Федеральное агентство связи
Федеральное государственное образовательное учреждение высшего профессионального образования
«Сибирский Государственный Университет Телекоммуникаций и Информатики»
(ФГБОУ ВО «СибГУТИ»)

Кафедра математического моделирования бизнес-процессов (ММБП)

ИССЛЕДОВАТЕЛЬСКАЯ РАБОТА

по курсу «Эконометрика»

Тема: «Причины ДТП - водители или обстоятельства? Исследование факторов, влияющих на летальные исходы в ДТП»

Выполнил: студент 3 курса Третяк А.Н.

Факультет: Информатики и Вычислительной Техники (ИВТ)

Группа: ИИ-461

Научный руководитель:

доцент Михалева М.М.

Новосибирск

2016 г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В Российской Федерации проблема безопасного движения является одной из наиболее значимых. Высокие темпы роста автомобильного парка страны создают дополнительные предпосылки ухудшения обстановки на дорогах. В настоящее время в Госавтоинспекции (ГАИ) состоит более 50.5 миллионов единиц транспортных средств. Причем основную часть из них – 76.7%, или 38.7 миллионов единиц - составляют легковые автомобили. По данным ГИБДД Новосибирской области, на 1 января 2016 года в Новосибирске было зарегистрировано 547 тысяч автомобилей, а 31 августа 2016 года машин в городе стало уже 556 тысяч. Если учесть, что значительная часть автомобилей зарегистрирована не в Новосибирске, то по оценкам экспертов в области автомобильного движения количество машин, находящихся в городе, составляет около 1 миллиона. Из них примерно 20-25% принадлежат юридическим лицам, остальные находятся в личном пользовании. Ежегодное увеличение числа автомобилей в Новосибирске ведет к частым пробкам, нехватке парковок, увеличению числа ДТП и другим сопутствующим проблемам, поскольку потенциал дорожной сети мегаполиса по вместимости уже давно исчерпан.

По сравнению с другими странами Россия имеет недопустимо высокие значения относительных показателей, характеризующих уровень опасности дорожного движения, прежде всего показателей количества погибших в дорожно-транспортных происшествиях (ДТП). В настоящее время для России этот риск оценивается величиной порядка 3,58%, в то время как для наиболее безопасных в этом смысле стран он на порядок меньше (Швеция, Великобритания, Финляндия, Германия, США). При том, что в России уровень автомобилизации оценивается в 210 транспортных средств на 1000 человек, в то время, как в названных странах он оценивается в интервале от 464 (Финляндия) до 790 (США). Из этих данных можно сделать вывод, что количество автомобилей на душу населения в России меньше чем в этих странах, а вероятность погибнуть в ДТП у граждан значительно выше. Ниже представлены данные в таблице (Таблица 1), отражающие количество летальных исходов в ДТП и вытекающие из этого другие показатели в России, США и некоторых европейских государствах. Таблица составлена на основании данных, опубликованных 13.11.2015 а сайте www.icebike.org

Таблица 1 - «Статистика по ДТП в мире в 2015 году»

Страна	**Погибло, чел.**	**Смертей на 100000 жителей, %**	**Риск погибнуть в ДТП, %**
Россия	23114	5.7	0.358
Финляндия	380	4.7	0.215
Швеция	471	2.7	0.202
Великобритания	3298	3.5	0.331
Германия	4949	4.4	0.258
США	31424	5.2	0.342

Проблема летальных исходов в результате ДТП всегда являлась актуальной, а в настоящее время в особенности: парк автомобилей расширяется, количество смертей растет, правительство принимает всевозможные меры по решению этой проблемы. Кроме того, проблема аварийности, связанной с автомобильным транспортом в последнее десятилетие приобрела особую остроту в связи с несоответствием дорожно-транспортной инфраструктуры потребностям общества и государства в безопасном

дорожном движении, недостаточной эффективностью функционирования системы обеспечения безопасности дорожного движения и крайне низкой дисциплиной участников дорожного движения. Эксперты считают положение России в мировом рейтинге ДТП неудовлетворительным, а принимаемые государством меры неэффективными.

Существует множество факторов, влияющих так или иначе на увеличение риска возникновения ДТП, но все факторы охватить невозможно, так как это может быть, как человеческий фактор, так и техническая неисправность автомобиля. Достоверную статистику по таким и подобным факторам составить практически невозможно, поэтому для исследования необходимо брать такие факторы, данные которых достоверны на 100% и ввод корректив в безопасность дорожного движения по которым помог бы уменьшить риск возникновения ДТП. Такие данные, в основном, находятся в открытом доступе на сайтах региональных управлений ГИБДД в виде сводок о месте, времени, количестве пострадавших, возрасте и стаже водителя, совершившего ДТП, и прочее. Данные о погодных условиях за последние 10 лет располагаются в открытом доступе в Интернете. Таким образом, исследовательская работа построена на основе поиска достоверной информации в открытых источниках в сети Интернет и литературе по соответствующей теме.

По моему мнению, наиболее очевидными факторами, влияющими на смертность на дорогах, являются – стаж и возраст водителя, погодные условия, время суток и качество дорожного покрытия. По моему предположению, риск возникновения ДТП с летальным исходом наиболее высок у водителей с высоким стажем, так как порой в нестандартных ситуациях они руководствуются своим стажем вождения и несколько самоуверенны, менее – у водителей с низким стажем из-за их неопытности. При погодных условиях – с наиболее высоким уровнем осадков и/или низкой температуре. А во время суток наиболее вероятно возникновение ДТП в вечернее время и час-пик из-за большого количество транспортных средств и людей на дорогах.

В большинстве случаев летальный исход в ДТП и само происшествие происходит по стечению многих обстоятельств. Таким образом, сделав исследование на эту тему, можно сделать выводы, из которых может стать ясно, что же является наибольшей опасностью для возникновения ДТП и летального исхода в нем, и выработать рекомендации, руководствуясь которыми можно предотвратить возникновение таких случаев или, по крайней мере, найти причину и способ снижения уровня риска возникновения ДТП.

Исследование было проведено на территории Новосибирской области, используя данные из сводок ГИБДД 2015 года. Анализ данных производится при помощи языка программирования R для статистической обработки данных и работы с графикой, а также свободной программной средой вычислений с открытым исходным кодом в рамках проекта GNU. Исследование проводится на основе регрессионного анализа.

ГЛАВА 1

1.1 ОБЗОР НАУЧНОЙ ЛИТЕРАТУРЫ ПО ТЕМЕ

При написании данной работы были использованы научная и учебно-методическая литература, статьи в периодических изданиях Российской Федерации и зарубежья, нормативно-законодательные акты Российской Федерации.

Основными источниками, раскрывающими суть возникновений ДТП, явилось научное исследование Йоахима Б. и Гвидо Б. «Дорожно-транспортные происшествия с летальным исходом», Хумлегарда Г. «Летальный исход в ДТП», Хайцмен А. и Шаде О. «Проблемы правоохранительной деятельности» и книга Якимова О. «Дорожно-транспортные происшествия». В данных источниках подробно рассмотрено понятие ДТП с летальным исходом, место данной проблемы в мире, виды ДТП и их особенности. Рассмотрены причины и виды дорожно-транспортных происшествий. Даны практические советы, как поступить в конкретных ситуациях, возникших при ДТП, с целью объективного решения вопросов.

В издании Якимова О. рассмотрены следствия возникновения дорожно-транспортных происшествий. Подробно описаны виды ответственности водителя за ДТП (дисциплинарная, материальная, гражданско-правовая, административная и уголовная). Даны советы, с целью решения вопросов по поводу вопросов предупреждения ДТП. Приведены образцы документов, составление которых может потребоваться каждому пострадавшему. Данное издание носит больше теоретический и общий характер знаний, которые доносятся читателю.

В статье Йоахима Б. и Гвидо Б. «Дорожно-транспортные происшествия с летальным исходом» представлены результаты исследования, в котором сделан сравнительный анализ ДТП с летальным исходом в Северном Рейн - Вестфалия за 12 месяцев. В программе по обеспечению безопасности дорожного движения 2004 г. в Северный Рейн - Вестфалия проведено планирование снижения случаев летального исхода в результате ДТП на половину до 2015 года. Основой для исследования стало рассуждение, что смертные случаи в результате ДТП не являются случайными результатами, а ситуативно обусловлены в результате ошибочного поведения отдельных участников дорожного движения.

Особенно актуальной для исследования была работа Хумлегарда Г. (Humlegard, 2009), который исследовал все зарегистрированные ДТП с летальным исходом за 2004-2005 гг. в Норвегии. Среди виновников ДТП он смог выделить 2 группы: одни совершали ДТП из-за неадекватного вождения, возникающего вследствие невнимательности, плохого самочувствия или усталости. 52% виновников ДТП имели заслуживающее порицания поведение. Почти все они были возрастом моложе, чем другая группа. 48% совершенных ДТП произошли по причине превышения скорости в связи с употреблением спиртных напитков или наркотиков, а также вследствие агрессивного поведения. Также исследования, касающиеся криминальной или специфической, связанной с дорожно-транспортными происшествиями судимости соответствуют представленным выше результатам. Результаты данного исследования имеют значение с двойной точки зрения. С одной стороны, они показывают, что виновники ДТП с летальным исходом образуют гетерогенную группу, которая легко раздваивается. С

другой стороны, тем самым подчеркивается, что дальнейшее исследование ДТП с летальным исходом необходимо, чтобы получить анализ лиц, которые посредством своего неправильного поведения приводят к катастрофам. Представленные результаты исследований содержат вывод о том, что, в том числе, имеет место зависимость от различных внешних факторов при совершении участником движения ДТП, в том числе с тяжкими последствиями.

Связь между преступностью в дорожном движении, соизмеримо с количеством зарегистрированных в Центральном реестре нарушений правил дорожного движения, с одной стороны, и повышенной вероятностью совершения ДТП, с другой стороны, освещается подробно Хайцмен и Шаде (Heizmann, Schade, 2004). Они смогли показать, что дифференцирование риска на основании данных Центрального реестра нарушений правил дорожного движения происходит в высокой степени из-за сильной дифференциации по полу и возрасту. Так, водители легковых автомобилей в возрасте 18-51 года, имеющие более трех зарегистрированных случаев нарушения правил дорожного движения, в противоположность к водителям-женщинам возрастной группы 41-60 лет, не имеющими зарегистрированных нарушений, имеют 25-кратный риск стать виновником совершения ДТП в следующие 12 месяцев.

Данные научные статьи очень помогли в осмыслении проблемы ДТП с летальным исходом, помогли понять, что количество факторов, влияющих на риск возникновения ДТП, намного больше, чем предполагалось. Кроме того, они дали понимание того, в каком направлении стоит двигаться в данном исследовании.

ГЛАВА 2

2.1 СБОР МАТЕРИАЛА. ЕГО ОПИСАНИЕ. СТРУКТУРИРОВАНИЕ И АНАЛИЗ

Почти 2.6 тысяч ДТП произошло в Новосибирской области в 2015 году. Такие данные привел НовосибирскСтат, основываясь на статистике ГИБДД. Погибших при ДТП оказалось 450, травмы получили более 2,9 тысячи. «По сравнению с 2014 годом число погибших в дорожно-транспортных происшествиях снизилось на 14%. Основная масса всех ДТП произошла из-за дорожной обстановки и нарушений ПДД водителями и пешеходами. Поскольку, составить статистику по количеству совершенных нарушений ПДД не предоставляется возможным из-за конфиденциальности информации в базе ГИБДД и страховых компаниях, то анализ летальных исходов в ДТП проводится по внешним обстоятельствам (погодные условия, дорожная обстановка) и опыту водителя. Здесь же ставится цель исследования – сделать вывод по тому, что больше влияет на летальные исходы в ДТП – опыт водителя или обстоятельства?

Все данные предоставляется возможным взять с сайтов региональных управлений ГИБДД из годовых отчетов или ежедневных сводок. В данном случае данные о случившихся ДТП и результаты их расследований взяты с сайта Управления ГИБДД ГУ МВД России по Новосибирской области. Данные о дорожной обстановке предоставляется возможным взять из сервиса «Яндекс.Пробки» с запросом на необходимое время с помощью веб-архива. Данные о погоде за определенную дату и время находятся практически на каждом сайте, предоставляющим информацию о погоде, либо это так же возможно сделать при помощи веб-архива с учетом функционирования сайта в запрашиваемое время. Данные были собраны за период 02.05.2015 – 24.12.2015, где в сводках ГИБДД НСО зафиксировано около 200 случаев ДТП.

Для оценки зависимости числа летальных исходов в ДТП от внешней обстановки и опыта водителей были взяты следующие критерии:

1. Количество летальных исходов в одном ДТП;

2. Стаж водителя;

3. Возраст водителя;

4. Температура воздуха;

5. Плотность трафика;

6. Время суток;

7. Количество осадков.

Кроме всех выбранных факторов существует необходимость взять в рассмотрение фактор «Индекс качества дорожного покрытия», но поскольку Министерство транспорта и дорожного хозяйства Новосибирской области не распространяет информацию данного характера, то провести сбор и анализ данных этого критерия не представляется возможным.

Далее проведено описание приведенных выше факторов и указание источников, откуда они были взяты.

1. Количество летальных исходов в одном ДТП

Данный показатель отражает количество человек с летальным исходом в ДТП. В данном показателе не учитывается количество человек, получивших травмы. Для каждого ДТП фиксировалось количество пострадавших как со стороны нарушителя, так и со стороны пострадавшего. Задача снижения количества жертв и раненых при ДТП является приоритетной для дорожных и правоохранительных ведомств.

Информация такого рода обязательна в каждой сводке совершенного ДТП как в СМИ, так и в базе ДТП ГИБДД, таким образом такая информация легкодоступна и имеется в открытом доступе, что в свою очередь позволяет составить достоверный набор данных.

2. Стаж водителя

Этот показатель отражает информацию о том, сколько лет водитель управлял ТС до совершенного им ДТП. Данная информация имеется в базе ГИБДД о сводках ДТП, реже информация такого рода оказывается в сводках СМИ.

Данный фактор один из самых важных в анализе совершенных ДТП, так как предыдущие исследования показываю прямую зависимость количества совершенных ДТП от стажа водителя.

3. Возраст водителя

Возраст виновника ДТП – ключевая информация в сводках ДТП ГИБДД. Она имеется в каждой сводке ДТП, и как стаж водителя является важным фактором в анализе совершенных ДТП и летальных исходов в них.

Кроме того, существует статистика, которая показывает, что ДТП наиболее часто совершают водители в юношеском и пожилом возрасте.

4. Температура воздуха

Данный фактор отражает температуру воздуха в Цельсиях (ºС). Экспериментальные данные, полученные многими исследователями в разное время, дают информацию о том, что на опорных поверхностях, покрытых снегом и льдом, коэффициент сцепления увеличивается с уменьшением температуры. Особенно значительные изменения происходят в диапазоне температур от 0 °С до –15 °С.

