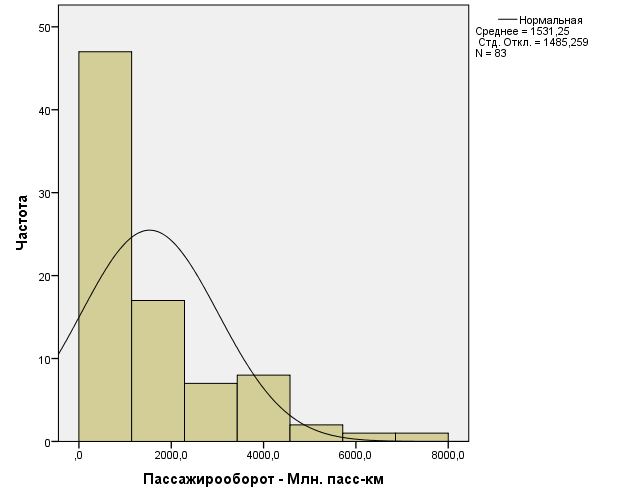
**Таблица 21 «Описательные статистики для переменной «количество автомобилей на 1000 человек»»**

По критерию Райта выбросами являются Камчатская область, где больше всего автомобилей на 1000 человек, и Чукотский автономный округ, где минимальное количество автомобилей. При этом Камчатская область проходит по критерию протяженности дорожного полотна, поэтому не исключается из исследования, тогда как другой регион Дальнего Востока, Чукотский автономный округ, был все же исключен, так как не имеет достаточной площади дорог.

### Пассажирооборот

Независимая переменная «пассажирооборот» представляет собой перевозки пассажиров и пассажирооборот автобусов общего пользования и измеряется в миллионах пассажирокиллометров.

На основании гистограммы можно сделать предположение о наличии асимметрии вправо и незначительном положительном эксцессе.



Предположения подтверждаются коэффициентом асимметрии и эксцесса: действительно имеет место асимметрия вправо, а эксцесс положителен и незначителен. Кроме того, распределение данных не является нормальным по тесту Колмогорова-Смирнова (таблица 22 и 23).

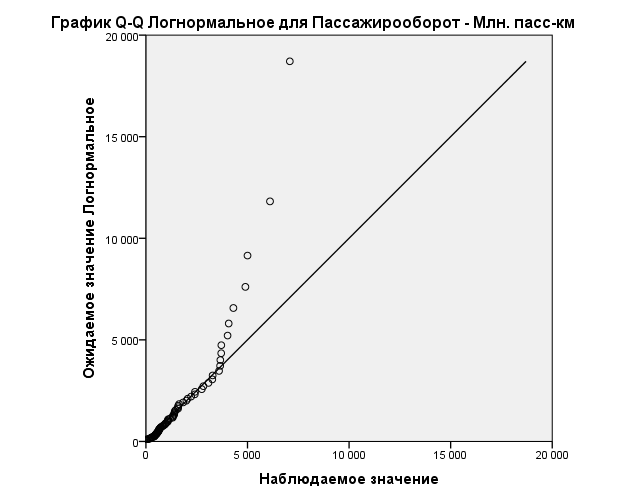
|  |  |  |
| --- | --- | --- |
| **Статистика** | | |
| Пассажирооборот – Млн. пасс-км | | |
| N | Допустимо | 83 |
| Пропущенные | 30 |
| Асимметрия | | 1,579 |
| Стандартная Ошибка асимметрии | | ,264 |
| Эксцесс | | 2,278 |
| Стандартная ошибка эксцесса | | ,523 |

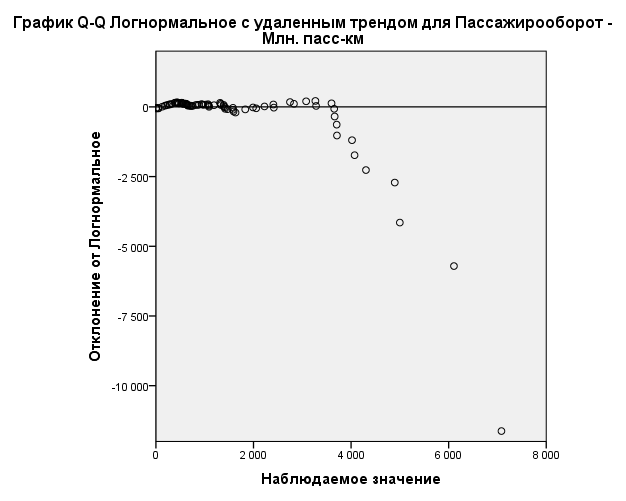
**Таблица 22 «Коэффициенты асимметрии и эксцесса для переменной «пассажирооборот»»**

|  |  |  |
| --- | --- | --- |
| **Одновыборочный критерий Колмогорова-Смирнова** | | |
|  | | Пассажирооборот – Млн. пасс-км |
| N | | 83 |
| Параметры нормального распределенияa,b | Среднее значение | 1531,253 |
| Среднеквадратичная отклонения | 1485,2590 |
| Наибольшие экстремальные расхождения | Абсолютная | ,196 |
| Положительные | ,196 |
| Отрицательные | -,152 |
| Статистика критерия | | ,196 |
| Асимптотическая значимость (2-сторонняя) | | ,000c |
| a. Проверяемое распределение является нормальным. | | |
| b. Вычислено из данных. | | |
| c. Коррекция значимости Лильефорса. | | |

**Таблица 23 «Тест Колмогорова-Смирнова для переменной «пассажирооборот»»**

По графику Q-Q было определено, что распределение скорее является логнормальным. Таким образом, если данная переменная будет входить в формулу расчета числа ДТП, то под математической операцией логарифм.





Значительная асимметрия вправо подтверждается разницей среднего и медианы (557,253). Дисперсия равна 2205994,167, а стандартная отклонения – 1485,2590.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Описательные статистики** | | | | |
|  | | | Статистика | Стандартная Ошибка |
| Пассажирооборот – Млн. пасс-км | Среднее значение | | 1531,253 | 163,0284 |
| 95% Доверительный интервал для среднего | Нижняя граница | 1206,938 |  |
| Верхняя граница | 1855,568 |  |
| Среднее по выборке, усеченной на 5% | | 1381,900 |  |
| Медиана | | 974,000 |  |
| Дисперсия | | 2205994,167 |  |
| Стандартная отклонения | | 1485,2590 |  |
| Минимум | | 7,0 |  |
| Максимум | | 7081,0 |  |
| Диапазон | | 7074,0 |  |
| Межквартильный диапазон | | 1520,0 |  |
| Асимметрия | | 1,579 | ,264 |
| Эксцесс | | 2,278 | ,523 |

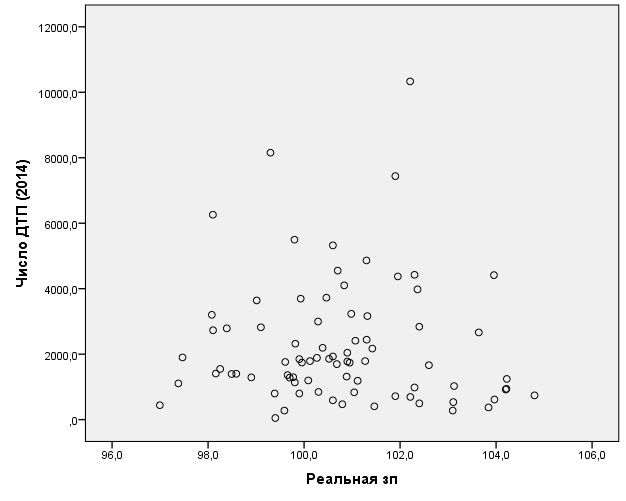
**Таблица 24 «Описательные статистики для переменной «пассажирооборот»»**

По критерию Райта выбросов по данной переменной нет.

## Анализ статистической связи

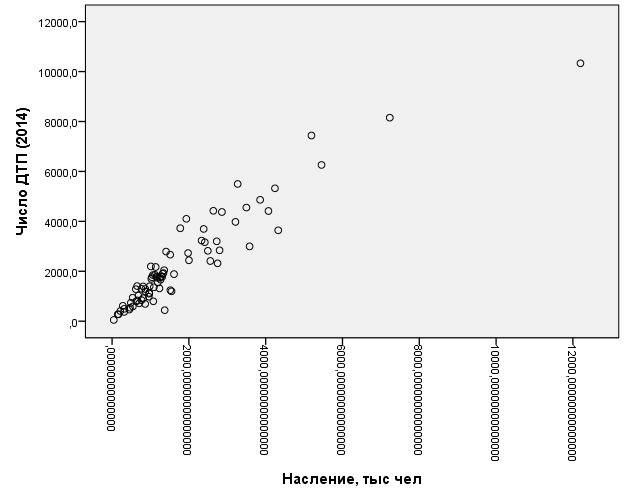
Для каждой пары «числовая зависимая переменная – независимая переменная» был проведен графический анализ (Scatter или Box-Whisker), были построены анализы – ANOVA или корреляционные матрицы с оценкой значимости.

