# ВВЕДЕНИЕ

В современном мире авиационная отрасль ежегодно претерпевает значительные изменения, что приводит к постоянному росту конкуренции и разнообразию предлагаемых услуг. Одним из ключевых факторов, определяющих успех авиакомпаний, является ценообразование на авиабилеты. Традиционные методы анализа цен часто сосредоточены на оценке среднего значения, что может не учитывать сложные зависимости и вариации в данных.

В этом контексте квантильная регрессия представляет собой мощный инструмент, позволяющий исследовать влияние различных факторов на стоимость авиаперелетов не только на уровне среднего, но и на различных квантилях распределения. Квантильная регрессия позволяет анализировать, как различные переменные, такие как время покупки билета, сезонность, расстояние между пунктами назначения и другие факторы, влияют на разные уровни цен. Это особенно важно для понимания динамики рынка авиаперевозок, так как влияние факторов может значительно варьироваться в зависимости от уровня цен. Например, некоторые факторы могут оказывать более выраженное влияние на высокие цены, чем на низкие, что может быть упущено при использовании традиционных методов регрессии.

Задачи, решаемые в данной курсовой работе:

* {old:[Выбор данных и описание набора данных] new:[выбор данных и описание набора данных]};
* {old:[Построение сценария и логики обработки данных] new:[построение сценария и логики обработки данных]};
* {old:[Построение модели квантильной регрессии на основе известных данных] new:[построение модели квантильной регрессии на основе известных данных]};
* {old:[Анализ получившейся квантильной регрессии] new:[анализ получившейся квантильной регрессии]};
* {old:[Оценка эффективности разработанной программы построения квантильной регрессии.] new:[оценка эффективности разработанной программы построения квантильной регрессии.]}

# 1 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

## **1.1 Использование регрессии для статистической обработки данных**

Перед тем, как перейти к выполнению основной задачи курсовой работы, следует дать несколько определений регрессии в математике и статистике, исследовать различные формы регрессии и различия между ними. После чего следует более подробно изучить вид регрессии, указанный в теме курсовой работы, а также подробно рассмотреть механизмы его создания для наиболее полного и корректного выполнения практической части настоящей работы.

Полученные знания будут использованы для выполнения основной задачи курсовой работы и написания практического этапа, заключающегося в разработке модели квантильной регрессии на примере данных стоимости пассажирских авиаперевозок.

Регрессия — это статистический метод, который используется для анализа взаимосвязи между одной зависимой переменной и одной или несколькими независимыми переменными. Данное понятие впервые было введено в конце 19 века английским статистиком Фрэнсисом Гальтоном. В 1885 году он опубликовал работу, в которой исследовал взаимосвязь между ростом родителей и их детей. Гальтон заметил, что дети, как правило, имеют рост, который ближе к среднему значению, чем рост их родителей. Он описал это явление как «регрессию к среднему» (regression towards mediocrity).

Термин «регрессия» был использован Гальтоном для обозначения того, что экстремальные значения имеют тенденцию возвращаться к среднему уровню в следующем поколении. Это наблюдение стало основой для дальнейшего развития статистических методов, связанных с анализом зависимостей между переменными. Позже, в начале 20 века, статистические методы, основанные на регрессии, были значительно усовершенствованы и расширены благодаря работам таких ученых, как Карл Пирсон и Рональд Фишер. Пирсон разработал метод линейной регрессии и коэффициент корреляции, а Фишер ввел концепцию множественной регрессии и другие статистические методы.

Математически регрессию можно определить через модель, которая описывает зависимость зависимой переменной Y от нескольких независимых переменных. В простейшем случае линейной регрессии эта зависимость может быть выражена следующим образом:

{old:[[] new:[​]}{old:[eq] new:[​]}] (1.1)

где — свободный член (или константа),

, , …, {old:[ — коэффициенты регрессии, которые показывают, как изменение независимых переменных влияет на зависимую переменную, ] new:[ — коэффициенты регрессии, которые показывают, как изменение независимых переменных влияет на зависимую переменную,]}

ϵ — ошибка модели, отражающая влияние факторов, не включенных в модель.