Изменение температуры воздуха влияет как на состояние водителя, так и на дорожное покрытие, как было описано выше. Таким образом, такая информация имеет

место быть как один из факторов, влияющих на количество летальных исходов и ДТП, в целом.

Данные о погоде за определенную дату и время находятся практически на каждом сайте, предоставляющим информацию о погоде, либо это так же возможно сделать при помощи веб-архива с учетом функционирования сайта в запрашиваемое время.

5. Плотность трафика

Плотность потока трафик представляет количество автомобилей, занимающих заданную длину полосы или дороги в определенный момент времени. Данный показатель измеряется в виде шкалы от 0 до 10 баллов: чем больше значение показателя, тем больше плотность трафика и наоборот. Данный показатель также является важным фактором, отражающим количество ДТП и количество летальных исходов, в частности.

Данные о дорожной обстановке предоставляется возможным взять из сервиса «Яндекс.Пробки» с запросом на необходимое время с помощью веб-архива.

6. Время суток

При регистрации каждой аварии указывается время, когда она произошла. Значение этой переменной недостаточно точно, из-за особенностей регистрации ДТП, но целью анализа является нахождение наиболее аварийного времени суток. Распределение количества аварий также ожидается неоднородным; логично предположить, что в ночные часы происходит меньше аварий.

06:00-12:00 – Утро = 0;

12:00-18:00 – День = 1;

18:00-23:00 – Вечер = 2;

23:00-06:00 – Ночь = 3.

7. Количество осадков

Одним из факторов, увеличивающих потенциальный риск ДТП, являются неблагоприятные погодные условия. Статистические данные подтверждают, что во время осадков число ДТП увеличивается. Выявлены закономерности, что неожиданные осадки после продолжительного сухого периода вызывают резкое увеличение риска ДТП, а затяжные осадки вызывают адаптацию водителей, в результате чего число ДТП постепенно уменьшается. На скользком дорожном покрытии, сразу после наступления гололеда, риск возникновения ДТП возрастает. По мере адаптации водителей к сложным дорожным условиям число ДТП постепенно уменьшается, влияние неблагоприятного внешнего фактора снижается.

Данный показатель измеряется в процентах (%) как количество выпавших осадков. Информация предоставляется совместно с температурой таким же способом, как описано в пункте 4 «Температура воздуха».

Краткое изложение описания данных для общей структурированности приведены в таблице 2.

Таблица 2 - «Описание исходных данных»

Код переменной	Описание переменной	Тип переменной	Единицы измерения
QT	Количество летальных исходов в одном ДТП	Количественная	Человек
EX	Стаж водителя	Количественная	Лет
AG	Возраст водителя	Количественная	Лет
TP	Температура воздуха	Количественная	Градусов, ^{o}C
DN	Плотность трафика	Количественная	Балл
TM	Время суток	Качественная	Время суток
PR	Количество осадков	Количественная	Процент, %

Всего же было проанализировано 100 наблюдений по 6 переменным.

2.2 ИЗУЧЕНИЕ ИСХОДНЫХ ДАННЫХ

Проведем нахождение значений описательных статистик по каждой переменной:

```
> summary(DTP_Base$EX)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    6.00   11.50   14.39   21.00   44.00
```

Средний стаж водителя-виновника ДТП составляет 14.39 лет. Минимальный стаж 0 лет, максимальный 44 года. 50% всех совершенных ДТП совершают водители со стажем от 6 до 21 года.

```
> summary(DTP_Base$AG)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  19.00   30.00   36.00   38.34   45.00   73.00
```

Средний возраст водителя-виновника ДТП составляет 38.34 лет. Минимальный возраст 19 лет, максимальный 73 года. 50% всех совершенных ДТП совершают водители в возрасте от 30 до 45 лет.

```
> summary(DTP_Base$TP)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -33.00   -5.00    2.00    3.94   15.00   27.00
```

Средняя температура в момент совершения ДТП составляет 3.94 градуса по Цельсию. Минимальная температура -33 градуса, максимальная 27 градусов. 50% всех ДТП происходит при температуре от -5 до 15 градусов.

```
> summary(DTP_Base$PR)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    0.00    0.00    3.24    5.00   30.00
```

Среднее количество выпавших осадков в момент совершения ДТП составляет 3.24%. Минимальное количество выпавших осадков 0%, максимальное 30%. 50% всех ДТП происходит от 0 до 5 процентов выпавших осадков.

```
> summary(DTP_Base$DN)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    1.00    3.00    3.46    5.00   10.00
```

Средняя плотность траффика, при котором совершаются ДТП составляет 3.46 баллов. Минимальная плотность траффика 0 баллов, максимальная 10 баллов. 50% всех совершенных ДТП происходит при плотности траффика от 1 до 5 баллов.

Данные характеристики представим на диаграмме «ящик с усами». (Рисунок 1,2,3,4,5)

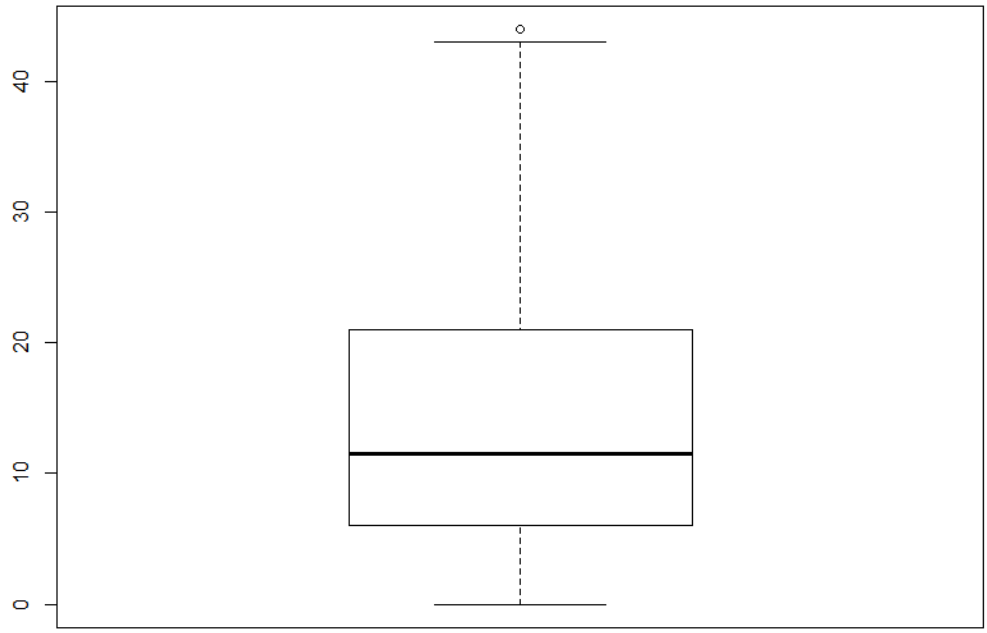

Рисунок 1 – «Ящик с усами переменной “Температура воздуха”»

На диаграмме «ящик с усами» переменной “Температура воздуха” видно, что значения смещены влево. Это означает, что большее количество наблюдений меньше среднего значения. На данной диаграмме также присутствует 1 выброс, выделяющийся из общей выборки.

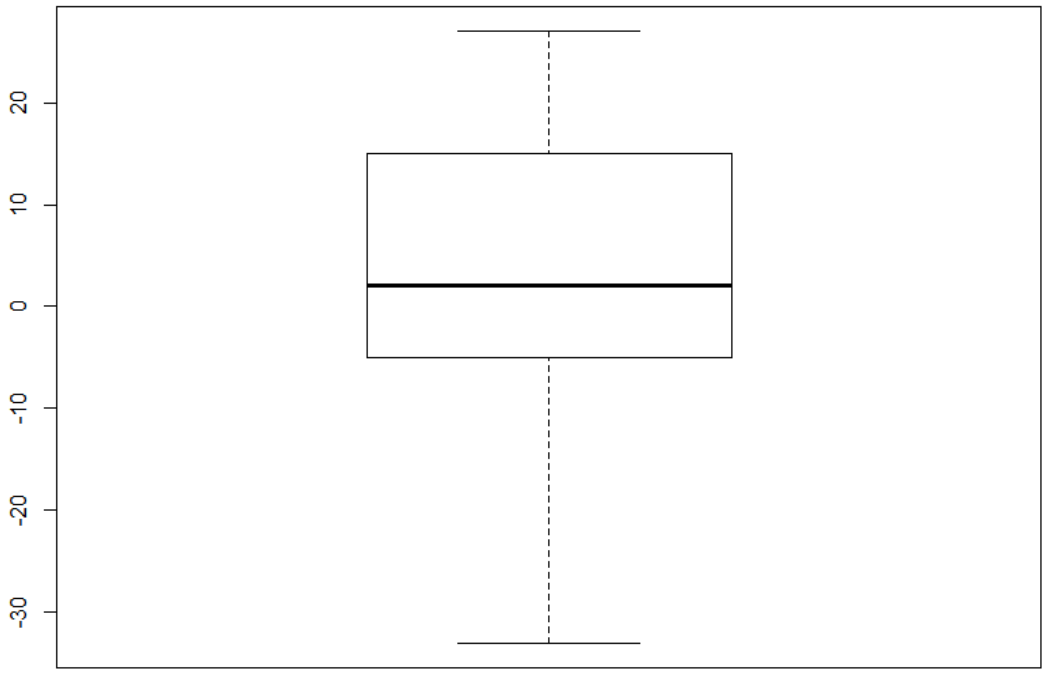

Рисунок 2 – «Ящик с усами переменной “Стаж водителя”»

На диаграмме «ящик с усами» переменной “Стаж водителя” видно, что значения смещены вправо. Это означает, что большее количество наблюдений больше среднего значения.

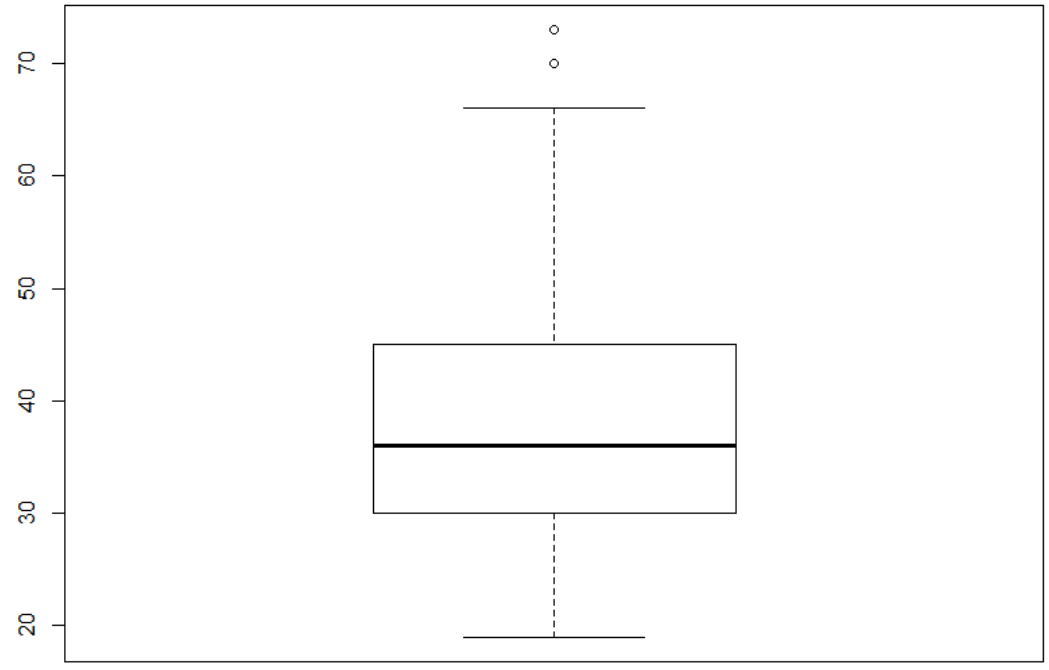

Рисунок 3 – «Ящик с усами переменной “Возраст водителя”»

На диаграмме «ящик с усами» переменной “Возраст водителя” видно, что значения смещены влево. Это означает, что большее количество наблюдений меньше среднего значения. На данной диаграмме также присутствует 2 выброса, выделяющихся из общей выборки.

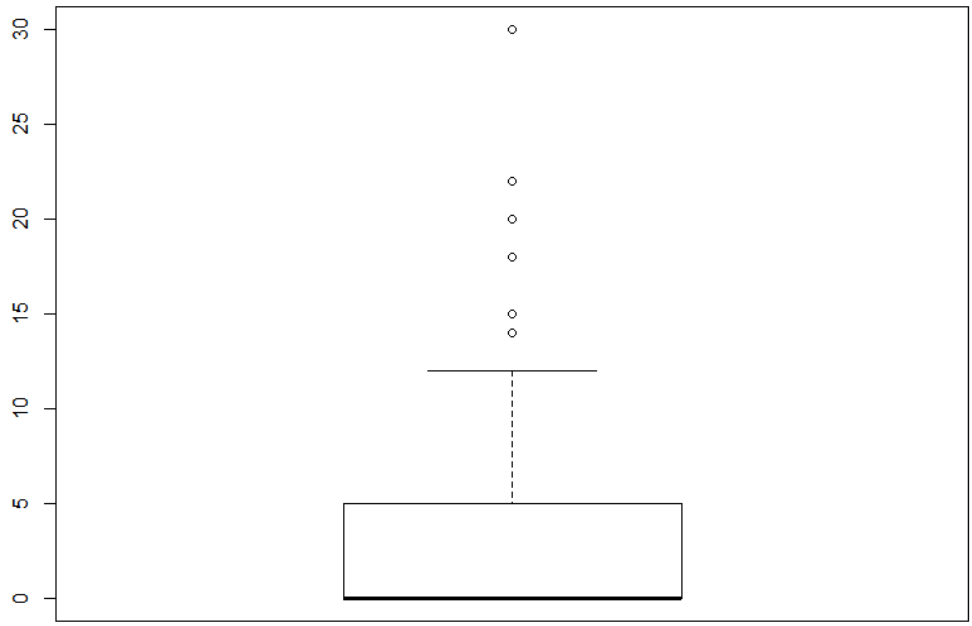

Рисунок 4 – «Ящик с усами переменной “Количество осадков”»

На диаграмме «ящик с усами» переменной “Количество осадков” видно, что значения значительно смещены влево. Это означает, что большее количество наблюдений меньше среднего значения. На данной диаграмме также присутствует 6 выбросов, выделяющихся из общей выборки.