### 2.2.1 Диаграмма рассеивания Реальная заработная плата и число ДТП



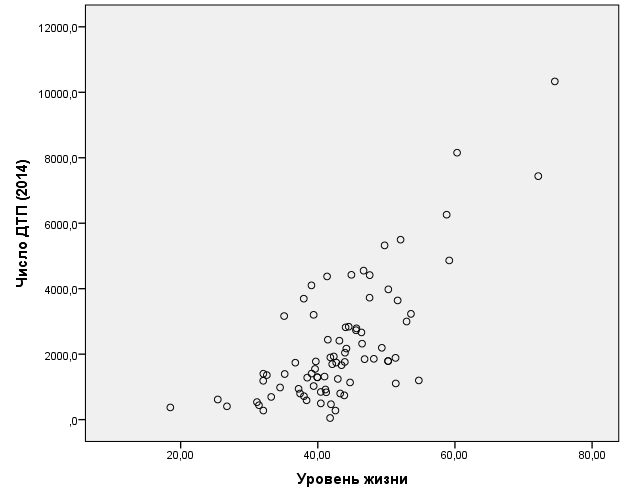
Тренд не наблюдаем, зависимость неочевидна. Исходя из гипотезы, более низкая реальная ЗП должна была вести к большему числу ДТП.

### 2.2.2. Диаграмма рассеивания Население и число ДТП



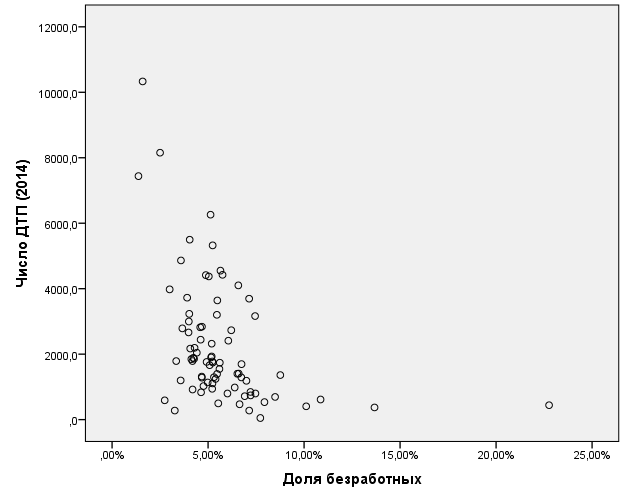
Тренд наблюдаем, зависимость линейна. Большее количество людей в регионе приводит к большему количеству транспортных средств, следовательно, и к большему числу ДТП. Данный тренд был ожидаем.

### 2.2.3. Диаграмма рассеивания Рейтинговый балл по уровню жизни и число ДТП



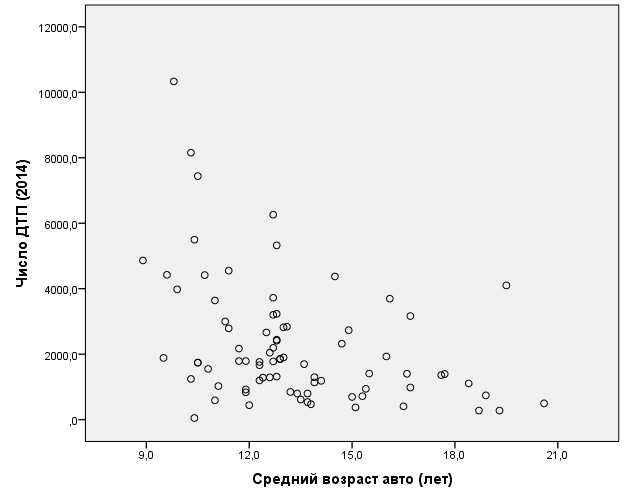
Тренд наблюдаем, зависимость, скорее всего, линейна. Высокий уровень жизни чаще всего в городах с большим количеством автомобилей, поэтому число ДТП в них также больше. Данный тренд не был ожидаем, выдвигалась гипотеза о том, что чем более низок уровень жизни, экономическое состояние региона, тем больше водители склонны к неаккуратной езде.

### 2.2.4. Диаграмма рассеивания Доля безработных и число ДТП



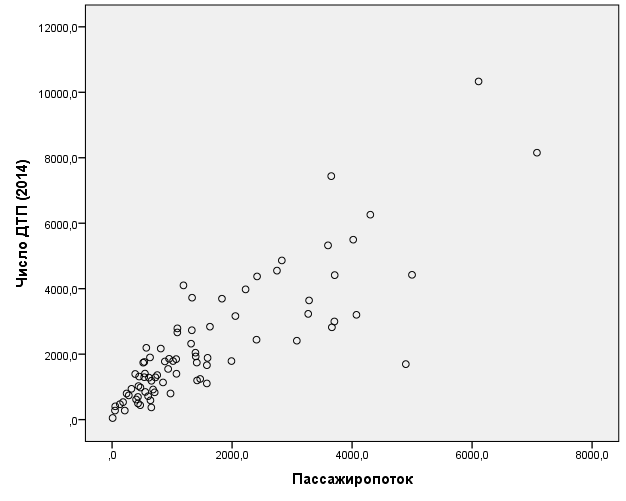
Тренд наблюдаем, зависимость нелинейна. Чем выше доля безработных, тем меньше в регионе наблюдаем тренд к покупке и вождению автомобиля, тем более неаккуратному. Данный тренд не был ожидаем, выдвигалась гипотеза о том, что чем ниже уровень жизни и экономическое состояние региона, тем больше водители склонны к неаккуратной езде. Нелинейность означает некоторую специфику вхождения переменной в модель, а именно: вероятнее всего, доля безработных будет входить в формулу расчета числа ДТП возведенная в отрицательную степень.

### 2.2.5. Диаграмма рассеивания Средний возраст автомобиля и число ДТП



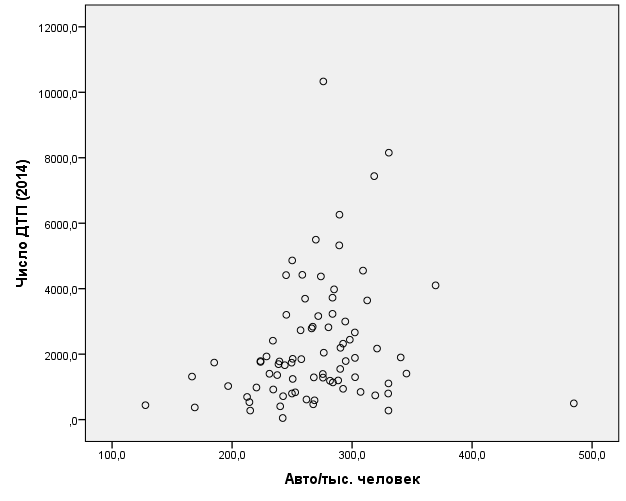
Тренд наблюдаем, зависимость нелинейна. Чем выше средний возраст автомобиля, тем меньше регион склонен к покупке новых машин, следовательно, уровень интеграции транспорта в жизнь жителей ниже, следовательно, и ниже число ДТП. В рамках данного анализа нелинейность означает специфику вхождения переменной в расчетную формулу. Вероятнее всего, средний возраст автомобиля будет возведен в отрицательную степень для расчета числа ДТП.

### 2.2.6. Диаграмма рассеивания Пассажиропоток и число ДТП



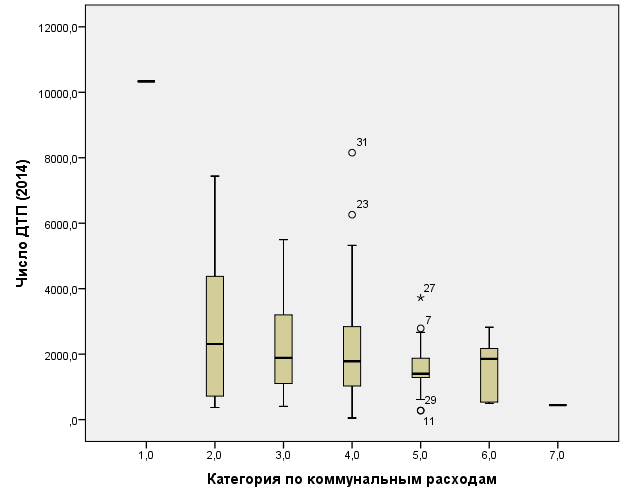
Тренд наблюдаем, зависимость линейна. Чем выше мобильность пассажиров, тем выше пассажиропоток в регионе, что приводит к частому использованию транспорта, следовательно, и к ДТП. Из линейной зависимости можем сделать предположение, что пассажиропоток будет прямопропорционально, линейно влиять на число ДТП.

### 2.2.7. Диаграмма рассеивания Количество автомобилей на 1000 человек и число ДТП



Тренд наблюдаем, зависимость нелинейна. Связь переменных идет через количество используемых машин. Чем выше их число, тем больше ДТП. В рамках данного анализа нелинейность означает специфику вхождения переменной в расчетную формулу. Вероятнее всего, переменная Число автомобилей на 1000 человек будет возведена в положительную (более единицы) степень для расчета числа ДТП.