Данное описание, чисто математическое, тем не менее точно отражает суть регрессии. Однако помимо теоретического, существует ее прикладное применение. Регрессия используется в статистике и обработке данных для различных целей. Она позволяет предсказывать значения зависимой переменной на основе значений независимых переменных. Использование аналитических моделей с применением регрессии можно заметить в экономике для прогнозирования продаж, в медицине для оценки влияния факторов на здоровье или в социальных науках для изучения взаимосвязей между различными показателями. Кроме того, регрессионный анализ помогает выявить и количественно оценить степень влияния независимых переменных на зависимую. Это может быть особенно важно при принятии решений на основе проанализированных данных.

## **1.2 Квантильная регрессия**

Рассмотрев некоторые основные понятия и определения, а также изучив историю появление терминологии регрессии, следует расширить исследуемый материал, представив различные виды регрессии и рассмотрев их достоинства и недостатки. В итоге данная часть теоретической части гарантирует комплексное понимание задачи практического этапа, где необходимо построить модель конкретного вида регрессии на основе известных данных — а именно квантильной регрессии.

В статистике существует несколько видов регрессии, каждый из которых подходит для различных типов данных и задач. Кратко рассмотрим основные из них для понимания основных различий и выявления достоинств и недостатков каждого вида:

1. Линейная регрессия.

Линейная регрессия предполагает линейную зависимость между независимой переменной X и зависимой переменной Y. Линейная регрессия, являясь самой простой моделью из представленных, имеет ряд достоинств, а именно:

* {old:[Простота интерпретации] new:[простота интерпретации]};
* {old:[Легкость в вычислении] new:[легкость в вычислении]};
* {old:[Высокая эффективность работы в линейных зависимостях.] new:[высокая эффективность работы в линейных зависимостях;]}

Однако ввиду простоты реализации линейная регрессия является менее точной по сравнению с более сложными моделями и имеет определенные недостатки:

* {old:[Чувствительность к выбросам] new:[чувствительность к выбросам]};
* {old:[Непригодн] new:[непригодн]}ость {old:[для нелинейных зависимостей. ] new:[для нелинейных зависимостей;]}

Сфера применения линейной регрессии включает в себя практически все отрасли, где применяется статистика, такие как экономика, социология, биология и прочие, так как ввиду своей эффективности при работе с линейными зависимостями, является крайне популярным решением.

1. Полиномиальная регрессия.

Полиномиальная регрессия, в отличие от линейной, используется для моделирования нелинейных зависимостей между переменными. Данный вид регрессии является более комплексным, чем линейная, однако по-прежнему весьма прост в реализации. Модель может выглядеть так:

(1.2)

где — свободный член (или константа),

, , …, {old:[ — коэффициенты регрессии, которые показывают, как изменение независимых переменных влияет на зависимую переменную, ] new:[ — коэффициенты регрессии, которые показывают, как изменение независимых переменных влияет на зависимую переменную,]}

ϵ — ошибка модели, отражающая влияние факторов, не включенных в модель.

Полиномиальная регрессия имеет большое количество достоинств, особенно по сравнению с линейной регрессией, а именно:

* {old:[Способн] new:[способн]}ость моделировать сложные нелинейные зависимости;
* {old:[Гибкость в выборе степени полинома] new:[гибкость в выборе степени полинома]};
* {old:[Относительная простота реализации] new:[относительная простота реализации]};
* {old:[Возможность анализа трендов] new:[возможность анализа трендов]}{old:[.] new:[;]}

Тем не менее, несмотря на очевидные преимущества, полиномиальная регрессия также обладает рядом недостатков, в частности:

* {old:[Риск переобучения (] new:[риск переобучения (]}overfitting) при высоких степенях полинома;
* {old:[Сложность интерпретации для высоких степеней. ] new:[сложность интерпретации для высоких степеней;]}

Полиномиальная регрессия, являясь своеобразным переходом от простой линейной регрессии к более сложным моделям, имеет более узкое применение, чем предыдущая, однако может быть использована в научных работах и диссертациях, где исследуемые данные имеют явные нелинейные паттерны и не поддаются обработке линейной регрессией.