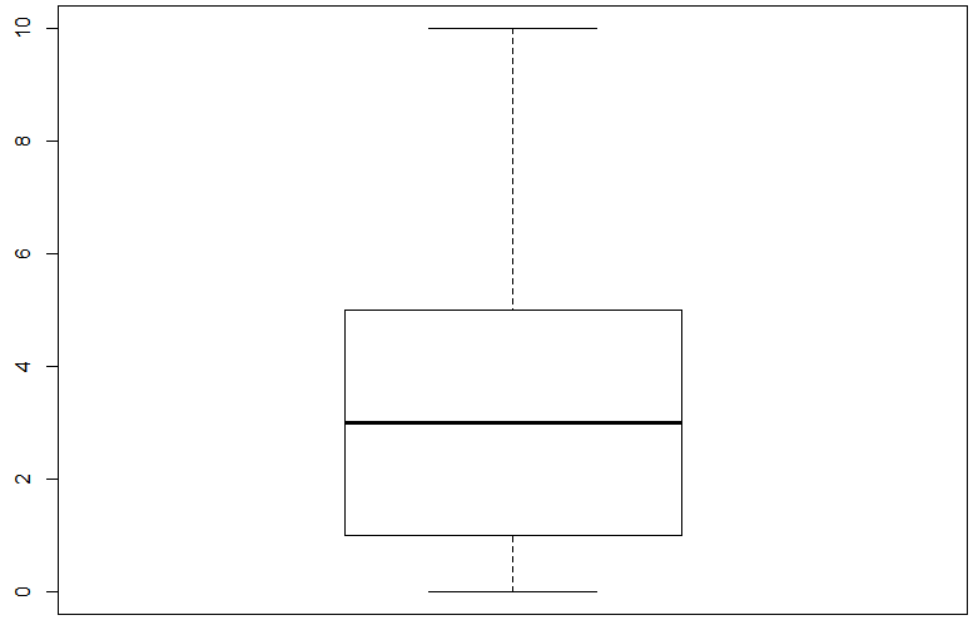

Рисунок 5 – «Ящик с усами переменной “Плотность трафика”»

На диаграмме «ящик с усами» переменной “Плотность трафика” видно, что значения смещены влево. Это означает, что большее количество наблюдений меньше среднего значения.

Асимметрия характеризует разброс значений: показывает перекос значений влево или вправо от среднего. По предыдущим диаграммам (Рисунок 1,2,3,4,5) видно, что распределение скошено влево и правый хвост распределения длиннее левого на Рисунке 1,2,3,4. На Рисунке 5 распределение скошено вправо и правый хвост распределения короче левого. Это можно наблюдать на гистограммах и диаграммах ядерной оценки функций плотности.

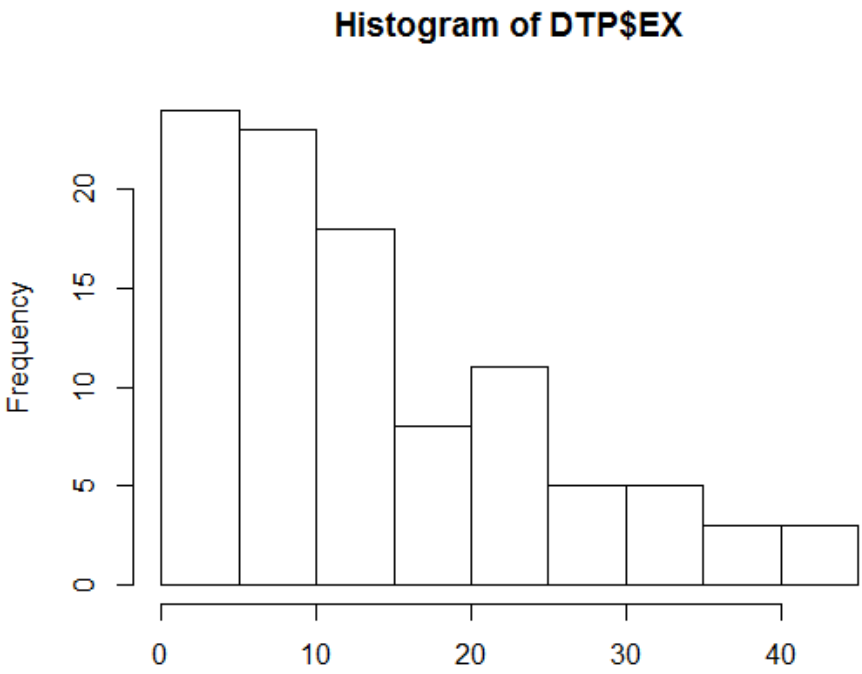

Рисунок 6 – «Гистограмма переменной “Стаж водителя”»

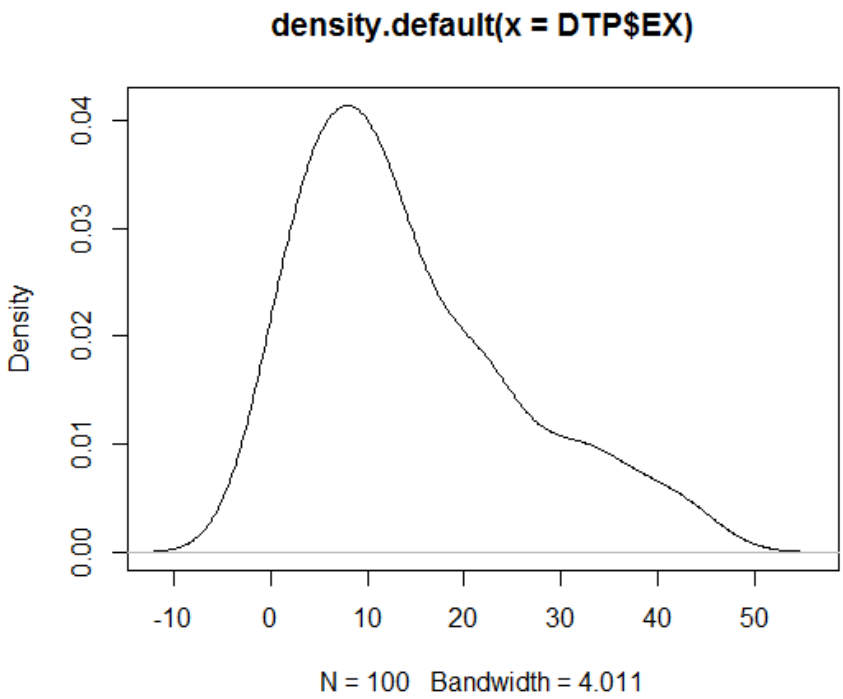

Рисунок 7 – «Кривая плотности переменной “Стаж водителя”»

Как видно из Рисунка 6 и 7, переменная «Стаж водителя» имеет асимметрию в распределении в отношении среднего значения.

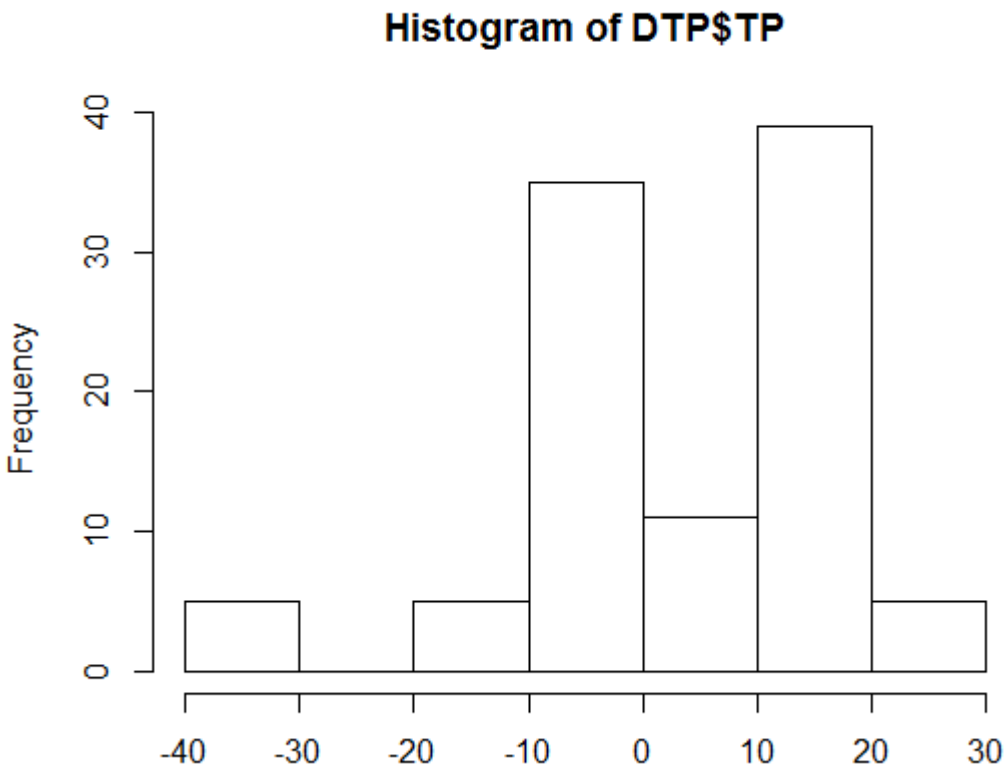

Рисунок 8 – «Гистограмма переменной “Температура воздуха”»

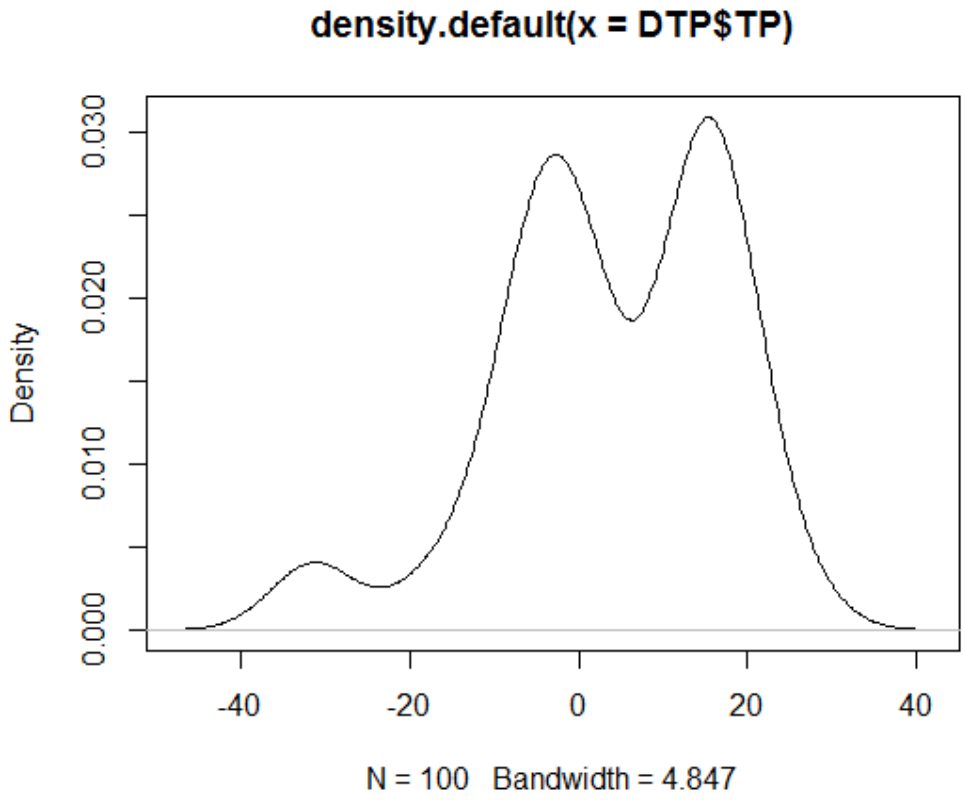

Рисунок 9 – «Кривая плотности переменной “Температура воздуха”»

Как видно из Рисунка 8 и 9, переменная «Возраст водителя» имеет бимодальное распределение.

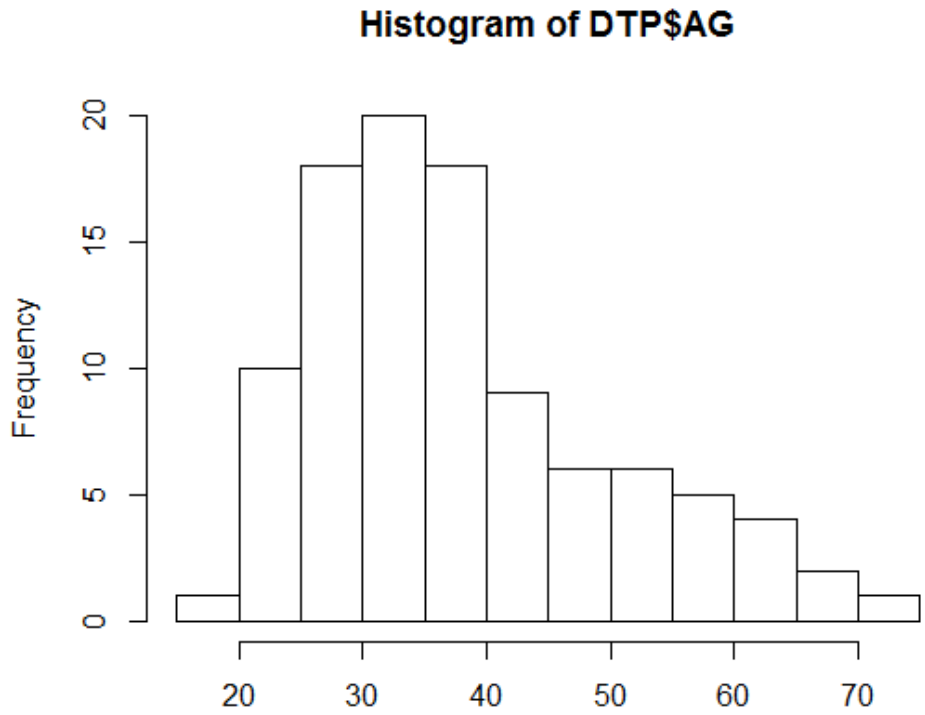

Рисунок 10 – «Гистограмма переменной “Возраст водителя”»

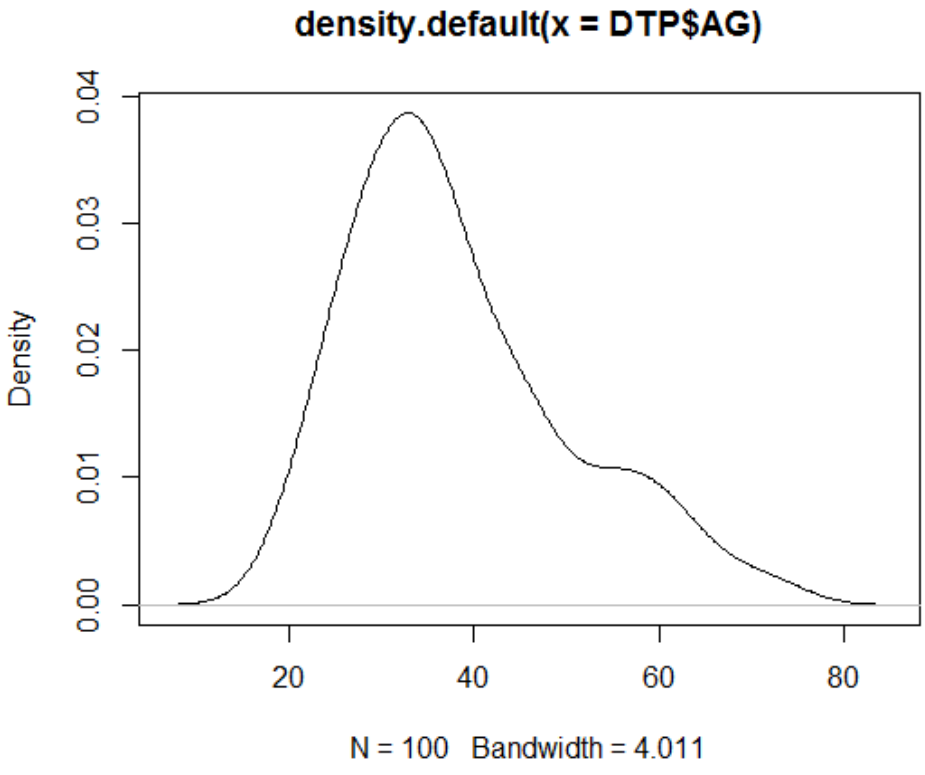

Рисунок 11 – «Кривая плотности переменной “Возраст водителя”»

Как видно из Рисунка 10 и 11 переменная «Возраст водителя» имеет асимметрию в распределении в отношении среднего значения.