### 2.2.8. Диаграмма Box-Whiskers для Коммунальных расходов и числа ДТП



Среднее во всех группах приблизительно на одном уровне. При этом размеры коробки различны. Это означает, что во всех категориях по коммунальным расходам среднее число ДТП одинаково, разница только в разбросе значений выше и ниже среднего значения. Теоретически, число групп можно было бы сократить, однако при этом возникнет неоднородность между группами, заключающаяся в числе элементов, входящих в каждую категорию.

### 2.2.9. Корреляционные матрицы и анализ ANOVA

Для всех числовых переменных была построена корреляционная матрица с оценкой значимости, основанной на критерии Пирсона и Спирмана. Результаты приведены в таблицах ниже.

|  |  |  |
| --- | --- | --- |
|  | | Число ДТП (2014) |
| Число ДТП (2014) | Корреляция Пирсона | 1 |
| Знач. (двухсторонняя) |  |
| N | 80 |
| Реальная зп | Корреляция Пирсона | -,046 |
| Знач. (двухсторонняя) | ,686 |
| N | 80 |
| Уровень жизни | Корреляция Пирсона | ,741\*\* |
| Знач. (двухсторонняя) | ,000 |
| N | 80 |
| Насление, тыс чел | Корреляция Пирсона | ,935\*\* |
| Знач. (двухсторонняя) | ,000 |
| N | 80 |
| Средний возраст авто (лет) | Корреляция Пирсона | -,390\*\* |
| Знач. (двухсторонняя) | ,000 |
| N | 80 |
| Авто/тыс. человек | Корреляция Пирсона | ,213 |
| Знач. (двухсторонняя) | ,058 |
| N | 80 |
| Пассажиропоток | Корреляция Пирсона | ,841\*\* |
| Знач. (двухсторонняя) | ,000 |
| N | 80 |
| Доля безработных | Корреляция Пирсона | -,420\*\* |
| Знач. (двухсторонняя) | ,000 |
| N | 80 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  |  |  | | --- | --- | --- | --- | | **Корреляции** | | | | |  | | | Число ДТП (2014) | | Ро Спирмана | Число ДТП (2014) | Коэффициент корреляции | 1,000 | | Знач. (2-х сторонняя) | . | | N | 80 | | Реальная зп | Коэффициент корреляции | -,109 | | Знач. (2-х сторонняя) | ,336 | | N | 80 | | Уровень жизни | Коэффициент корреляции | ,644\*\* | | Знач. (2-х сторонняя) | ,000 | | N | 80 | | Насление, тыс чел | Коэффициент корреляции | ,914\*\* | | Знач. (2-х сторонняя) | ,000 | | N | 80 | | Средний возраст авто (лет) | Коэффициент корреляции | -,400\*\* | | Знач. (2-х сторонняя) | ,000 | | N | 80 | | Авто/тыс. человек | Коэффициент корреляции | ,283\* | | Знач. (2-х сторонняя) | ,011 | | N | 80 | | Пассажиропоток | Коэффициент корреляции | ,842\*\* | | Знач. (2-х сторонняя) | ,000 | | N | 80 | | Доля безработных | Коэффициент корреляции | -,484\*\* | | Знач. (2-х сторонняя) | ,000 | | N | 80 | | | \*. Корреляция значима на уровне 0,05 (двухсторонняя). |   \*\*. Корреляция значима на уровне 0,01 (двухсторонняя). |

Из полученных результатов можно сказать о значимости значимости для числа ДТП в регионе следующих переменных: рейтинговая оценка уровня жизни, население, доля безработных, средний возраст автомобиля по региону, пассажиропоток. Для всех перечисленных переменных уровень значимости меньше 0.01 и по критерию Спирмана, и по критерию Пирсона, что свидетельствует о наличии зависимости между переменными.

При этом, в результате построения корреляционных матриц, удалось установить тесную корреляцию переменной Уровень жизни с переменными Население, Средний возраст автомобиля, Количество автомобилей на 1000 человек, Пассажиропоток и Доля безработных. Население коррелирует с переменными Средний возраст автомобиля, Пассажиропоток и Доля безработных. Переменная Средний возраст автомобиля коррелирует с переменной Количество автомобилей на 1000 человек и с Пассажиропотоком, однако Количество автомобилей на 1000 человек не имеет высокой степени корреляции с переменной Число ДТП. Переменная Пассажиропоток коррелирует с Долей безработных.

Таким образом, в дальнейшем анализе, в результате исключение мультиколлинеарности, будут использоваться числовые переменные: Население, Доля безработных, Средний возраст автомобиля.

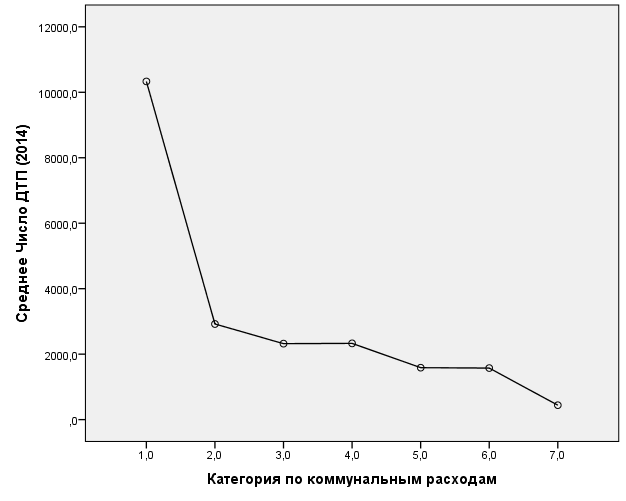
Далее, для номинальных переменных был проведен ANOVA – анализ.

Прежде всего, была рассмотрена переменная категории по коммунальным расходам (удельной части коммунальных расходов).

|  |  |  |  |
| --- | --- | --- | --- |
| **Критерий однородности дисперсий** | | | |
| Число ДТП | | | |
| Статистика Ливиня | ст.св.1 | ст.св.2 | Знач. |
| 3,401a | 4 | 73 | ,013 |

|  |
| --- |
| a. Группы с единственным наблюдением не учитываются при вычислении критерия однородности дисперсий для Число ДТП (2014). |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ANOVA** | | | | | |
| Число ДТП (2014) | | | | | |
|  | Сумма квадратов | ст.св. | Средний квадрат | F | Знач. |
| Между группами | 80892809,833 | 6 | 13482134,972 | 4,907 | ,000 |
| Внутри групп | 200588582,917 | 73 | 2747788,807 |  |  |
| Всего | 281481392,750 | 79 |  |  |  |
|  |  |  |  |  |  |

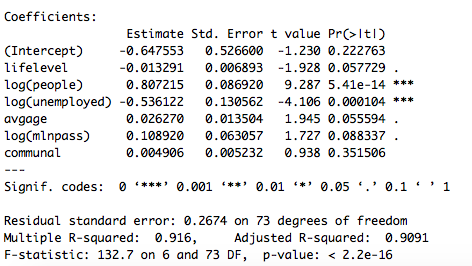


В результате однофакторного дисперсионного анализа удалось выяснить, что значимыми являются переменные категории коммунальных расходов, так как их уровень значимости 0 из ANOVA-анализа.

Таким образом, результатом анализа статистической связи стало определение наличия или отсутствия связи между числом дорожно-транспортных происшествий и некоторыми из описанных в первом пункте переменными. Например, зависимость между числом ДТП и населением в регионе линейна и прямопропоциональна так же, как и для переменных рейтинговой оценки уровня жизни и пассажиропотока. Для доли безработных и среднего возраста автомобиля связь с числом ДТП оказалась нелинейна и обратнонаправлена. Зависимость между количеством автомобилей на 1000 человек в регионе и числом ДТП нелинейна, однако находится в прямой, а не в обратной связи. Связь реальная заработная плата и число ДТП связаны нелинейно, кроме того, тренд не очевиден из диаграммы рассеивания. Более того, построение корреляционных матриц показало, что реальная заработная плата и число ДТП слабо коррелируют. Для остальных переменных корреляция присутствует со значимостью 0,000. Для номинальной переменной (удельная часть расходов на коммунальные платежи) был проведен ANOVA-анализ, который показал связь между данной переменной и числом ДТП. В результате данных операций нам удалось определить переменные, которые влияют на число ДТП и будут формировать данное значение.

# Линейная регрессия

По результатам произведённого исследования было предположено, что логарифм количества ДТП (так как данная переменная распределена логнормально) зависит от логарифма населения, логарифма уровня безработицы в регионе, количества автомобилей на 1000 населения, стоимости коммунальных услуг, уровня жизни, логарифма пассажиропотока, средней зарплаты в регионе и среднего возраста автомобилей. То есть формула будет выглядеть следующим образом:

В результате построения линейной регрессии в программной среде R studio были исключены незначимые переменные и получены следующие коэффициенты: (см. изображение), в результате формула выглядит следующим образом

# Эссе по найденным публикациям

Существует множество исследований, посвященных зависимости количества ДТП от времени суток, погоды, освещения и прочих параметров, то есть обращающих пристальное внимание на непосредственно предшествующие каждому происшествию события и влияние их комбинации на вероятность ДТП. В то же самое время исследований, подобных нашему, то есть учитывающих социально-экономические параметры, такие как средние доходы населения, уровень безработицы, средний возраст автомобиля в регионе и проч, найдено не было. Такие исследования не призваны устранить конкретные недостатки дорожного покрытия или каких-либо других отдельных параметров дорог или автомобилей, они нужны для выявления более ранних причин большего или меньшего количества ДТП.