1. Логистическая регрессия.

Логистическая регрессия используется для моделирования бинарных исходов. Она оценивает вероятность события по следующей формуле:

(1.3)

где — вероятность того, что событие произойдет,

— свободный член (или константа),

, , …, — коэффициенты регрессии, которые показывают, как изменение независимых переменных влияет на зависимую переменную.

Логистическая регрессия, обладающая уникальной особенностью работать с вероятностями, несомненно, имеет большое количество достоинств, а именно:

* {old:[Способность] new:[способность]} предсказывать вероятности;
* {old:[Высокая эффективность] new:[высокая эффективность]} работы{old:[ с бинарными исходами. ] new:[ с бинарными исходами;]}

Тем не менее, данный вид регрессии имеет и свои недостатки, обусловленные особенностями функционирования конкретной модели, среди которых:

* {old:[Не подходит для ] new:[не подходит для ]}многоклассовых задач без модификаций;
* {old:[Предполагает линейную связь между ] new:[предполагает линейную связь между ]}логитом{old:[ вероятности и независимыми переменными. ] new:[ вероятности и независимыми переменными;]}

Сфера применения логистической регрессии включает в себя чаще всего проведение медицинских исследований с необходимостью работы с вероятностями, а также отрасль маркетинговых исследований.

1. Регрессия с использованием деревьев решений (Decision Tree Regression).

Данный метод регрессии использует дерево решений для предсказания значений. Дерево разбивает данные на подгруппы по значениям независимых переменных.

Регрессия с использованием дерева решений имеет определенные достоинства, такие как:

* {old:[Хорошая эффективность в работе] new:[хорошая эффективность в работе]} с нелинейными зависимостями;
* {old:[Простота] new:[простота]} интерпретации и визуализации{old:[.] new:[;]}

Тем не менее, недостатки также присутствуют:

* {old:[Шанс переобучения ] new:[шанс переобучения ]}на малых выборках;
* {old:[Чувствительно] new:[чувствительно]}сть{old:[ к изменениям в данных. ] new:[ к изменениям в данных;]}

Чаще всего Decision Tree Regression применяется в финансовом анализе и маркетинге, где применяется для исследования рынка и обработки статистики

1. Регрессия с регуляризацией (Ridge и Lasso).

Rigle и Lasso — методы, используемые при построении регрессии с регуляцией. Основная их задача — добавление штрафа к коэффициентам регрессии, и достигаемое таким образом отсутствие переобучения модели.

Ridge использует L2-норму:

(1.4)

где a — гиперпараметр, корректирующий величину штрафа

,— коэффициенты регрессии, которые показывают, как изменение независимых переменных влияет на зависимую переменную.

Lasso использует L1-норму:

, (1.5)

где a — гиперпараметр, корректирующий величину штрафа

,— коэффициенты регрессии, которые показывают, как изменение независимых переменных влияет на зависимую переменную.

Достоинства применения данных методов для моделирования регрессии неоспоримы:

* {old:[Уменьш] new:[уменьш]}ение переобучения;
* {old:[Возможность ] new:[возможность ]}Lasso {old:[отбирать переменные (обнулять коэффициенты). ] new:[отбирать переменные (обнулять коэффициенты);]}

Тем не менее, у регрессии с регуляризацией существуют и недостатки, а именно:

* {old:[Осложнение выбора ] new:[осложнение выбора ]}гиперпараметра регуляризации;
* {old:[Затрудненная интерпретация модели] new:[затрудненная интерпретация модели]}{old:[. ] new:[;]}

1. Квантильная регрессия.

Рассмотрим еще один вид регрессии, который будет нами использован при выполнении практического этапа курсовой работы — а именно квантильную регрессию. Перед изучением основных понятий и механизмов действия данного вида регрессии, необходимо изучить его историю.

Квантильная регрессия появилась в 1970-х годах как ответ на необходимость более гибкого подхода к анализу данных, чем традиционная линейная регрессия. Основным инициатором этого направления стал статистик Рoger Koenker, который вместе с Гоффридом Баском разработал методы, позволяющие оценивать условные квантильные функции. Традиционная линейная регрессия фокусируется на среднем значении зависимой переменной при заданных значениях независимых переменных.