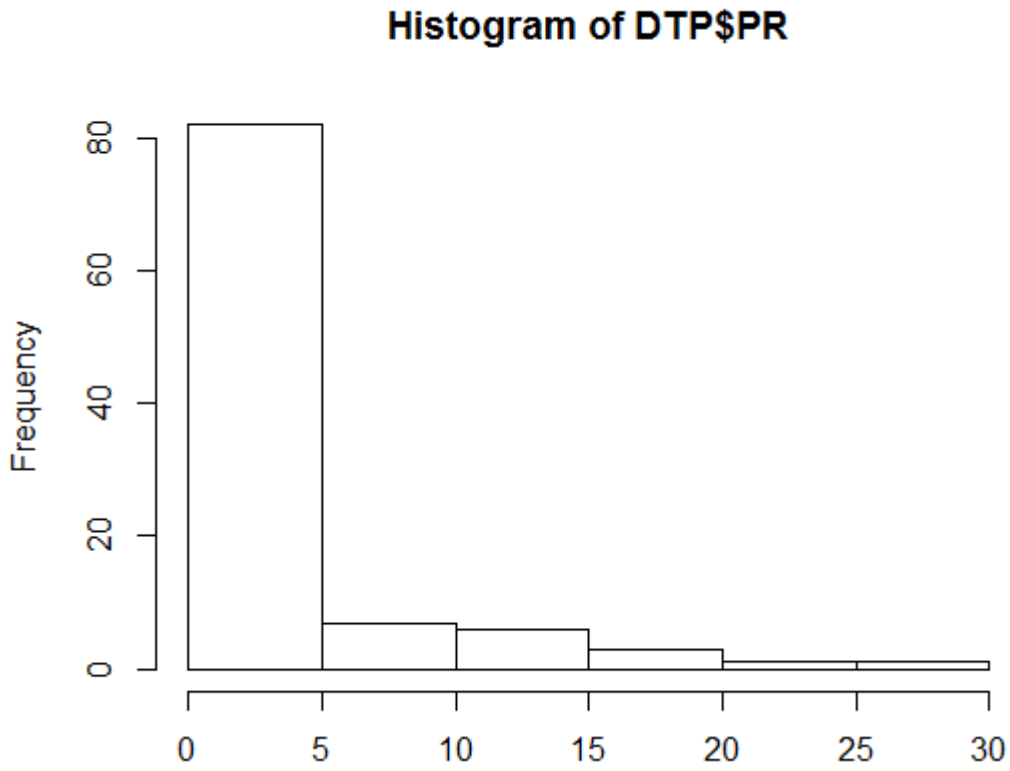

Рисунок 12 – «Гистограмма переменной “Количество осадков”»

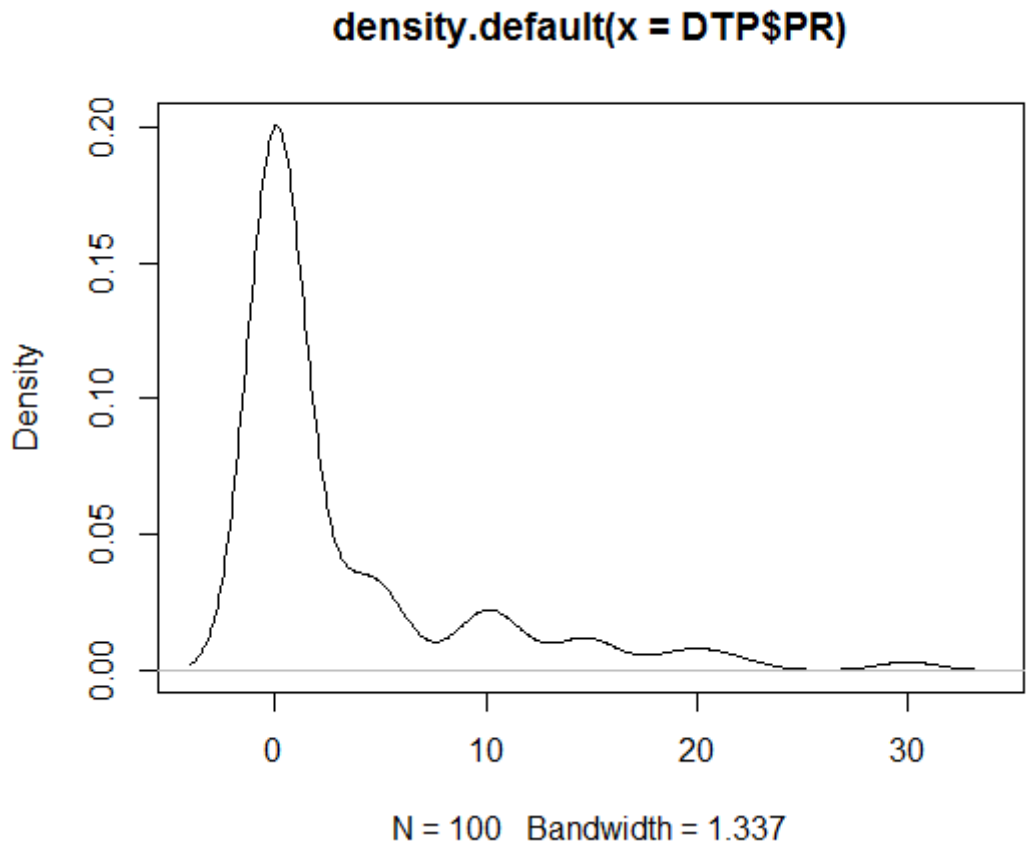

Рисунок 13 – «Кривая плотности переменной “Количество осадков”»

Как видно из Рисунка 12 и 13 переменная «Количество осадков» имеет распределение, значительно скошенное влево.

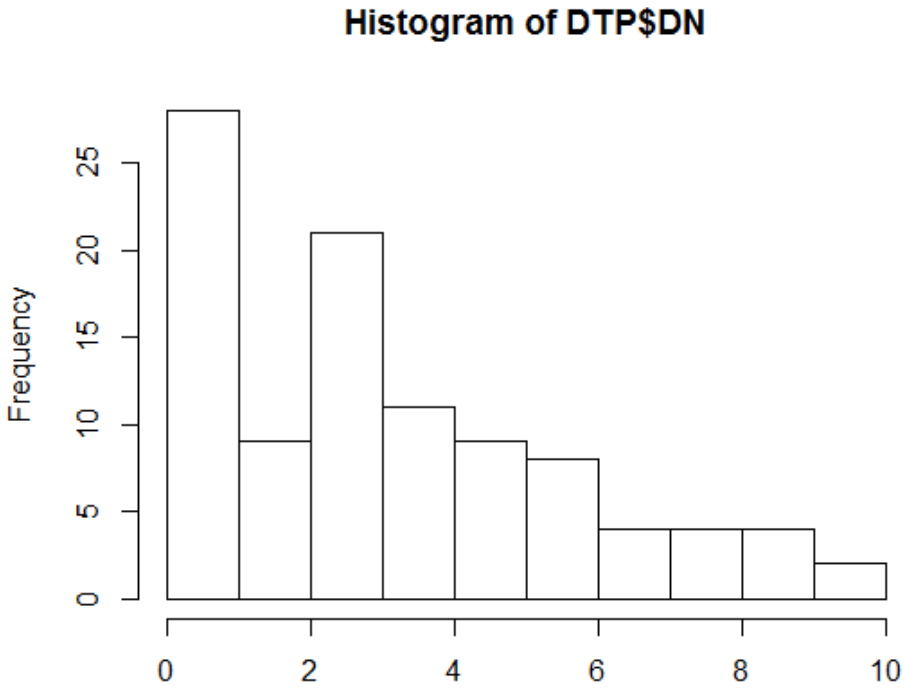

Рисунок 14 – «Гистограмма переменной “Плотность трафика”»

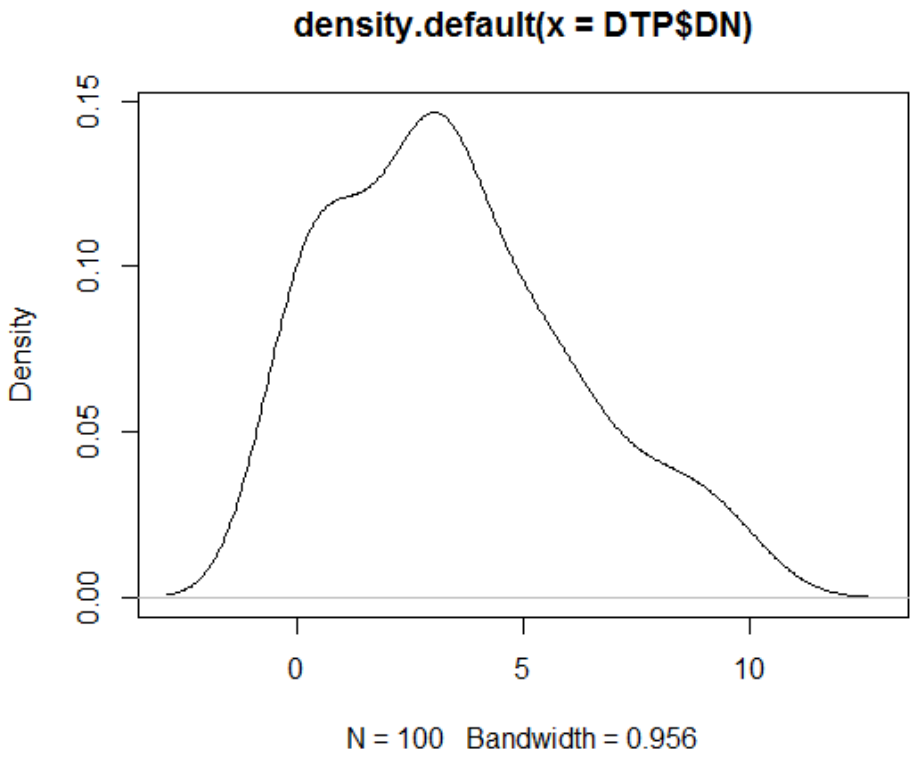

Рисунок 15 – «Кривая плотности переменной “Плотность трафика”»

Как видно из Рисунка 14 и 15 переменная «Плотность трафика» имеет асимметрию в распределении в отношении среднего значения.

Далее проанализируем парные взаимосвязи между переменными с помощью:
1) Корреляционной матрицы;
2) Матрицы диаграмм рассеяния;
3) Показателя VIF

1) Корреляционная матрица.

```
> cor(DTP)
     QT       EX         AG       TP       TM       PR       DN
QT    1    -0.07     -0.09   -0.127    0.019    0.17    0.16
EX -0.07      1       0.91    -0.08    -0.16     0.1     0.0003
AG -0.09     0.91       1     -0.15    -0.18     0.13   -0.07
TP -0.127  -0.08      -0.15      1     -0.086   -0.259   0.079
TM  0.019  -0.16      -0.18   -0.086      1      0.15   -0.44
PR  0.17    0.1        0.13   -0.259    0.15       1    -0.116
DN  0.16    0.000357 -0.076    0.079   -0.44    -0.116     1
```

Выделяется сильная положительная корреляция переменных AG и EX (Возраст водителя и стаж водителя). Данная ситуация носит название мультиколлинеарность — наличие линейной зависимости между объясняющими переменными (факторами) регрессионной модели.

Между другими переменными корреляция ниже среднего, значит переменные слабо влияют друг на друга.

2) Матрица диаграмм рассеяния.

```
> scatterplot.matrix(DTP, spread=FALSE, lty.smooth=2)
```

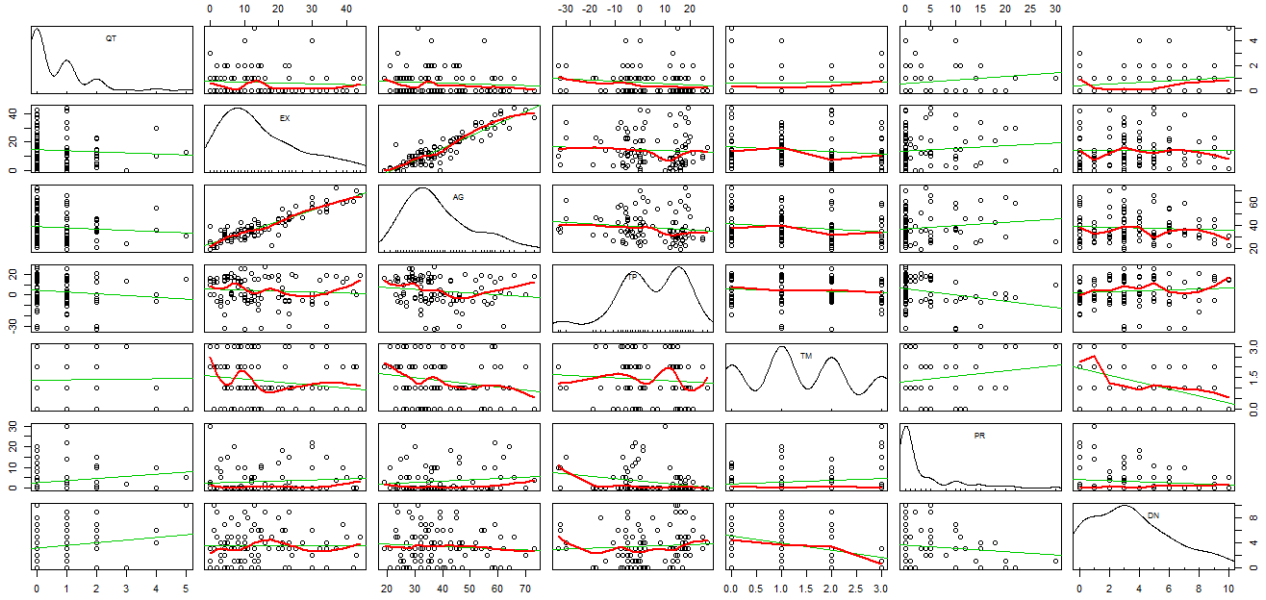

Рисунок 16 – «Матрица диаграмм рассеяния»

Функция создает диаграммы рассеяния (Рисунок 16) для всех пар переменных с нало женными сглаженной кривой и регрессионной прямой. На главной диагонали представлены диаграммы плотности и графики-щетки для каждой переменной.