Основной целью исследований, найденных нами на просторах Интернета, являлось выявление параметров, влияющих на количество ДТП в определенный день, что позволяет вовремя предупреждать экстренные службы в случае определенных погодных или иных условий, а также рекомендовать водителям соблюдать повышенную осторожность, например, в случае снегопада. В свою очередь, исследование влияния социально-экономических параметров на количество ДТП может помочь в решении проблемы происшествий на дорогах с помощью воздействия на значимые параметры. Например, если выяснится, что количество ДТП зависит от среднего возраста автомобилей в регионе, можно будет запустить программу утилизации, чтобы мотивировать людей обновлять автопарк.

В идеальном случае модель должна объяснять относительный показатель ДТП на 1000 человек через остальные параметры, однако на текущий момент мы столкнулись со сложностями, так как при подсчете R2 для относительных параметров, получилось число, меньшее 0.35, тогда как при использовании абсолютных показателей R2 повышается до 0.89. Это может быть объяснено тем, что количество ДТП, очевидно, напрямую зависит от населения региона.

# Спецификация базовой модели

## Связь структуры базовой модели с основными гипотезами исследования.

Основными гипотезами, сформулированными в рамках решения прикладной задачи, являются:

1. Количество ДТП в регионах прямо пропорционально зависит от численности населения и количества автомобилей на тысячу человек;
2. Число ДТП в регионах находится в зависимости от рейтингового балла качества жизни. А именно, чем выше балл, тем меньше количество ДТП, что может объясняться тем, что при более высоком уровне благосостояния семей в регионе жители покупают более дорогие автомобили и более трепетно к ним относятся, что отражается на их стиле вождения;
3. Число ДТП в регионах имеет прямую зависимость от таких экономических показателей региона, как прирост реальной заработной платы и уровень безработицы. Чем выше уровень безработицы и меньше прирост реальной заработной платы, тем меньше средств у жителей для поддержания автомобиля на должном уровне и, как результат, тем больше количество ДТП;
4. Уровень безработицы имеет прямое влияние на средний возраст автомобилей, и в совокупности переменные при своем увеличении также увеличивают число ДТП.

Перечисленные гипотезы можно преобразовать в статистические для их дальнейшей проверки. Так, на языке статистики они будут звучать следующим образом:

1. Переменная «население» является значимой для итоговой модели, и ее коэффициент будет положительным;
2. Переменная «авто на тыс. человек» является значимой для итоговой модели, и ее коэффициент будет положительным;
3. Переменная «уровень жизни» значима для итоговой модели, и ее коэффициент будет отрицательным;
4. Переменная «уровень безработицы» является значимой для итоговой модели, и ее коэффициент будет положительным;
5. Переменная «прирост реальной зп» является значимой для итоговой модели, и ее коэффициент будет отрицательным;
6. В модели будет присутствовать корреляция между переменными «средний возраст авто» и «уровень безработицы», то есть коэффициент корреляции будет близок к 1.

## Связь структуры базовой модели с результатами предварительного анализа данных

В рамках предварительного анализа данных были определены функциональная форма вхождения независимых переменных в итоговую модель. В ходе построения линейной регрессии были получены действительные коэффициенты и уровень значимости каждой переменной. Таким образом, существует возможность сравнения структуры базовой модели с результатами предварительного анализа данных.

Так, при помощи анализа наличия статистической взаимосвязи между зависимой переменной (число ДТП) и независимыми параметрами по результатам предварительного анализа было получено, что связь числа ДТП и:

* Реальной Заработной платы – отсутствует
* Доли безработных – нелинейная, обратная
* Рейтинговый балл региона по уровню жизни – линейная, прямая
* Население – линейная, прямая
* Возраст авто – нелинейная, прямая
* Пассажирооборот – линейная, прямая
* Авто / тыс. Человек – нелинейная, обратная.

В результате построения корреляционных таблиц и регрессии оказалось, что между реальной заработной платой и числом ДТП нет зависимости, точно так же, как и у переменной авто/ тыс. человек. В случае с переменными население, пассажирооборот– зависимость прямая, линейная, как и было предположено в ходе первичного анализа данных. Уровень безработицы и уровень жизни имеют обратную зависимость относительно числа ДТП. В случае с переменной средний возраст авто в ходе предварительного анализа было предположено, что зависимость обратная, однако, оказалось, что более низкий средний возраст авто говорит о низком уровне числа ДТП. Таким образом, большинство выводов, сделанных в ходе первичного анализа данных, оказались верными.

Также, в ходе анализа базовых данных было определено, что некоторое число переменных распределено скорее логнормально, чем нормально. Действительно, при использовании некоторых данных под знаком логарифма была получена более высокая степень достоверности модели.

# Оценивание базовой модели

## Проверка мультиколлинеарности

### 6.1.1 VIF-анализ

В рамках проверки модели на мультиколлинеарность были проведены некоторые анализы. В частности, VIF-анализ, позволяющий определить корреляцию между переменными, которые должны быть независимы. Для каждой пары переменных был получен коэффициент VIF, который ни для одной из пар не превысил границы в 10. Однако, для переменной Население VIF поднимался на уровень около 8. Следовательно, по результатам данного анализа не было выявлено присутствия мультиколлинеарности в модели. Результаты анализа для каждой пары переменных представлены ниже.

|  |  |  |  |
| --- | --- | --- | --- |
| **Coefficientsa** | | | |
| Model | | Collinearity Statistics | |
| Tolerance | VIF |
| 1 | Пассажирооборот - Млн. пасс-км | .301 | 3.318 |
| Насление, тыс чел | .231 | 4.334 |
| Рейтинговый балл региона по уровню жизни | .260 | 3.840 |
| Доля безработных | .526 | 1.899 |
| Средний возраст автомобиля | .753 | 1.328 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| a. Dependent Variable: 1 Категория по коммунальным расходам | | | | |
| **Coefficientsa** | | | | |
| Model | | | Collinearity Statistics | |
| Tolerance | VIF |
| 1 | | 1 Категория по коммунальным расходам | .423 | 2.363 |
| Пассажирооборот - Млн. пасс-км | .237 | 4.228 |
| Насление, тыс чел | .120 | 8.340 |
| Рейтинговый балл региона по уровню жизни | .271 | 3.687 |
| Доля безработных | .525 | 1.905 |

|  |
| --- |
| a. Dependent Variable: Средний возраст автомобиля |

|  |  |  |  |
| --- | --- | --- | --- |
| **Coefficientsa** | | | |
| Model | | Collinearity Statistics | |
| Tolerance | VIF |
| 1 | Доля безработных | .857 | 1.167 |
| Средний возраст автомобиля | .794 | 1.259 |
| 1 Категория по коммунальным расходам | .429 | 2.333 |
| Пассажирооборот - Млн. пасс-км | .235 | 4.258 |
| Насление, тыс чел | .148 | 6.744 |

|  |
| --- |
| a. Dependent Variable: Рейтинговый балл региона по уровню жизни |

|  |  |  |  |
| --- | --- | --- | --- |
| **Coefficientsa** | | | |
| Model | | Collinearity Statistics | |
| Tolerance | VIF |
| 1 | Рейтинговый балл региона по уровню жизни | .317 | 3.154 |
| Доля безработных | .549 | 1.822 |
| Средний возраст автомобиля | .751 | 1.332 |
| 1 Категория по коммунальным расходам | .812 | 1.232 |
| Пассажирооборот - Млн. пасс-км | .551 | 1.813 |

|  |
| --- |
| a. Dependent Variable: Насление, тыс чел |

|  |  |  |  |
| --- | --- | --- | --- |
| **Coefficientsa** | | | |
| Model | | Collinearity Statistics | |
| Tolerance | VIF |
| 1 | Насление, тыс чел | .282 | 3.542 |
| Рейтинговый балл региона по уровню жизни | .257 | 3.890 |
| Доля безработных | .527 | 1.896 |
| Средний возраст автомобиля | .758 | 1.319 |
| 1 Категория по коммунальным расходам | .543 | 1.842 |

|  |
| --- |
| a. Dependent Variable: Пассажирооборот - Млн. пасс-км |

|  |  |  |  |
| --- | --- | --- | --- |
| **Coefficientsa** | | | |
| Model | | Collinearity Statistics | |
| Tolerance | VIF |
| 1 | Средний возраст автомобиля | .751 | 1.331 |
| 1 Категория по коммунальным расходам | .424 | 2.361 |
| Пассажирооборот - Млн. пасс-км | .235 | 4.247 |
| Насление, тыс чел | .125 | 7.972 |
| Рейтинговый балл региона по уровню жизни | .419 | 2.388 |

|  |
| --- |
| a. Dependent Variable: Доля безработных |

### 6.1.2 Индекс обусловленности

Кроме VIF-анализа был подсчитан индекс обусловленности. Первоначальный вектор собственных значений выглядит следующим образом:

Индекс обусловленности (*k*) – имеет следующее значение:

Так как k < 30, то был сделан вывод, что в нашем случае присутствует слабая мультиколлинеарность. Для снижения степени выявленной мультиколлинеарности переменная «население» может быть исключена из анализа, так как по методу VIF ее коэффициент поднимается до уровня 8.

## Проверка гетероскедастичности

### 6.2.1 Проверка гетероскедастичности произвольной формы. Тест Breusch-Pagan

После построения регрессии и нахождения коэффициентов были найдены следующие значения коэффициентов в регрессии :

|  |  |  |  |
| --- | --- | --- | --- |
|  | Свободный  коэффициент | 0,268396457 |  |
|  | Население | -0,002561201 |  |
|  | Уровень жизни | -0,000227469 |  |
|  | Безработица | 0,045837844 |  |
|  | Средний возраст авто | -0,003170762 |  |
|  | 1 категория по комм  расх | -0,013989086 |  |
|  | Пассажирооборот | 0,000866203 |  |

Значение получилось равным 21,67, вычисление дало в результате 1, что говорит об отсутствии необходимости отвергать основную гипотезу. Иными словами, гетероскедастичности нет.

### 6.2.2 Проверка гетероскедастичности по конкретной переменной. Тест Park

Для проведения теста были выбраны несколько переменных: «уровень жизни», «уровень безработицы», «пассажирооборот» и «средний возраст авто», чтобы исследовать их отдельное влияние на гетероскедастичность. Для каждой из этих переменных была построена регрессия , результаты которой следующие:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Коэффициентыa** | | | | | | |
| Модель | | Нестандартизованные коэффициенты | | Стандартизованные коэффициенты | т | Знач. |
| B | Стандартная Ошибка | Бета |
| 1 | (Константа) | -5,771 | 1,784 |  | -3,236 | ,002 |
| Население | ,227 | ,250 | ,103 | ,908 | ,367 |
| a. Зависимая переменная: LN(e^2) | | | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Коэффициентыa** | | | | | | |
| Модель | | Нестандартизованные коэффициенты | | Стандартизованные коэффициенты | т | Знач. |
| B | Стандартная Ошибка | Бета |
| 1 | (Константа) | -4,139 | 4,123 |  | -1,004 | ,319 |
| Рейтинговый балл региона по уровню жизни | -,007 | 1,102 | -,001 | -,006 | ,995 |
| a. Зависимая переменная: LN(e^2) | | | | | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | **Коэффициентыa** | | | | | | | | Модель | | Нестандартизованные коэффициенты | | Стандартизованные коэффициенты | т | Знач. | | B | Стандартная Ошибка | Бета | | 1 | (Константа) | -3,296 | 1,750 |  | -1,884 | ,063 | | Безработица | ,296 | ,592 | ,057 | ,500 | ,618 | | a. Зависимая переменная: LN(e^2) | | | | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Коэффициентыa** | | | | | | |
| Модель | | Нестандартизованные коэффициенты | | Стандартизованные коэффициенты | т | Знач. |
| B | Стандартная Ошибка | Бета |
| 1 | (Константа) | -3,538 | 1,169 |  | -3,026 | ,003 |
| Средний возраст автомобиля | -,047 | ,086 | -,062 | -,546 | ,587 |
| a. Зависимая переменная: LN(e^2) | | | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Коэффициентыa** | | | | | | |
| Модель | | Нестандартизованные коэффициенты | | Стандартизованные коэффициенты | т | Знач. |
| B | Стандартная Ошибка | Бета |
| 1 | (Константа) | -5,098 | 1,364 |  | -3,737 | ,000 |
| Пассажирооборот | ,136 | ,196 | ,079 | ,694 | ,490 |
| a. Зависимая переменная: LN(e^2) | | | | | | |

Коэффициенты значимости получились следующие:

|  |  |
| --- | --- |
| **Переменная** | **Значимость** |
| Уровень жизни | 0,995 |
| Уровень безработицы | 0,618 |
| Пассажирооборот | 0,49 |
| Средний возраст авто | 0,587 |
| Население | 0,367 |

Таким образом, ни по какой из переменных не была выявлена значимость для регрессии и, как итог, гетероскедастичность.

### 6.2.3 Тест Goldfeld-Quandt

Ключевым фактором в рамках данного теста была выбрана переменная «пассажирооборот». После деления данных на три группы и выброса средней были построены линейные регрессии по первым 30 записям данных и по последним 30.   
В 1 группе получились следующие коэффициенты:

|  |  |
| --- | --- |
| Свободный коэффициент | -1,369907834 |
| Население | 0,893461478 |
| Уровень жизни | -0,001527003 |
| Безработица | -0,560938688 |
| Средний возраст авто | 0,044144085 |
| 1 категория по комм расх | 0 |
| Пассажирооборот | 0,039110234 |

При этом =0,880563449, а = 1,798042131.

Во 2 группе получились следующие коэффициенты:

|  |  |
| --- | --- |
| Свободный коэффициент | -0,868695964 |
| Население | 0,773165563 |
| Уровень жизни | -0,00620069 |
| Безработица | -0,196932461 |
| Средний возраст авто | 0,008013572 |
| 1 категория по комм расх | -0,24206126 |
| Пассажирооборот | 0,303680386 |

в этом случае равен 0,898715032, а = 0,920891946.

Получается , и по распределению Фишера

Так как , то нет оснований отвергать основную гипотезу, а значит, нет гетероскедастичности.

## Проверка гипотез

Для проверки гипотез использовались разные уровни значимости – 5%, 10% и 1%. В итоге получились следующие результаты:

1. Переменная «население» является значимой для итоговой модели, и ее коэффициент будет положительным; Рассчитанная значимость – 1,32619E-14, коэффициент – 0,835235067, как результат, при всех уровнях значимости нет оснований отвергать гипотезу.
2. Переменная «авто на тыс. человек» является значимой для итоговой модели, и ее коэффициент будет положительным; Рассчитанная значимость – 0,058919333, коэффициент – 0,001670283, как результат, при всех уровнях значимости гипотеза отвергается.
3. Переменная «уровень жизни» значима для итоговой модели и ее коэффициент будет отрицательным; Рассчитанная значимость – -0,016887837, коэффициент – 0,021886143, как результат, при всех уровнях значимости нет оснований отвергать гипотезу.
4. Переменная «уровень безработицы» является значимой для итоговой модели, и ее коэффициент будет положительным; Рассчитанная значимость – 0,000473664, коэффициент – -0,491699374, как результат, при всех уровнях значимости нет оснований отвергать гипотезу.
5. Переменная «прирост реальной заработной платы» является значимой для итоговой модели, и ее коэффициент будет отрицательным; Рассчитанная значимость – 0,91104928, коэффициент – -0,001967606, как результат, при всех уровнях значимости гипотеза отвергается.
6. В модели будет присутствовать корреляция между переменными «средний возраст авто» и «уровень безработицы», коэффициент корреляции будет близок к 1.

Корреляция между переменными по критерию Пирсона равна 0,409, по критерию Спирмана равна 0,529. Как результат, основная гипотеза отвергается.

# Оптимизация базовой модели

Базовая модель имеет следующие коэффициенты и :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Коэффициентыa** | | | | | | |
| Модель | | Нестандартизованные коэффициенты | | Стандартизованные коэффициенты | т | Знач. |
| B | Стандартная Ошибка | Бета |
| 1 | (Константа) | -,538 | ,464 |  | -1,159 | ,250 |
| LN Население | ,828 | ,088 | ,853 | 9,441 | ,000 |
| LN Безработица | -,544 | ,130 | -,242 | -4,178 | ,000 |
| LN пассажирооборот | ,090 | ,063 | ,116 | 1,426 | ,158 |
| Рейтинговый балл региона по уровню жизни | -,012 | ,007 | -,124 | -1,792 | ,077 |
| Средний возраст автомобиля | ,026 | ,013 | ,077 | 1,959 | ,054 |
| 1 Категория по коммунальным расходам | -,382 | ,297 | -,048 | -1,285 | ,203 |
| a. Зависимая переменная: LN число ДТП | | | | | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Сводка для модели** | | | | |
| Модель | R | R-квадрат | Скорректированный R-квадрат | Стандартная ошибка оценки |
| 1 | ,958a | ,917 | ,910 | ,266003747137536 |
| a. Предикторы: (константа), 1 Категория по коммунальным расходам, Средний возраст автомобиля, LN пассажирооборот, LN Безработные, Рейтинговый балл региона по уровню жизни, LN Население | | | | |

В качестве альтернатив для базовой модели были рассмотрены следующие варианты:

* Модель с логарифмами и всеми категориями регионов по коммунальным расходам (логарифм обусловлен логнормальным распределением параметров «уровень безработицы», «пассажирооборот», «население» и зависимой переменной «число ДТП»; номинальная переменная «категория региона по коммунальным расходам» используется в базовой модели не полностью вследствие проведенного ранее ANOVA-анализа)
* Модель с логарифмами и без населения (переменная может быть исключена из соображений о сокращении уровня мультиколлинеарности);
* Модель без логарифмов и со всеми категориями регионов по коммунальным расходам.

Для проведения оптимизации базовой модели использовались два метода:

* Расчет модифицированного коэффициента детерминации;
* Применение принципа экономии с помощью информационных критериев.