В основном имеются унимодальные распределения, но также есть и бимодальные.

3) Показатель VIF

Мультиколлинеарность можно выявить при помощи статистики, называемой фактором инфляции дисперсии. Квадратный корень, извлеченный из этой статистики для любой независимой переменной, указывает на степень увеличения доверительного

интервала для параметра регрессии данной переменной по сравнению с моделью без скоррелированных независимых переменных (отсюда название статистики). Фактор инфляции дисперсии можно вычислить при помощи функции vif() из пакета car. Обычно значения квадратного корня этой статистики, превышающие 2, указывают на наличие мультиколлинеарности.

```
> fit<-lm(QT~EX+AG+TP+TM+PR+DN, data=DTP)
> vif(fit)
      EX       AG       TP       TM       PR       DN
6.234487 6.597793 1.114938 1.379408 1.111060 1.339069

> sqrt(vif(fit)) > 2
   EX    AG    TP    TM    PR    DN
 TRUE  TRUE FALSE FALSE FALSE FALSE
```

Из модели видно, что присутствует мультиколлинеарность. Уберем переменную AG (Возраст водителя) из модели, так как она сильно коррелирует с переменной EX (Стаж водителя). После создания новой модели без переменной AG проблема мультиколлинеарности решится.

```
> fit1<-lm(QT~EX+TP+TM+PR+DN, data=DTP)
> vif(fit1)
      EX       TP       TM       PR       DN
1.055580 1.080212 1.312950 1.106919 1.254997

> sqrt(vif(fit1)) > 2
   EX    TP    TM    PR    DN
FALSE FALSE FALSE FALSE FALSE
```

ГЛАВА 3

3.1 ПОСТРОЕНИЕ И ИНФЕРЕНЦИЯ О МОДЕЛИ РЕГРЕССИИ

Далее представлена оценка модели пуассоновской регрессии и выводы о качестве построенной модели.

Проанализируем зависимую переменную более подробно. Следующие команды позволяют получить диаграмму, представленную на рисунке 17.

```
> opar<-par(no.readonly = TRUE)
> par(mfrow=c(1,2))
> attach(DTP)
> hist(QT, breaks=20, xlab="Число летальных исходов", main="Частота")
> boxplot(QT~EX, xlab="Стаж водителя", main="Сравнение групп")
> par(opar)
```

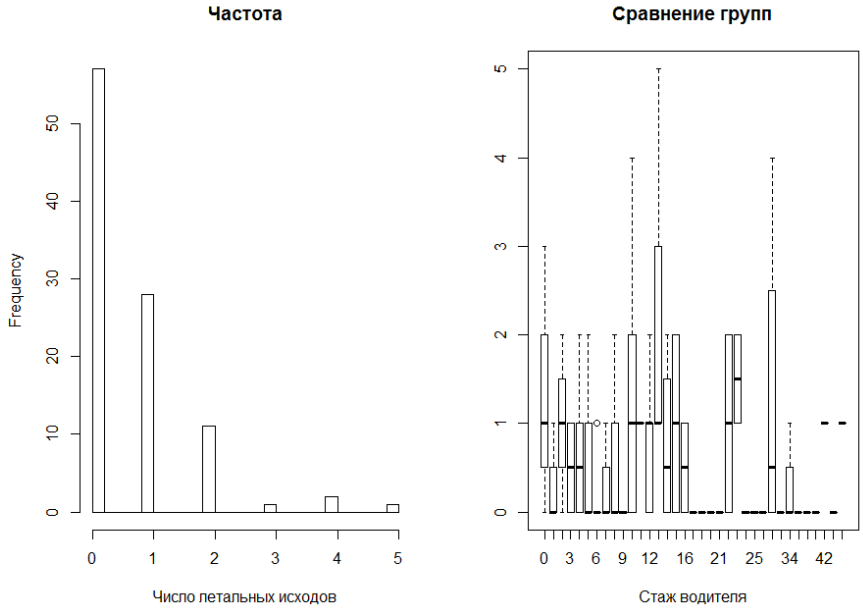

Рисунок 17 – «Распределение значений числа летальных исходов»

Хорошо заметно асимметричное распределение значений зависимой переменной, а также возможное наличие выбросов. На первый взгляд, число летальных исходов в ДТП, со стажем более 3-х лет, кажется меньшим и имеет меньший разброс данных. Для данных с пуассоновским распределением можно ожидать, что меньшая дисперсия будет сопряжена с меньшим средним значением. Гетерогенность дисперсии не представляет проблемы для пуассоновской регрессии, в отличие от стандартной МНК-регрессии.

Следующий шаг – это подгонка пуассоновской регрессионной модели:

```
> fit<-glm(QT~EX+AG+TP+TM+PR+DN, data=DTP, family = poisson())
> summary(fit)

Call:
glm(formula = QT ~ EX + AG + TP + TM + PR + DN, family = poisson(),
    data = DTP)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7725  -1.0328  -0.8634   0.5451   2.9301
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.350117   0.893509  -0.392   0.6952
EX           0.008608   0.029748   0.289   0.7723
AG          -0.021803   0.028727  -0.759   0.4479
TP          -0.012140   0.008864  -1.370   0.1708
TM           0.072549   0.148191   0.490   0.6244
PR           0.038764   0.018690   2.074   0.0381 *
DN           0.108740   0.051186   2.124   0.0336 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 130.21  on 99  degrees of freedom
Residual deviance: 117.83  on 93  degrees of freedom
AIC: 229.59

Number of Fisher Scoring iterations: 6
```

Выводятся девиаты, параметры регрессионной модели, среднеквадратичные отклоне ния и тесты, проверяющие, что параметры не равны нулю. Обратим внимание на то, что влияние независимой переменной PR и DN значимо на уровне $p < 0.05$.

Проведем интерпретацию параметров модели.

Коэффициенты модели можно вывести на экран при помощи функции coef() или в составе таблицы Coefficients, создаваемой функцией summary().

```
> coef(fit)
 (Intercept)            EX            AG            TP            TM            PR
-0.350117265  0.008607606 -0.021802955 -0.012140161  0.072548508  0.038764041
          DN
 0.108740053
```

В пуассоновской регрессии зависимая переменная моделируется как логарифм условного среднего $\log_e(\lambda)$. Регрессионный коэффициент для возраста (AG), равный -0.021802955, показывает, что каждый дополнительный год возраста при постоянных значениях числа летальных исходов в ДТП сопряжен с уменьшением логарифма среднего значения числа ДТП на 0.03.

Обычно гораздо проще интерпретировать регрессионные коэффициенты, когда они выражены в исходных единицах зависимой переменной (число летальных исходов, а не логарифм этого числа). Для этого коэффициенты нужно потенцировать:

```
> exp(coef(fit))
(Intercept)          EX          AG
  0.7046055   1.0086448   0.9784330
          TP          TM          PR
  -0.9879332   1.0752450   1.0395252
          DN
  1.1148725
```

Теперь видно, что увеличение возраста на один год увеличивает число летальных исходов в ДТП в 0.9784330 раз. Это значит, что старший возраст сопряжен с большим числом летальных исходов.

Из полученных результатов следует, что уравнение регрессии имеет следующий вид:

$$QT = 0.7046055 + 1.0086448 * EX + 0.9784330 * AG - 0.9879332 * TP +$$

$$+ 1.0752450 * TM + 1.0395252 * PR + 1.1148725 * DN$$

Далее приводится объяснение изменения каждого фактора.

Регрессионный коэффициент для переменной EX (стаж) равен 1.0086448. Это означает, что при увеличении стажа водителя на 1 год, количество летальных исходов увеличивается на 1.0086448. Этот коэффициент статистически не значим, так как Pr(>|t|) = 0.7723, больше 0.05.

Регрессионный коэффициент для переменной TP (температура) равен -0.9879332. Это означает, что при увеличении температуры на 1 градус по Цельсию, количество летальных исходов уменьшается на 0.9879332, а если температура снижается на 1 градус, то количество летальных исходов увеличивается на 0.9879332. Этот коэффициент статистически не значим, так как Pr(>|t|) = 0.1708, больше 0.05.

Регрессионный коэффициент для переменной TM (время) равен 1.0752450, это означает, что если временной промежуток времени отличается от базового (0 - Утро), то количество летальных исходов будет увеличиваться на 1.0752450. Этот коэффициент статистически не значим, так как Pr(>|t|) = 0.6244 , больше 0.05.

Регрессионный коэффициент для переменной PR (количество осадков) равен 1.0395252. Это означает, что при увеличении количества выпавших осадков на 1%, количество летальных исходов увеличивается на 1.0395252. Этот коэффициент статистически значим на уровне меньше 0.05, так как Pr(>|t|) = 0.0381.

Регрессионный коэффициент для переменной DN (плотность трафика) равен 1.1148725. Это означает, что при увеличении плотности трафика на 1 балл, количество летальных исходов при ДТП увеличивается на 1.1148725. Этот коэффициент статистически значим на уровне меньше 0.05, так как Pr(>|t|) = 0.0336.

Коэффициент, равный 0.7046055 формально показывает прогнозируемый уровень зависимой переменной (количество летальных исходов в ДТП), но только в том случае, если объясняющие переменные равны 0 и находятся близко с выборочными значениями. Но если они находятся далеко от выборочных значений, то буквальная интерпретация может привести к неверным результатам.

Важно помнить, что, как и экспоненциальные параметры в логистической регрессии, экспоненциальные параметры в пуассоновской регрессии оказывают мультипликативный, а не аддитивный эффект на зависимую переменную.

3.2 ПРОВЕРКА ДАННЫХ НА НАЛИЧИЕ НЕОБЫЧНЫХ НАБЛЮДЕНИЙ

Выброс – это значение, которое плохо предсказывается подобранной моделью (то есть имеет большой положительный или отрицательный остаток).

Необходимо построить графики остатков.

```
> rstud <- rstandard(fit1)
> rstud <- rstandard(fit1)
> rjack <- rstudent(fit1)
> par(mfrow=c(2,2))
> plot(fit1$res,ylab="raw residuals")
> plot(rstud,ylab="studentized residuals")
> plot(rjack,ylab="jacknife residuals")
```

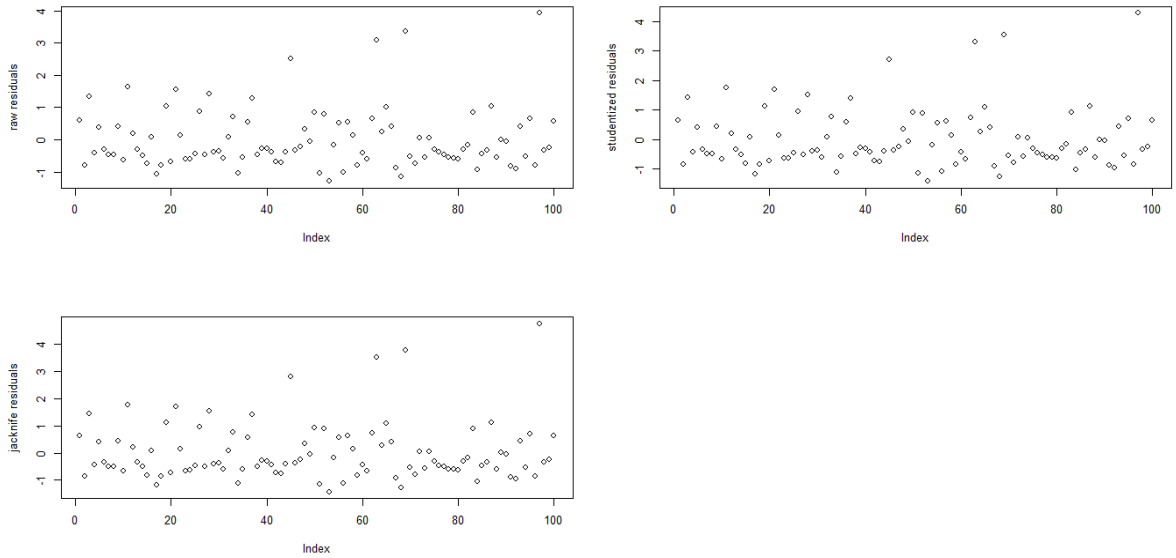

Рисунок 18 – «Стандартные и Стьюденизированные остатки по методу складного ножа от номера наблюдения»

Стьюдентизированные остатки (Рисунок 18) представляют собой частное от деления обычного остатка на оценку его стандартного отклонения. Также на графике видно, что имеется несколько выбросов, которые выходят за пределы основной массы наблюдений.

Степень уверенности в результатах зависит от степени соответствия данных допущениям, лежащим в основе статистических тестов. Данные не очень хорошо укладываются в 95% доверительные границы, это значит, что требование нормального распределения выполняется недостаточно хорошо.

На диаграмме qqPlot (Рисунок 19) видно, что в основном все значения находятся близко к линии уравнения регрессии. Есть несколько значений, которые достаточно отдалены от этой линии и выходят за пределы доверительных границ.