## 7.1 Модифицированный коэффициент детерминации

Модифицированный рассчитывался по формуле

Для базовой модели . Таким образом, равен 0,911391892. Получается, что в модели используются лишние факторы, которые при этом не влияют на ее эффективность. Именно из-за этих дополнительных переменных «оштрафовал» базовую модель на 0,005608108, снизив расчетное число с 0,917 до 0,911.

Для модели с с логарифмами и всеми 6 категориями регионов по коммунальным расходам , а «штраф» составляет 0,003918919.

Для модели с логарифмами и без населения , а , «штраф» составляет 0,012475441.

Для модели без логарифмов и со всеми 6 категориями регионов по коммунальным расходам , а «штраф» составляет 0,005337838.

Таким образом, наибольший коэффициент детерминации, модифицированный коэффициент детерминации и наименьший «штраф» имеет модель с логарифмами и со всеми категориями регионов по коммунальным платежам.

## 7.2 Информационные критерии

**Критерий Акаике** рассчитывался по формуле

где )), n – количество регрессоров, T – количество записей.

Для базовой модели критерий Акаике получился равным -0,146440068. Для базовой модели, дополненной всеми категориями регионов по коммунальным платежам, критерий Акаике составляет -0,323662143. Для базовой модели без населения критерий Акаике составляет 0,222986831. Для модели без натуральных логарифмов, но с полной номинальной переменной вычисленное число – 7,49025427.

**Критерий Шварца** рассчитывался по формуле

.

Для базовой модели он составляет 0,360863928, для модели со всеми категориями – 0,183641852, для модели без населения - 0,730290827, для модели без логарифмов и с полной номинальной переменной – 7,997558265.

Итоги проделанного анализа оптимизации модели представлены в следующей таблице:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Модель |  |  |  |  |
| Базовая модель |  | 0,9114 | -0,146440068 | 0,360863928 |
| Базовая модель + все категории по коммунальным платежам |  |  | **-0,323662143** | **0,183641852** |
| Базовая модель - население |  |  | 0,222986831 | 0,730290827 |
| Модель без логарифмов, но со всеми категориями по коммунальным платежам и населением |  |  | 7,49025427 | 7,997558265 |

Как результат, наиболее эффективной моделью (наибольшие , и наименьшие , ) является модель с логарифмами, населением и всеми категориями регионов по коммунальным платежам. Данный результат никак не противоречит поставленной задаче, так как позволяет описать влияние показателей на число ДТП наиболее полно в отличие от базовой модели, не имеющей всех категорий номинальной переменной «категория региона по коммунальным платежам» вследствие их незначимости по ANOVA-анализу. Полезность данной оптимизации при этом имеется, но не особо значительная, так как точность базовой модели достаточно высокая () и при оптимизации увеличивается на 0,025.

## 7.3 Проверка гипотез для оптимальной модели

Для проверки гипотез в рамках оптимальной модели вновь использовались разные уровни значимости – 5%, 10% и 1%. В итоге получились следующие результаты:

1. Переменная «население» является значимой для итоговой модели, и ее коэффициент будет положительным; Рассчитанная значимость – 1,2264E-17, коэффициент – 0,902993664639085, как результат, при всех уровнях значимости нет оснований отвергать гипотезу.
2. Переменная «авто на тыс. человек» является значимой для итоговой модели, и ее коэффициент будет положительным; Данная переменная была исключена из анализа в рамках других тестов, соответственно, гипотеза была отвергнута ранее и более не учитывалась в решении прикладной задачи (незначимая переменная по корреляционным матрицам).
3. Переменная «уровень жизни» значима для итоговой модели и ее коэффициент будет отрицательным; Рассчитанная значимость – -0,003903486, коэффициент – 0,04355316, как результат, при всех уровнях значимости нет оснований отвергать гипотезу.
4. Переменная «уровень безработицы» является значимой для итоговой модели, и ее коэффициент будет положительным; Рассчитанная значимость – 0,020832287, коэффициент – -0,174035022, как результат, при всех уровнях значимости нет оснований отвергать гипотезу.
5. Переменная «прирост реальной зп» является значимой для итоговой модели, и ее коэффициент будет отрицательным; Данная переменная также была исключена из анализа в рамках других тестов, соответственно, гипотеза была отвергнута ранее и более не учитывалась в решении прикладной задачи (незначимая переменная по корреляционным матрицам).
6. В модели будет присутствовать корреляция между переменными «средний возраст авто» и «уровень безработицы», коэффициент корреляции будет близок к 1.

Корреляция между переменными по критерию Пирсона равна 0, по критерию Спирмана равна 0,529. Как и раньше, гипотеза может быть отвергнута.

Итоговая наиболее оптимальная модель может быть представлена в виде формулы:

## 7.4 Проверка мультиколлинеарности и гетероскедастичности в рамках оптимальной модели

### 7.4.1 Проверка мультиколлинеарности с помощью VIF-анализа

Для проверки итоговой модели на мультиколлинеарность были проведены VIF-анализы. В результате было определено, что мультиколлинеарность присутствует у первой категории по коммунальным расходам и 3, 4 и 5 категорий. Возможно, это вызвано тем, что в первой категории присутствует только одно наблюдение. При проведении анализов для данных переменных был обнаружен коэффициент VIF выше 10.

|  |  |  |  |
| --- | --- | --- | --- |
| **Coefficientsa** | | | |
| Model | | Collinearity Statistics | |
| Tolerance | VIF |
| 1 | 2 Категория по коммунальным расходам | .128 | 7.815 |
| 3 Категория по коммунальным расходам | .046 | 21.947 |
| 4 Категория по коммунальным расходам | .045 | 22.329 |
| 5 Категория по коммунальным расходам | .058 | 17.106 |
| 6 Категория по коммунальным расходам | .145 | 6.884 |
| Средний возраст автомобиля | .693 | 1.443 |
| Рейтинговый балл региона по уровню жизни | .218 | 4.589 |
| LOG Население | .132 | 7.601 |
| LOG Безработные | .230 | 4.343 |
| LOG пассажиро поток | .162 | 6.175 |

|  |
| --- |
| a. Dependent Variable: 1 Категория по коммунальным расходам |

|  |  |  |  |
| --- | --- | --- | --- |
| **Coefficientsa** | | | |
| Model | | Collinearity Statistics | |
| Tolerance | VIF |
| 1 | 3 Категория по коммунальным расходам | .278 | 3.592 |
| 4 Категория по коммунальным расходам | .271 | 3.688 |
| 5 Категория по коммунальным расходам | .320 | 3.127 |
| 6 Категория по коммунальным расходам | .565 | 1.769 |
| Средний возраст автомобиля | .693 | 1.443 |
| Рейтинговый балл региона по уровню жизни | .218 | 4.589 |
| LOG Население | .132 | 7.601 |
| LOG Безработные | .230 | 4.343 |
| LOG пассажиро поток | .162 | 6.175 |
| 1 Категория по коммунальным расходам | .719 | 1.391 |

|  |
| --- |
| a. Dependent Variable: 2 Категория по коммунальным расходам |

|  |  |  |  |
| --- | --- | --- | --- |
| **Coefficientsa** | | | |
| Model | | Collinearity Statistics | |
| Tolerance | VIF |
| 1 | 4 Категория по коммунальным расходам | .705 | 1.418 |
| 5 Категория по коммунальным расходам | .669 | 1.494 |
| 6 Категория по коммунальным расходам | .857 | 1.167 |
| Средний возраст автомобиля | .693 | 1.443 |
| Рейтинговый балл региона по уровню жизни | .218 | 4.589 |
| LOG Население | .132 | 7.601 |
| LOG Безработные | .230 | 4.343 |
| LOG пассажиро поток | .162 | 6.175 |
| 1 Категория по коммунальным расходам | .793 | 1.261 |
| 2 Категория по коммунальным расходам | .862 | 1.160 |

|  |
| --- |
| a. Dependent Variable: 3 Категория по коммунальным расходам |

|  |  |  |  |
| --- | --- | --- | --- |
| **Coefficientsa** | | | |
| Model | | Collinearity Statistics | |
| Tolerance | VIF |
| 1 | 5 Категория по коммунальным расходам | .680 | 1.471 |
| 6 Категория по коммунальным расходам | .845 | 1.183 |
| Средний возраст автомобиля | .693 | 1.443 |
| Рейтинговый балл региона по уровню жизни | .218 | 4.589 |
| LOG Население | .132 | 7.601 |
| LOG Безработные | .230 | 4.343 |
| LOG пассажиро поток | .162 | 6.175 |
| 1 Категория по коммунальным расходам | .796 | 1.256 |
| 2 Категория по коммунальным расходам | .857 | 1.166 |
| 3 Категория по коммунальным расходам | .720 | 1.388 |