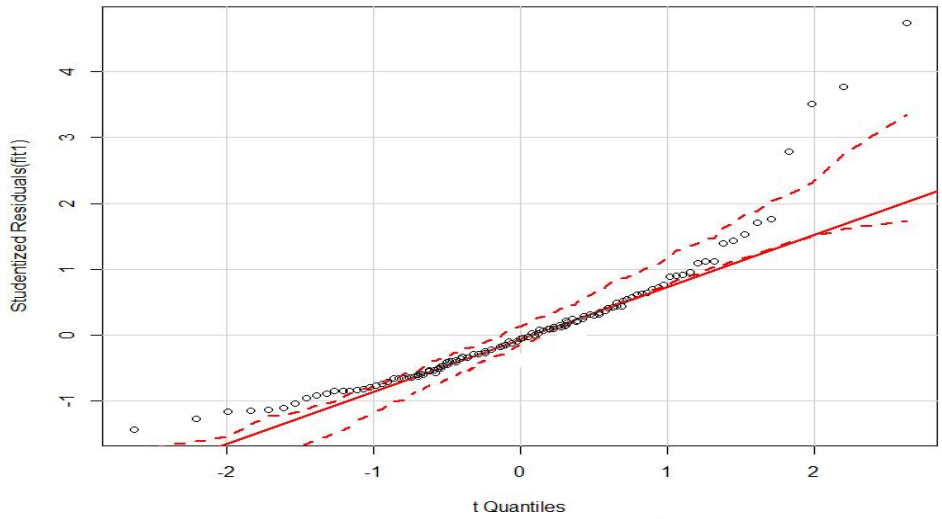

Рисунок 19 – «Диграмма qqPlot»

Вычисляем значение вероятности статистической ошибки первого рода с поправкой Бонферрони для наибольших остатков Стьюдента:

```
> library(car)
> fit1<-lm(QT~EX+TP+TM+PR+DN, data=DTP)
> outlierTest(fit1)

   rstudent unadjusted p-value Bonferonni p
97 4.737172         7.7568e-06   0.00077568
69 3.768123         2.8846e-04   0.02884600
```

Функция outlierTest проверяет на значимость самый большой выброс в указанной модели методом Бонферрони, в котором двусторонняя вероятность нулевой гипотезы умножается на размер выборки. Смысл процедуры Бонферрони в том, что по теории вероятности мы должны ожидать какое-то количество выбросов, например, 5% из 100%; чем больше выборка, тем больше будет выбросов. Если у нас всего одно наблюдение и вероятность выброса 5%, то при 2-х наблюдениях вероятность выброса уже 10% и т.д. Вероятность по Бонферрони поэтому правильнее. После удаления данного выброса проведем тест еще раз:

```
rstudent unadjusted p-value Bonferonni p
65   4.28035         2.2207e-05     0.011836
```

Видим, что вероятность по Бонферрони значима (< 5%) значит имеется выброс ID которого равен 65. После удаления данного выброса проведем тест еще раз:

```
rstudent unadjusted p-value Bonferonni p
49  3.999395         7.2692e-05     0.038672
```

Видим, что вероятность по Бонферрони значима (< 5%) значит имеется выброс ID которого равен 49. После удаления данного выброса проведем тест еще раз:

```
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
rstudent unadjusted p-value Bonferonni p
40  3.908549         0.00010513     0.055826
```

Как видно, вероятность по Бонферрони не значима (5.58% > 5%) и выбросов больше не имеется.

Проверим данные на наличие необычных наблюдений с помощью расстояния Кука и диаграммы добавленных переменных.

Существуют два метода обнаружения влиятельных наблюдений: расстояние Кука (или D-статистика) и диаграммы добавленных переменных. Грубо говоря, значения расстояния Кука, превышающие 4/(n – k – 1), где n – объем выборки, а k – число независимых переменных, свидетельствуют о влиятельных наблюдениях. Построить диаграмму расстояний Кука можно при помощи следующего программного кода:

```
>cutoff <- 4/(nrow(DTP)-length(fit1$coefficients)-2)
>plot(fit1, which=4, cook.levels=cutoff)
>abline(h=cutoff, lty=2, col="red")
```

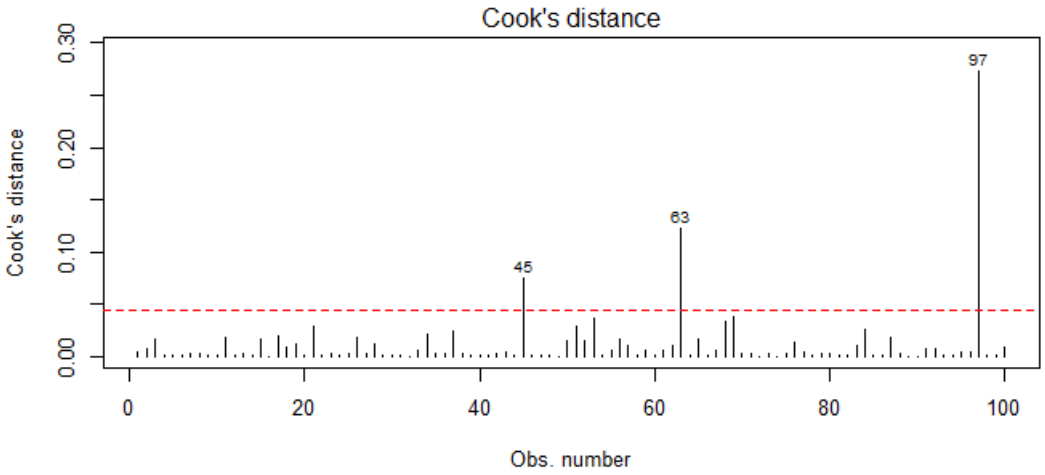

Рисунок 20 – «Диаграмма расстояний Кука»

Из рисунка 20 видно, что сильно превышают линию значения 45, 63, 97. Это влиятельные наблюдения. Удаление этих моделей заметно влияет на значения свободного члена и угловых коэффициентов в регрессионной модели.

Диаграммы дистанции Кука помогают обнаружить влиятельные наблюдения, но они не позволяют понять, как эти наблюдения влияют на модель. В этой ситуации приходят на выручку диаграммы добавленных переменных. Для одной зависимой и k независимых переменных описанным ниже способом создается k диаграмм добавленных переменных.

Для каждой независимой переменной X_k отображаются остатки от регрессии зависимой переменной по остальным k – 1 независимым переменным. Такие диаграммы добавленных переменных (Рисунок 21) можно построить при помощи функции avPlots() из пакета car:

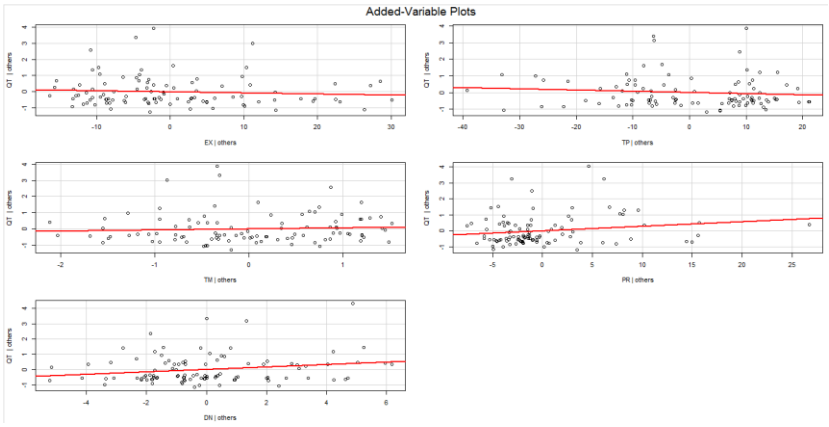

Рисунок 21 – «Диаграммы добавленных переменных»

Прямая на каждой диаграмме – это регрессионный коэффициент для данной независимой переменной. Вклад влиятельных наблюдений можно оценить, если представить, как изменится линия, если удалить точку, соответствующую данному наблюдению.

Точки с высокой напряженностью – это выбросы в отношении других независимых переменных. Иными словами, они характеризуются необычным сочетанием значений независимых переменных. Значение зависимой переменной не используется при вычислении напряженности.

Наблюдения с высокой напряженностью идентифицируются при помощи показателя влияния (hatstatisctic). Для определенного набора данных среднее значение этой статистики вычисляется как p/n, где p –это число параметров в модели (включая свободный член), а n – размер выборки. Наблюдения, для которых значение этой статистики превышает среднее в два или три раза, должны быть проанализированы.

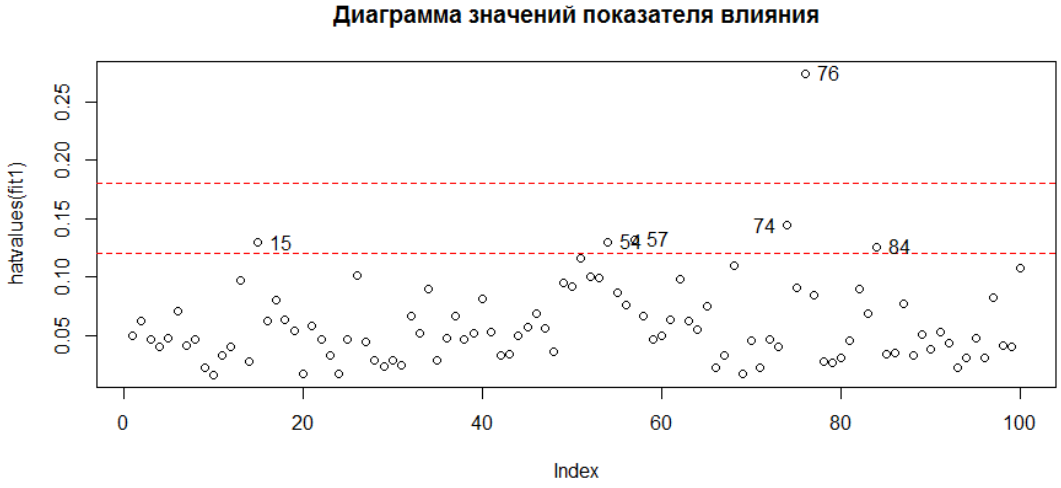

Рисунок 22 – «Диаграмма значений показателя влияния»

Здесь видно (Рисунок 22), что особо необычными являются наблюдения 15, 76, 74, 54, 57, 84. Можно предположить, что эти модели выделяются из-за отсутствия стажа водителя, а, следовательно, и отсутствия водительского удостоверения.

Наблюдение 15 характерно тем, что виновник ДТП имеет большой опыт (30 лет), но при этом совершил ДТП в нормальных дорожных условиях.

Наблюдение 76 характерно тем, что виновник ДТП не имеет водительских прав, при этом совершил ДТП в дождливую погоду.

Наблюдение 74 характерно тем, что виновник ДТП имеет большой опыт (30 лет), но при этом совершил ДТП в нормальных дорожных условиях.

Наблюдение 54 характерно тем, что виновник ДТП совершил ДТП с летальным исходом в холодную погоду (-32 градуса) в плотном трафике.

Наблюдение 57 характерно тем, что виновник ДТП совершил ДТП с летальным исходом в спокойной дорожной обстановке, практически без других участников движения.

Наблюдение 84 характерно тем, что виновник ДТП с большим стажем (39 лет) совершил ДТП в плотном трафике в благоприятных погодных условиях.

Можно свести информацию о выбросах, точках с высокой напряженностью и влиятельных наблюдениях на одну чрезвычайно информативную диаграмму при помощи функции influencePlot.

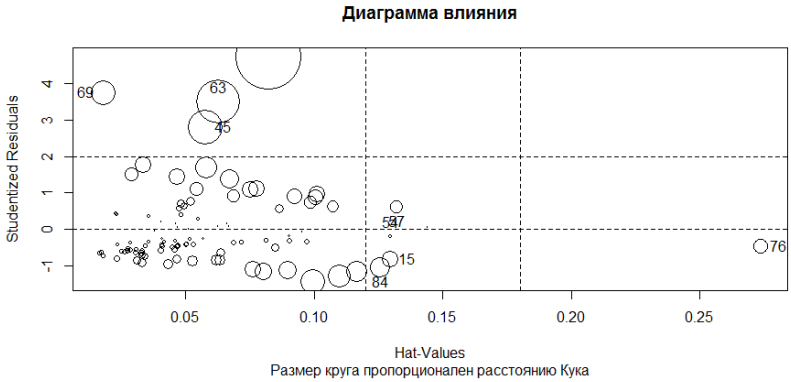

Рисунок 23 – «Диаграмма влияния»

На полученной диаграмме (Рисунок 23) видно, что 63, 69 и 45 наблюдение – это выбросы, а 76, 15, 84 и 87 наблюдения характеризуются высокой напряженностью, а 63 и 45 – влиятельные наблюдения.

3.3 ДИАГНОСТИКА РЕГРЕССИОННЫХ МОДЕЛЕЙ НА ВЫПОЛНЕНИЕ СТАНДАРТНЫХ УСЛОВИЙ НА ОСТАТКИ

1. Далее проведен тест на гомоскедастичность (тест Уайта).

```
> ncvTest(fit1)

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 8.915447    Df = 1     p = 0.11427686
```

Функция ncvTest() позволяет проверить гипотезу о постоянстве дисперсии остатков как альтернативу тому, что дисперсия остатков изменяется в зависимости от подобранных значений. Статистически значимый результат свидетельствует о гетероскедастичности (неоднородности дисперсии остатков).

Так как p-value>0.05, то гипотезу о гетероскедастичность отклоняем - модель гомоскедастична. В данном случае признаки гетероскедастичности отсутствуют, и предлагаемая степень близка к 1 (никакого преобразования не требуется).

2. Кроме того, осуществлен тест Гольдфельдта-Квандта. Тест Голдфелда-Квандта — процедура тестирования гетероскедастичности случайных ошибок регрессионной модели, применяемая в случае, когда есть основания полагать, что стандартное отклонение ошибок может быть пропорционально некоторой переменной. Тест также основывается на предположении нормальности распределения случайных ошибок регрессионной модели. Фактически это F-тест, поскольку статистика теста имеет распределение Фишера.

```
>gqTest(fit1)

Goldfeld-Quandt test
data:  reg
GQ = 0.86487, df1 = 217, df2 = 232, p-value = 0.7344
```

Так как результаты получили незначимые (p-value = 0.7344), то гипотеза о гетероскедастичности отклоняется, отсюда следует, что модель гомоскедастична.

3. Проведен тест на автокорреляцию, а именно тест Дарбина - Уотсона и Бройша - Годфри (LM test).

Тест Дарбина-Уотсона – статистический критерий, используемый для нахождения автокорреляции остатков первого порядка регрессионной модели. Наблюдения, сделанные в короткие отрезки времени, более сильно коррелируют друг с другом, чем разнесенные во времени наблюдения.

```
> durbinWatsonTest(fit1)
 lag Autocorrelation D-W Statistic p-value
   1      -0.1858299      2.363579   0.114
 Alternative hypothesis: rho != 0
```

Высокое значение p-value = 0.114 свидетельствует об отсутствии автокорреляции и, следовательно, о независимости остатков.

Бройша-Годфри LM test - в данном тесте случайные ошибки не обязательно должны быть нормально распределены. Тест является асимптотическим, то есть для достоверности выводов требуется большой объём выборки. Особенность данного теста заключается в том, что его можно использовать практически всегда. Если значение статистики превышает критическое значение, то автокорреляция признаётся значимой, в противном случае она незначима.

```
> bgtest(fit1)

Breusch-Godfrey test for serial correlation of order up to 1
data:  reg
LM test = 0.72961, df = 1, p-value = 0.223
```

Высокое значение p-value = 0.223 свидетельствует о том, что нулевая гипотеза не отвергается. За нулевую гипотезу принимали гипотезу об отсутствии автокорреляции. Значит автокорреляция отсутствует.

1. Проведем проверку на нормальность распределения остатков.

Для проверки нормальности распределения нужно использовать Q-Q-диаграмму:

```
> qqPlot(fit1)
```

Результат представлен на рисунке 24.

Функция qqPlot() является более аккуратным методом проверки предположения о нормальности. Она изображает связь между остатками Стьюдента и квантилями распределения Стьюдента с n – p – 1 степенями свободы, где n –это объем выборки, а p –число параметров регрессии (включая свободный член).

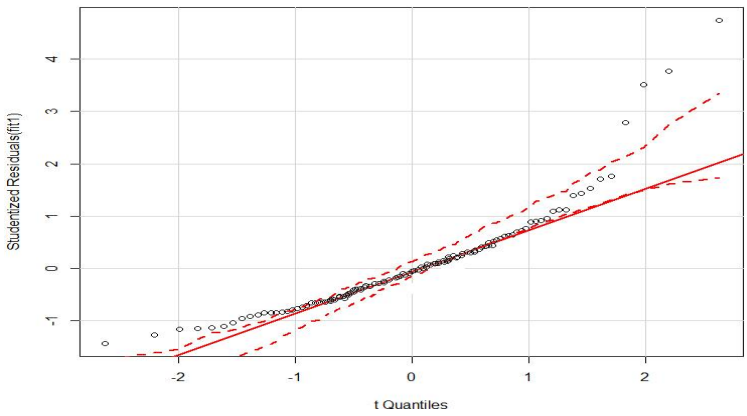

Рисунок 24 – «Диаграмма qqPlot»

Большая часть наблюдений попадает в рамки доверительного интервала, что свидетельствует о нормальности распределения.

Функция residplot() создает гистограмму остатков Стьюдента с наложенными кривой нормального распределения, кривой ядерной оценки функции плотности и графиком-щеткой.

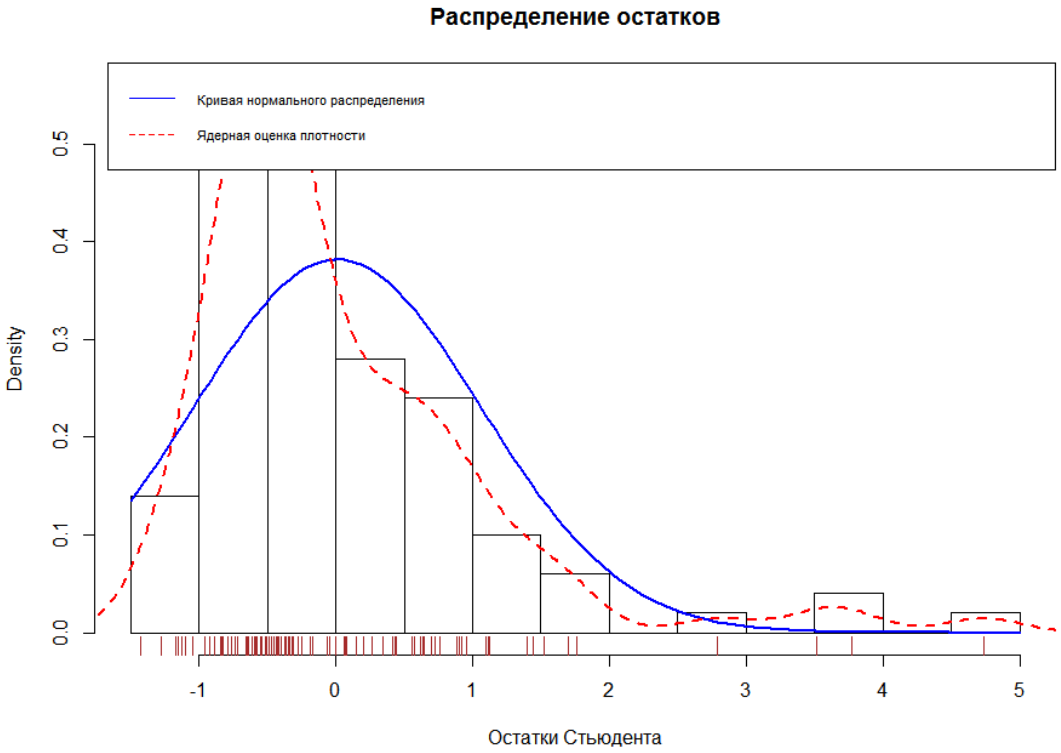

Рисунок 25 – «Распределение остатков Стьюдента, изображенное при помощи функции residplot()»

Как видно из рисунка 25, остатки не имеют нормального распределения – наблюдается значительная асимметрия.

3.4 ВЫБОР «ЛУЧШЕЙ РЕГРЕССИОННОЙ МОДЕЛИ»

Проведем проверку нелинейной связи между зависимой и независимыми переменными с помощью диаграммы частных остатков (Рисунок 26).

Наличие нелинейной связи между зависимой и независимыми переменными можно проверить при помощи диаграмм компонент и остатков (также известных под названием диаграммы частных остатков). Нелинейность свидетельствует о том, что возможно некорректно смоделирована функциональная форма этой независимой переменной в уравнении.

```
> crPlots(fit1)
```

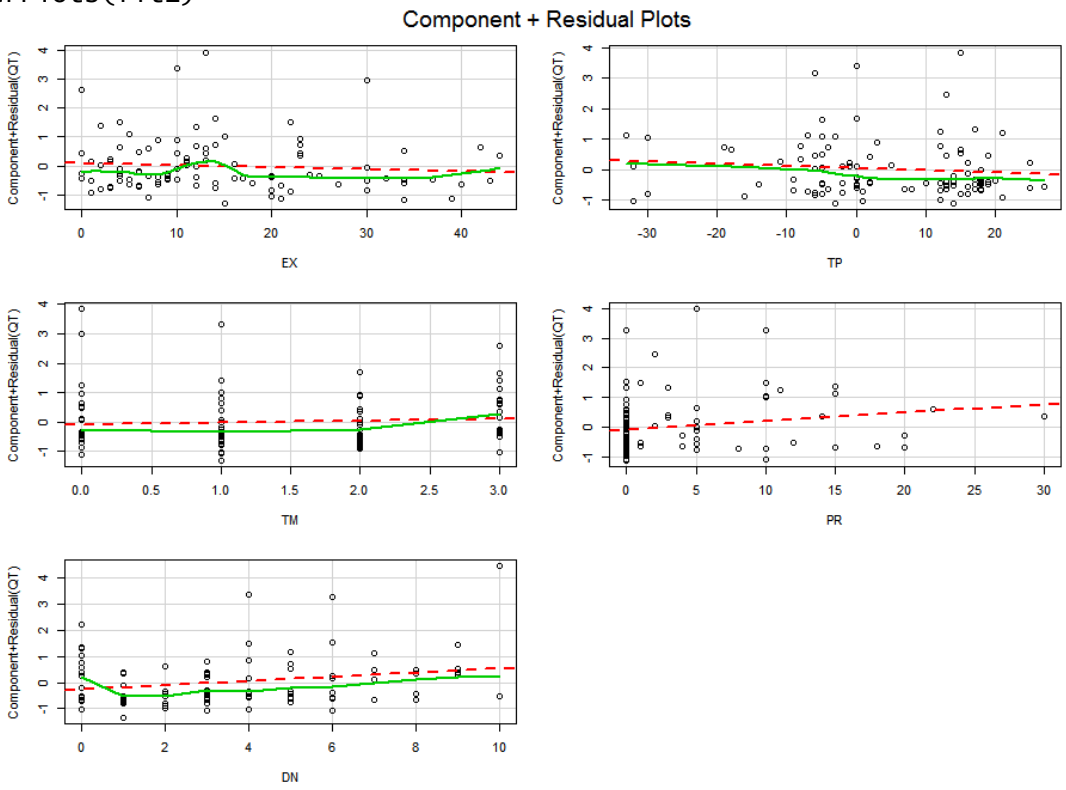

Рисунок 26 – «Диаграммы компонент и остатков»

1. RESET тест Рамсея

Необходимо проверить гипотезу: H_0: $a_2=a_3=\ldots=a_m=0$. Если значение статистики больше критического, то нулевая гипотеза отвергается, и спецификация модели признается неверной. В противном случае функциональная форма модели является приемлемой.

```
> resettest(fit1)

RESET test
data:  reg
RESET = 1.9912, df1 = 2, df2 = 454, p-value = 0.05763
```

Результат данного теста говорит о том, что p-value незначим, откуда следует вывод о том, что нулевая гипотеза отвергается, спецификация модели признается неверной.

2. J-тест Дэвидсона

Пусть даны две невложенные регрессионные модели, то есть ни одна из них не является частным случаем другой.

Идея критерия Давидсона состоит в следующем: если первая модель содержит верный набор свободных переменных, то включение восстановленных значений второй модели в этот набор не должно приводить к значимым улучшениям. Но если это так, то, возможно, первая модель не является верной.

Таким образом, для сравнения обеих моделях необходимо добавить значения первой модели во вторую и наоборот. Тестовая статистика критерия проверяет значимость восстановленных значений в расширенной модели, обычно для этого используется t-статистика Стьюдента.

```
J test

Model 1: QT ~ EX + AG + TP + TM + PR + DN
Model 2: WAGE ~ AG + TP + TM + PR + DN
                 Estimate Std. Error t value Pr(>|t|)
M1 + fitted(M2) -0.70109    0.42802 -1.6380  0.04203 *
M2 + fitted(M1)  1.00000    0.44748  2.2347  0.03585 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Как видно из результатов, модель является значимой.

3. Тест Кокса-Бокса

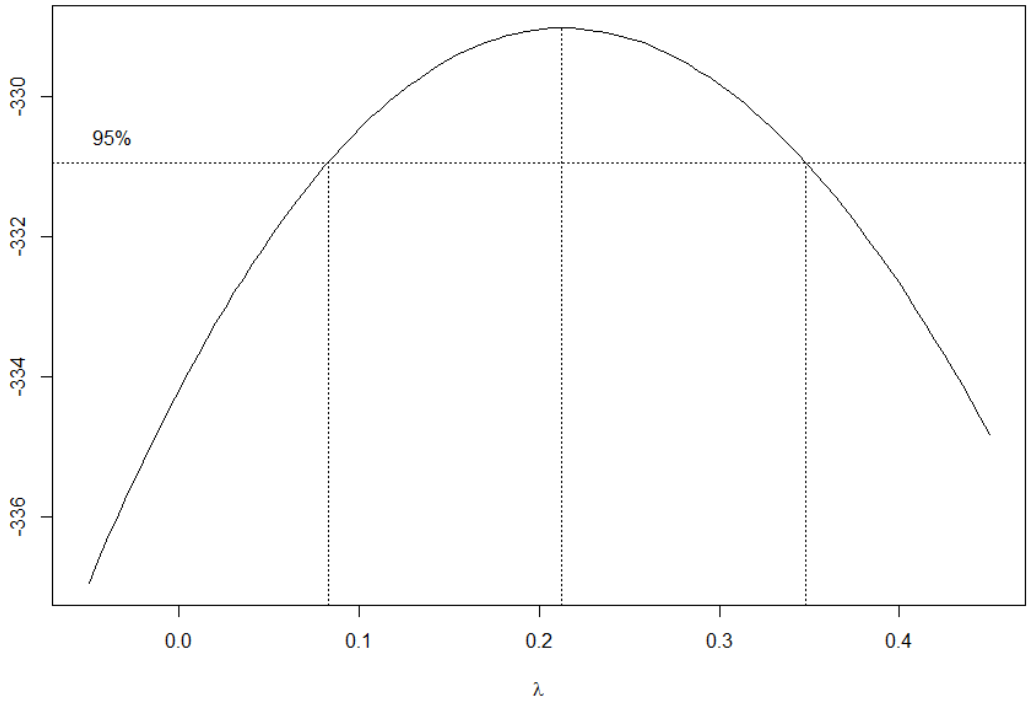

Рисунок 27 – «График теста Бокса-Кокса»

Если параметр равен единице, то данная функция принимает вид F=y-1. В том случае, если параметр стремиться к нулю, то данная функция принимает вид F=log(y).

Как видно из Рисунка 27, функция стремится к 1, значит данная функция принимает вид F=y-1.

4. Далее проведем оценку альтернативной спецификации модели

```
> RMSE1 = sqrt(sum((DTP$QT - predict(fit1))^2)/nrow(DTP))
RMSE1 = 0.929111646690042
> RMSE2 = sqrt(sum((DTP$QT - predict(fit))^2)/nrow(DTP))
RMSE2 = 0.927266986878504
```

RMSE показывает, насколько велика в среднем разница между действительными наблюдениями и значениями, предсказанными моделью. То есть, чем меньше RMSE, тем точнее предсказания.

Из полученных результатов видно, что вторая модель лучше.

В модели множественной регрессии коэффициенты для переменных были значимыми. Функция AIC не требует вложенных моделей, поэтому далее будем использовать именно ее.

```
> AIC(fit, fit1)
     df      AIC
fit   8 284.6850
fit1  7 283.0824
```

Как видно, модель без незначимых переменных (fit1) лучше полной модели.

5. Далее проведено сформирование окончательного набора независимых переменных при помощи пошагового метода исключения переменных – «backward».