|  |
| --- |
| a. Dependent Variable: 4 Категория по коммунальным расходам |

|  |  |  |  |
| --- | --- | --- | --- |
| **Coefficientsa** | | | |
| Model | | Collinearity Statistics | |
| Tolerance | VIF |
| 1 | 6 Категория по коммунальным расходам | .801 | 1.249 |
| Средний возраст автомобиля | .693 | 1.443 |
| Рейтинговый балл региона по уровню жизни | .218 | 4.589 |
| LOG Население | .132 | 7.601 |
| LOG Безработные | .230 | 4.343 |
| LOG пассажиро поток | .162 | 6.175 |
| 1 Категория по коммунальным расходам | .758 | 1.320 |
| 2 Категория по коммунальным расходам | .738 | 1.356 |
| 3 Категория по коммунальным расходам | .499 | 2.006 |
| 4 Категория по коммунальным расходам | .496 | 2.017 |

|  |
| --- |
| a. Dependent Variable: 5 Категория по коммунальным расходам |

|  |  |  |  |
| --- | --- | --- | --- |
| **Coefficientsa** | | | |
| Model | | Collinearity Statistics | |
| Tolerance | VIF |
|  | Средний возраст автомобиля | .693 | 1.443 |
| Рейтинговый балл региона по уровню жизни | .218 | 4.589 |
| LOG Население | .132 | 7.601 |
| LOG Безработные | .230 | 4.343 |
| LOG пассажиро поток | .162 | 6.175 |
| 1 Категория по коммунальным расходам | .690 | 1.450 |
| 2 Категория по коммунальным расходам | .477 | 2.094 |
| 3 Категория по коммунальным расходам | .234 | 4.280 |
| 4 Категория по коммунальным расходам | .226 | 4.430 |
| 5 Категория по коммунальным расходам | .293 | 3.411 |

|  |
| --- |
| a. Dependent Variable: 6 Категория по коммунальным расходам |

|  |  |  |  |
| --- | --- | --- | --- |
| **Coefficientsa** | | | |
| Model | | Collinearity Statistics | |
| Tolerance | VIF |
| 1 | Рейтинговый балл региона по уровню жизни | .219 | 4.565 |
| LOG Население | .132 | 7.569 |
| LOG Безработные | .239 | 4.182 |
| LOG пассажиро поток | .162 | 6.172 |
| 1 Категория по коммунальным расходам | .796 | 1.256 |
| 2 Категория по коммунальным расходам | .865 | 1.157 |
| 4 Категория по коммунальным расходам | .713 | 1.403 |
| 5 Категория по коммунальным расходам | .670 | 1.494 |
| 6 Категория по коммунальным расходам | .862 | 1.160 |

|  |
| --- |
| a. Dependent Variable: Средний возраст автомобиля |

|  |  |  |  |
| --- | --- | --- | --- |
| **Coefficientsa** | | | |
| Model | | Collinearity Statistics | |
| Tolerance | VIF |
| 1 | LOG Население | .142 | 7.042 |
| LOG Безработные | .473 | 2.116 |
| LOG пассажиро поток | .163 | 6.136 |
| 1 Категория по коммунальным расходам | .809 | 1.236 |
| 2 Категория по коммунальным расходам | .861 | 1.161 |
| 3 Категория по коммунальным расходам | .720 | 1.388 |
| 5 Категория по коммунальным расходам | .703 | 1.423 |
| 6 Категория по коммунальным расходам | .849 | 1.178 |
| Средний возраст автомобиля | .697 | 1.435 |

|  |
| --- |
| a. Dependent Variable: Рейтинговый балл региона по уровню жизни |

|  |  |  |  |
| --- | --- | --- | --- |
| **Coefficientsa** | | | |
| Model | | Collinearity Statistics | |
| Tolerance | VIF |
| 1 | LOG пассажиро поток | .162 | 6.175 |
| 1 Категория по коммунальным расходам | .798 | 1.252 |
| 2 Категория по коммунальным расходам | .863 | 1.158 |
| 4 Категория по коммунальным расходам | .707 | 1.415 |
| 5 Категория по коммунальным расходам | .705 | 1.419 |
| 6 Категория по коммунальным расходам | .870 | 1.150 |
| Средний возраст автомобиля | .720 | 1.389 |
| Рейтинговый балл региона по уровню жизни | .447 | 2.236 |
| LOG Население | .132 | 7.593 |

|  |
| --- |
| a. Dependent Variable: LOG Безработные |

|  |  |  |  |
| --- | --- | --- | --- |
| **Coefficientsa** | | | |
| Model | | Collinearity Statistics | |
| Tolerance | VIF |
| 1 | LOG Безработные | .231 | 4.338 |
| LOG пассажиро поток | .632 | 1.582 |
| 1 Категория по коммунальным расходам | .812 | 1.232 |
| 2 Категория по коммунальным расходам | .858 | 1.166 |
| 3 Категория по коммунальным расходам | .727 | 1.376 |
| 5 Категория по коммунальным расходам | .684 | 1.462 |
| 6 Категория по коммунальным расходам | .857 | 1.167 |
| Средний возраст автомобиля | .696 | 1.437 |
| Рейтинговый балл региона по уровню жизни | .235 | 4.251 |

|  |
| --- |
| a. Dependent Variable: LOG Население |

|  |  |  |  |
| --- | --- | --- | --- |
| **Coefficientsa** | | | |
| Model | | Collinearity Statistics | |
| Tolerance | VIF |
| 1 | 1 Категория по коммунальным расходам | .810 | 1.234 |
| 2 Категория по коммунальным расходам | .859 | 1.164 |
| 3 Категория по коммунальным расходам | .733 | 1.364 |
| 5 Категория по коммунальным расходам | .680 | 1.471 |
| 6 Категория по коммунальным расходам | .852 | 1.173 |
| Средний возраст автомобиля | .694 | 1.442 |
| Рейтинговый балл региона по уровню жизни | .219 | 4.560 |
| LOG Население | .513 | 1.948 |
| LOG Безработные | .230 | 4.343 |

|  |
| --- |
| a. Dependent Variable: LOG пассажиро поток |

### 7.4.2 Проверка мультиколлинеарности с помощью индекса обусловленности

Для оптимизированной модели были вычислены собственные значения вектора

На основании которого был вычислен индекс обусловленности

Данное значение индекса говорит о слабой мультиколлинеарности в модели.

### 7.4.3 Проверка гетероскедастичности произвольной формы. Тест Breusch-Pagan

После построения регрессии и нахождения коэффициентов были найдены следующие значения коэффициентов в регрессии :

|  |  |  |  |
| --- | --- | --- | --- |
|  | Свободный  коэффициент | -0,053874474 |  |
|  | Население | 0,024315919 |  |
|  | Уровень жизни | 8,97464E-05 |  |
|  | Безработица | -0,000923225 |  |
|  | Средний возраст авто | -6,17545E-05 |  |
|  | 1 категория по комм  расх | 0 |  |
|  | 2 категория по комм  расх | 0,055497325 |  |
|  | 3 категория по комм  расх | 0,10565144 |  |
|  | 4 категория по комм  расх | 0,045197291 |  |
|  | 5 категория по комм  расх | 0,051551077 |  |
|  | 6 категория по комм  расх | 0,06634362 |  |
|  | Пассажирооборот | -0,021201786 |  |

Значение равно 0,172, значение получилось равным 13,76, вычисление дало в результате 1, что говорит об отсутствии необходимости отвергать основную гипотезу. Иными словами, гетероскедастичности нет.

### 7.4.4 Проверка гетероскедастичности по конкретной переменной. Тест Park

Для проведения теста были выбраны несколько переменных: «уровень жизни», «уровень безработицы», «пассажирооборот» и «средний возраст авто», чтобы исследовать их отдельное влияние на гетероскедастичность. Для каждой из этих переменных была построена регрессия , результаты которой совпадают с базовой моделью:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Коэффициентыa** | | | | | | |
| Модель | | Нестандартизованные коэффициенты | | Стандартизованные коэффициенты | т | Знач. |
| B | Стандартная Ошибка | Бета |
| 1 | (Константа) | -5,771 | 1,784 |  | -3,236 | ,002 |
| Население | ,227 | ,250 | ,103 | ,908 | ,367 |
| a. Зависимая переменная: LN(e^2) | | | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Коэффициентыa** | | | | | | |
| Модель | | Нестандартизованные коэффициенты | | Стандартизованные коэффициенты | т | Знач. |
| B | Стандартная Ошибка | Бета |
| 1 | (Константа) | -4,139 | 4,123 |  | -1,004 | ,319 |
| Рейтинговый балл региона по уровню жизни | -,007 | 1,102 | -,001 | -,006 | ,995 |
| a. Зависимая переменная: LN(e^2) | | | | | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | **Коэффициентыa** | | | | | | | | Модель | | Нестандартизованные коэффициенты | | Стандартизованные коэффициенты | т | Знач. | | B | Стандартная Ошибка | Бета | | 1 | (Константа) | -3,296 | 1,750 |  | -1,884 | ,063 | | Безработица | ,296 | ,592 | ,057 | ,500 | ,618 | | a. Зависимая переменная: LN(e^2) | | | | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Коэффициентыa** | | | | | | |
| Модель | | Нестандартизованные коэффициенты | | Стандартизованные коэффициенты | т | Знач. |
| B | Стандартная Ошибка | Бета |
| 1 | (Константа) | -3,538 | 1,169 |  | -3,026 | ,003 |
| Средний возраст автомобиля | -,047 | ,086 | -,062 | -,546 | ,587 |
| a. Зависимая переменная: LN(e^2) | | | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Коэффициентыa** | | | | | | |
| Модель | | Нестандартизованные коэффициенты | | Стандартизованные коэффициенты | т | Знач. |
| B | Стандартная Ошибка | Бета |
| 1 | (Константа) | -5,098 | 1,364 |  | -3,737 | ,000 |
| Пассажиропоток | ,136 | ,196 | ,079 | ,694 | ,490 |
| a. Зависимая переменная: LN(e^2) | | | | | | |

Коэффициенты значимости получились следующие:

|  |  |
| --- | --- |
| **Переменная** | **Значимость** |
| Уровень жизни | 0,995 |
| Уровень безработицы | 0,618 |
| Пассажирооборот | 0,49 |
| Средний возраст авто | 0,587 |
| Население | 0,367 |

Таким образом, ни по какой из переменных не была выявлена значимость для регрессии и, как итог, гетероскедастичность

### 7.4.5 Проверка гетероскедастичности по конкретной переменной. Тест Goldfeld-Quandt

Ключевым фактором в рамках данного теста была выбрана переменная «пассажирооборот». После деления данных на три группы и выброса средней были построены линейные регрессии по первым 30 записям данных и по последним 30.   
В 1 группе получились следующие коэффициенты:

|  |  |
| --- | --- |
| Свободный коэффициент | 0,788891581 |
| Население | 0,826307117 |
| Уровень жизни | 0,008168167 |
| Безработица | 0,181742195 |
| Средний возраст авто | 0,004373665 |
| 1 категория по комм расх | 0 |
| 2 категория по комм расх | -0,219416145 |
| 3 категория по комм расх | 0,272773511 |
| 4 категория по комм расх | -0,085541566 |
| 5 категория по комм расх | 0,312753872 |
| 6 категория по комм расх | 0 |
| Пассажирооборот | 0,117186936 |

При этом =0,956729904405346, а = 0,83627732.

Во 2 группе получились следующие коэффициенты:

|  |  |
| --- | --- |
| Свободный коэффициент | -1,117705468 |
| Население | 0,808368967 |
| Уровень жизни | -0,01112817 |
| Безработица | -0,242101668 |
| Средний возраст авто | 0,007284731 |
| 1 категория по комм расх | -0,027502534 |
| 2 категория по комм расх | 0,277321204 |
| 3 категория по комм расх | 0,172566287 |
| 4 категория по комм расх | 0,259386406 |
| 5 категория по комм расх | 0 |
| 6 категория по комм расх | 0 |
| Пассажирооборот | 0,291142203 |

в этом случае равен 0,904691677, а = 0,976319142.

Получается , и по распределению Фишера

Так как , то нет оснований отвергать основную гипотезу, а значит, нет гетероскедастичности.

Таким образом, оптимизированная модель была проверена на наличие мультиколлинеарности и гетероскедастичности, которые не были выявлены в ходе анализов.

# Построение доверительных интервалов

## 8.1 Построение доверительных интервалов для истинного значения регрессии

Для построения доверительных интервалов для параметров модели был выбран уровень значимости в 0,05, что означает, с вероятностью 0,95 истинное значение параметра будет накрыто вычисленным интервалом.

Все доверительные интервалы рассчитывались по формуле

При этом

,

,

.

Так, были получены следующие интервальные значения:

* Для параметра «население» – (0,746579831; 1,059407498);
* Для параметра «рейтинговый балл региона по уровню жизни» – (-0,016411366; 0,008604394);
* Для параметра «уровень безработицы» – (-0,447430018; 0,099359974);
* Для параметра «средний возраст автомобиля» – (-0,007860164; 0,040238544);
* Для параметра «пассажирооборот» – (-0,05216542;0,174502991);
* Для параметра «1-ая категория по коммунальным расходам» – (-0,031262954; 1,558679583);
* Для параметра «2-ая категория по коммунальным расходам» – (0,375141488; 1,586005583);
* Для параметра «3-ая категория по коммунальным расходам» – (0,529471898; 1,693350752);
* Для параметра «4-ая категория по коммунальным расходам» – (0,492475636; 1,640885618);
* Для параметра «5-ая категория по коммунальным расходам» – (0,751874477;1,956059657);
* Для параметра «6-ая категория по коммунальным расходам» – (0,545458622; 1,801679213);
* Для свободного члена – (-1,993707959; -0,283360296).

## 8.2 Построение доверительных интервалов для параметров модели

Кроме того, были найдены доверительные интервалы для каждого параметра модели при уровне значимости 5%.

После проведения расчетов получились следующие значения:

* Для параметра «население» – (0,810740854; 0,952333671);
* Для параметра «рейтинговый балл региона по уровню жизни» – (0,04542022; 0,076122957);
* Для параметра «уровень безработицы» – (-1,530524301; -0,636502587);
* Для параметра «средний возраст автомобиля» – (-0,206656052; -0,080093626);
* Для параметра «пассажирооборот» – (0,561942141; 0,736531072);
* Для параметра «1-ая категория по коммунальным расходам» – (0,289161156; 3,273769968);
* Для параметра «2-ая категория по коммунальным расходам» – (-0,378532952; 1,036601244);
* Для параметра «3-ая категория по коммунальным расходам» – (-0,327495529; 0,423214494);
* Для параметра «4-ая категория по коммунальным расходам» – (-0,291238648; 0,45879523);
* Для параметра «5-ая категория по коммунальным расходам» – (-0,818763809; 0,027117121);
* Для параметра «6-ая категория по коммунальным расходам» – (-0,922379429; 0,656651724).

# Обнаружение выбросов и влиятельных наблюдений в модели

## Расстояние Махаланобиса

Чтобы вычислить меру удаления наблюдения от центра системы в рамках текущего многопеременного анализа, в IBM SPSS были рассчитаны расстояния Махаланобиса, общая формула которых выглядит следующим образом:

Критическое значение, с которым сравниваются рассчитанные расстояния, вычисляется при уровне значимости 0,05 и имеет формулу

Так, для текущий данных , а .

При сравнении с критическим значением расстояний, посчитанных в SPPS, выбросами являются: Камчатский край (25,36292), Москва (78,0125), Ненецкий автономный округ (37,86413), Республика Тыва (28,11178), Санкт-Петербург (30,87561), Чеченская республика (78,0125) и Ямало-Ненецкий автономный округ (23,13219).

## Расстояние Кука

Расстояние Кука, которое является общей мерой влияния наблюдений, вычисляется по формуле

В рамках анализа оно было рассчитано для каждого наблюдения с помощью IBM SPSS. Критическое значение находится как медиана F-распределения со степенями свободы k и n-k. В данном случае оно равно 1,869186905. В результате, ни одно из наблюдений не превышает рассчитанное критическое значение, что говорит о том, что ни одно значение не может считаться значимо выделяющимся своим влиянием на регрессионную зависимость.

## Ковариационное отношение

Ковариационное отношение также является общей мерой влияния наблюдений и рассчитывается по формуле

Влиятельными наблюдениями являются значимо отличающиеся от 1. Если , то наблюдение существенно влияет на регрессионную зависимость.

По итогам расчетов получилось, что влиятельными являются следующие наблюдения: Белгородская область, Калужская область, Камчатский край, Магаданская область, Москва, Ненецкий автономный округ, Республика Тыва, Санкт-Петербург, Томская область, Чеченская республика.

## Меры DFFIT и DFFTIS

*DFFIT* – разность между ожидаемым значением для данного наблюдения и удаленным ожидаемым значением для него. Она рассчитывается для каждого регрессора по формуле

,

тогда как стандартизированное значение вычисляется как

Так как выборка имеет средний объем, влиятельными являются те, значения мер DFFIT и DFFTIS которых больше 1. В рамках анализа данные меры были посчитаны в IBM SPSS и ни одна из них не достигла 1, значит, ни одно из наблюдений не является значительным.

## Меры DFBETA и DFBETAS

Меры DFBETA и DFBETAS - специфические меры влияния, оценивающие степень изменения отдельных параметров регрессионной модели при исключении из i-того наблюдения.

DFBETA может быть найдена по формуле

,

где – ожидаемое значение для j-ого параметра регрессии.

Стандартизированная мера DFBETA в случае линейной регрессии вычисляется как

Как и ранее, данные меры, вычисленные в IBM SPSS, сравнивались с 1, так как выборка не является большой.

Как результат, ни одно из значений не достигает единицы, что означает, ни одно из наблюдений не влияет значительно на регрессию.

По итогам проведенных анализа были найдены пересечения влиятельных наблюдений и выбросов по всем тестам. Так, пересечениями по тесту Махаланобиса и ковариационным отношениям (единственным тестам, где были найдены выбросы) являются: Камчатский край, Москва, Ненецкий автономный округ, Республика Тыва, Санкт-Петербург и Чеченская республика.

Чтобы оценить эффект от исключения данных записей из анализа, была построена линейная модель без выявленных наблюдений. Так, получилось, что в этом случае коэффициент детерминации (как обычный, так и модифицированный) значительно снижаются по сравнению с исходной моделью – с 0,942 и 0,938 соответственно до 0,91798484 и 0,889982508. Таким образом, было решено сохранить все выявленные как выбросы наблюдения.