```
> stepAIC(fit, direction="backward")
Start:  AIC=-1.1
QT ~ EX + AG + TP + TM + PR + DN
       Df Sum of Sq    RSS      AIC
- EX    1   0.04052 86.023 -3.05563
- TM    1   0.18133 86.164 -2.89208
- AG    1   0.34244 86.325 -2.70527
- TP    1   1.07516 87.058 -1.86006
<none>              85.982 -1.10275
- PR    1   2.65053 88.633 -0.06667
- DN    1   2.86978 88.852  0.18039

Step:  AIC=-3.06
QT ~ AG + TP + TM + PR + DN

       Df Sum of Sq    RSS     AIC
- TM    1   0.20547 86.228 -4.8171
- AG    1   0.95312 86.976 -3.9537
- TP    1   1.03848 87.061 -3.8557
<none>              86.023 -3.0556
- PR    1   2.64306 88.666 -2.0294
- DN    1   3.13132 89.154 -1.4802

Step:  AIC=-4.82
QT ~ AG + TP + PR + DN

       Df Sum of Sq    RSS     AIC
- TP    1    1.1026 87.331 -5.5464
- AG    1    1.3009 87.529 -5.3197
<none>              86.228 -4.8171
- PR    1    2.8918 89.120 -3.5184
- DN    1    3.0821 89.310 -3.3051

Step:  AIC=-5.55
QT ~ AG + PR + DN

       Df Sum of Sq    RSS     AIC
- AG    1    1.0468 88.378 -6.3549
<none>              87.331 -5.5464
- DN    1    2.9259 90.257 -4.2509
```

```
- PR    1    4.0379 91.369 -3.0264

Call:
lm(formula = QT ~ PR + DN, data = DTP)

Coefficients:
(Intercept)           PR           DN
    0.32246      0.03217      0.06743
```

В столбце AIC приведено значение одноименного критерия для модели, из которой удалена указанная в соответствующей строке переменная. Постепенно по шагам удалились переменные EX (стаж), TM (время), TP (температура), что привело к улучшению модели.

Регрессии по всем подмножествам.

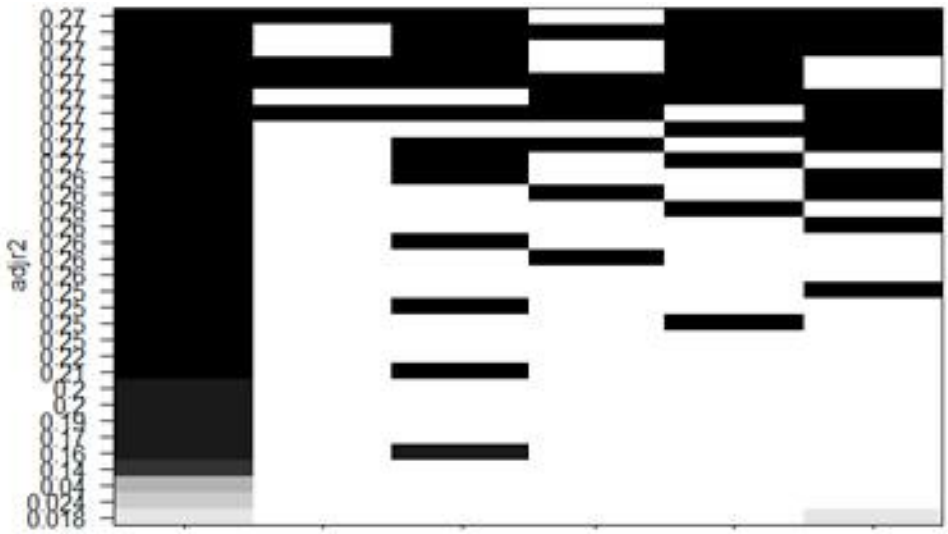

Рисунок 28 – «Диаграмма регрессии по подмножествам»

Кроме того, проведена регрессия по всем подмножествам. Регрессия по всем подмножествам проводится (Рисунок 28) при помощи функции regsubsets() из пакета leaps. В качестве критерия «лучшей» модели можно выбрать коэффициент детерминации, скорректированный коэффициент детерминации или Cp-статистику Мэллоуса (Mallows Cp statistic).

```
>library(leaps)
>leaps <-regsubsets(QT ~ EX + AG + TP + TM + PR + DN, data=DTP,
nbest=4)
>plot(leaps, scale="adjr2")
>library(car)
>subsets(leaps, statistic="cp",
main="Статистика Мэллоуса для регрессии по всем подмножествам")
>abline(1,1,lty=2,col="red")
```

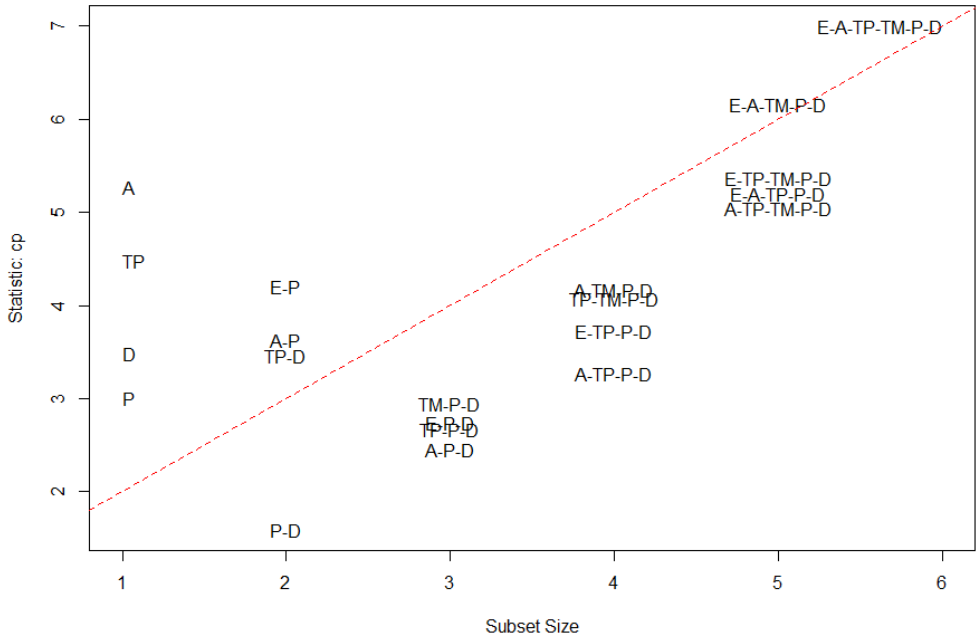

Рисунок 29 – «График теста Бокса-Кокса»

Из диаграммы (Рисунок 29) видно, что модель, включающая все переменные, является наилучшей и имеет наибольшей R^2.

6. Далее проведен тест на устойчивость параметров модели

Нулевая гипотеза формулируется как утверждение о том, что качество общей модели регрессии без ограничений лучше качества частных моделей регрессии.

```
F value        d.f.1        d.f.2      P value
1.084875e+02 2.000000e+00 1.064000e+03 1.322216e-43
```

P-value имеет малое значение, отсюда можно сделать вывод о том, что качество общей модели регрессии лучше, чем частных моделей.

3.5 АНАЛИЗ ОТНОСИТЕЛЬНОЙ ВАЖНОСТИ НЕЗАВИСИМЫХ ПЕРЕМЕННЫХ

Проведено сравнение стандартизованных коэффициентов регрессии. Стандартизированные коэффициенты регрессии - это коэффициенты, деленные на стандартное отклонение.

Таким образом, приведенное сравнение абсолютных величин, стандартизованных коэффициентов регрессии позволяет получить пусть и довольно грубое, но достаточно наглядное представление о важности рассматриваемых факторов.

```
> QT <- as.data.frame(scale(DTP))
> zfit<-lm(QT~EX+AG+TP+TM+PR+DN, data=QT)
> coef(zfit)
  (Intercept)            EX            AG            TP
-6.664481e-17  5.172046e-02 -1.546724e-01 -1.126636e-01
           TM            PR            DN
 5.146417e-02  1.765863e-01  2.017193e-01
```

На основе полученных данных (Рисунок 29) видно, что наибольшую степень важности имеют переменные PR (количество осадков) и DN (плотность трафика) и TP (температура).

```
    Weights
EX  4.007357
AG  7.975801
TP 16.101756
TM  3.346378
PR 33.874980
DN 34.693728
```

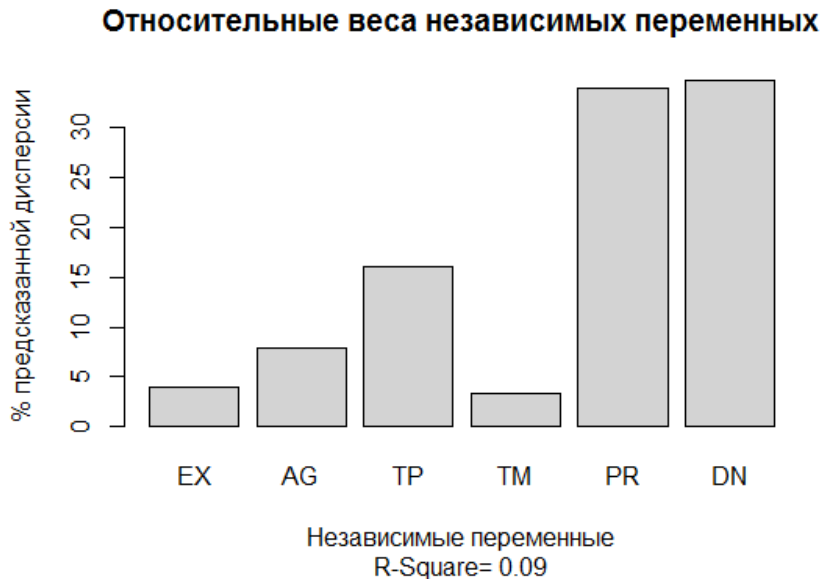

Рисунок 30 – «Относительные веса независимых переменных»

В программном коде функция relweights() применяется для предсказания летальных исходов в ДТП по стажу водителя, возрасту водителя, количеству осадков и температурам, плотности трафика, времени суток в наборе данных DTP. Из рисунка 30 следует, что общая доля объясненной моделью дисперсии (коэффициент детерминации, R-Square = 0.09) распределяется между независимыми переменными. На долю количества осадков приходится 33.9% коэффициента детерминации, на долю плотность трафика – 34.7% и так

далее. Если применить метод относительных весов, то количество осадков и плотность трафика имеет наибольшую относительную важность, а дальше следуют (в порядке убывания важности) температура, возраст водителя, опыт водителя и время суток.

Показатели относительной важности переменных (и в особенности метод относительных весов) имеют широкое применение. Они находятся гораздо ближе к интуитивной концепции относительной важности по сравнению со стандартизованными коэффициентами регрессии. Я уверен, что в ближайшие годы интенсивность использования показателей относительной важности резко возрастет.

ЗАКЛЮЧЕНИЕ

Современное понимание истинных причин ДТП приводит к выводу, что снижение вероятности совершения ошибок участниками дорожного движения – весьма перспективное направление для деятельности организаций, отвечающих за безопасность дорожного движения, а также для международного сотрудничества.

Из исследования следует, что около 3/4 всех учетных ДТП, зарегистрированных в Управлении ГИБДД ГУ МВД России по Новосибирской области, происходит по вине человека. Неблагоприятное сочетание факторов “человек + дорожные условия” (то есть человеческая ошибка + неблагоприятные сопутствующие дорожные условия) является причиной еще четверти ДТП с погибшими. Однако, по имеющимся статистическим данным из учета ДТП оказалось невозможным определить чистое влияние таких факторов, как «водитель» и «дорожные условия» а также влияние их сочетания.

Из проведенного исследования можно сделать следующие выводы:

Средний стаж водителя-виновника ДТП составляет 14.39 лет. Минимальный стаж 0 лет, максимальный 44 года. 50% всех совершенных ДТП совершают водители со стажем от 6 до 21 года, то есть, выводы о том, что водители со стажем менее 3-х лет является наиболее опасным участником дорожного движения, является ошибочным. Водители даже с большим опытом могут совершить ДТП, которое может повлечь за собой летальные исходы. Здесь можно сделать предположение о том, что такие водители бывают чересчур самоуверенными, ссылаясь на свой стаж, что и приводит к ошибкам. Предположение о наиболее высоком риске возникновения ДТП с летальным исходом у водителей с высоким стажем оказалось верным.

Кроме того, оказались верными предположения о том, что возникновение ДТП с летальным исходом при погодных условиях наиболее высоко с увеличением осадков и/или при низкой температуре. Наиболее же важными переменными оказались переменная «Плотность трафика» и «Количество осадков», то есть кроме количества выпавших осадков, также плотность трафика значительно влияет на количество ДТП, что в свою очередь увеличивает вероятность летального исхода. Другими словами, перегруженность дороги транспортными средствами повышает число ошибок участников дорожного движения, конфликтных ситуаций, что приводит к росту числа ДТП. А что касается количества осадков как важного фактора, влияющего на увеличение ДТП, то можно сделать следующий вывод: во время осадков число ДТП увеличивается; если осадки затяжные, то водители адаптируются и число ДТП постепенно снижается.

Другие переменные (факторы) как возраст водителя, время суток также влияют на количество летальных исходов в ДТП и самих ДТП, в целом. О времени суток из данного исследования можно сделать следующие выводы: в темное время суток относительное число ДТП примерно в 1.5-3.5 раза выше по сравнению со светлым временем - условия видимости хуже, может быть больше водителей в состоянии алкогольного опьянения, утомленных водителей.

Рекомендации (мероприятия) по уменьшению/предотвращению количества ДТП:

- Улучшение качества образования и обучения, во-первых - детей и школьных учителей безопасности дорожного движения; во-вторых, обязательно нужно учить подростков принципам безопасного вождения и серьезному отношению к безопасности дорожного движения; в-третьих, нужны курсы повышения квалификации для пожилых

водителей, чтобы обновить знания правил дорожного движения; и конечно же нужна пропаганда через газеты, радио и телевидение, чтобы привлечь внимание всех участников дорожного движения и к опасностям на дороге.

- Обеспечение выполнения и принятие разумных и правил дорожного движения, которые сами по себе прежде всего предназначены для предотвращения несчастных случаев; улучшение материально технической базы дорожной полиции; и тщательного тестирования новых транспортных средств, чтобы они не становились причиной несчастных случаев.

- Плановая проверка транспортных средств, включающая регулярный осмотр автомобиля, для того чтоб быть уверенным, что основные компоненты транспортного средства работоспособны и безопасны; совершенствование конструкции автомобиля производителем, улучшение систем пассивной безопасности.

- Постройка новых дорог, которые являются безопасными в результате разделения противоположных транспортных потоков, уменьшение перекрестных трафиков, а также проектирование и постройка широких полос движения с хорошей видимостью; совершенствование существующих дорог.

В любом случае, причинами ДТП в большинстве случаев являются ошибки человека. Исследования показывают, что на счету ошибок водителей более 80% всех смертельных и травмоопасных случаев.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. R в действии. Анализ и визуализация данных на яз. R Автор: Кабаков Роберт И., Издательство: ДМК-Пресс, 2014г.
2. Девятов В.М. Обоснование мероприятий по снижению аварийности дорожного движения на участках концентрации ДТП// Вестник ВолгГАСУ. Сер.: Стр-во и архит, 2009г.
3. Луценко Е.В., Коржаков В.Е. Адаптивная семантическая информационная модель прогнозирования рисков совершения ДТП // Вестник Адыгейского Государственного Университета. Серия 4: Естественно-математические и технические науки, 2008г.
4. Энглези И.П., Пахно А.Е. Моделирование вероятности возникновения ДТП на участке транспортной сети // Вестник Донецкой академии автомобильного транспорта № 4, 2010г.
5. Каминский А. 100 способов избежать аварии. Спецкурс для водителей категории «В». Общероссийский проект «Безопасность дорожного движения», 2010г.
6. Голиусов Ю. Приоритеты Российской транспортной политики //Власть, 2009г.
7. Думанян Г.Д., Давидянц В.А. Дорожно-транспортные происшествия: последствия для общественного здравоохранения. Ереван, 2008г.
8. Попов П. Л. Дорожно-транспортные происшествия в Российской Федерации в 1997-2007 гг. Географический аспект // В мире научных открытий, 2010г.
9. Состояние безопасности дорожного движения. Партнерский обзор по стране: Российская Федерация – ЕКМТ, 2006г.