Однако в некоторых случаях исследователи заинтересованы не только в среднем, но и в других аспектах распределения, таких как медиана или другие квантильные значения. Квантильная регрессия позволяет это сделать, предоставляя возможность оценивать влияние независимых переменных на различные части распределения зависимой переменной. С течением времени квантильная регрессия приобрела популярность в различных областях, включая экономику, экологии и медицину. Исследователи начали применять этот метод для анализа данных, где традиционные подходы не давали полной картины. Квантильная регрессия также оказалась полезной для изучения асимметричных эффектов и выявления устойчивых трендов в данных.

Таким образом, квантильная регрессия, возникшая в ответ на потребность более глубокого анализа распределений данных, предоставляет исследователям мощный инструмент для изучения влияния независимых переменных на различные квантильные значения зависимой переменной. Понимание исторических аспектов и мотивации, стоящих за развитием этого метода, подготавливает нас к более детальному рассмотрению его математической основы. В следующем подразделе мы погрузимся в формальные определения и ключевые концепции квантильной регрессии, которые позволят нам лучше понять, как этот подход реализует свои преимущества на практике.

Квантильная регрессия — это статистический метод, который позволяет исследовать связь между переменной-ответом и набором независимых переменных, фокусируясь не только на среднем значении, как это делает традиционная линейная регрессия, но и на различных квантилях распределения зависимой переменной. Это означает, что квантильная регрессия может предоставить более детализированное представление о том, как факторы влияют на разные части распределения, например, на его нижнюю, среднюю и верхнюю части.

Квантили представляют собой точки, которые делят распределение на равные части. Например, медиана делит данные пополам, в то время как 25-й процентиль показывает значение, ниже которого находится 25% наблюдений. Квантильная регрессия позволяет исследовать влияние независимых переменных на эти и другие квантили, что делает её особенно полезной в ситуациях, когда интересуют не только средние значения, но и крайние случаи.

Достоинства квантильной регрессии включают в себя прежде всего:

1. Гибкость.

Позволяет моделировать различные аспекты зависимой переменной (например, медиану, верхние и нижние квантилы), что полезно для понимания распределения данных.

1. Устойчивость к выбросам.

Квантильная регрессия менее чувствительна к выбросам по сравнению с обычной линейной регрессией, так как фокусируется на определенных квантилях.

1. Информация о распределении.

Предоставляет более полное представление о зависимости между переменными, особенно когда распределение зависимой переменной асимметрично или имеет тяжелые хвосты.

1. Подходит для неравномерных данных.

Квантильная регрессия эффективна при наличии изменчивой дисперсии данных, что повышает надежность данного метода построения регрессии.

Проанализировав представленные достоинства, можно с уверенностью заявить, что квантильная регрессия является одним из лучших методов анализа статистических данных в рамках данной темы. Однако с учетом объективности необходимо также рассмотреть недостатки:

1. Сложность интерпретации.

Интерпретация результатов может быть сложнее, чем в случае обычной линейной регрессии, особенно если необходимо учитывать несколько квантилей одновременно.

1. Высокие требования к данным.

Для корректного применения метода необходимо достаточное количество данных для каждого квантиля, что может быть проблемой при малом размере выборки.

1. Вычислительная сложность.

Решение задач квантильной регрессии может требовать больше вычислительных ресурсов, чем использование обычной линейной регрессии.

Представленные недостатки, несомненно, существенны, однако не так важны в рамках решаемой задачи, поэтому квантильная регрессия является идеальным выбором.

Таким образом, квантильная регрессия представляет собой мощный инструмент для глубокого анализа данных, позволяя исследователям лучше понимать сложные зависимости и вариации в данных.

# 2 ПРАКТИЧЕСКАЯ ЧАСТЬ

## **2.1 Предобработка данных для моделирования квантильной регрессии**

Для моделирования квантильной регрессии использовались данные о пассажирских авиаперевозках, предоставленные в датасете “Flight Price Prediction 2018-2022” с сайта “Kaggle”, который содержит данные о стоимости пассажирских авиарейсов крупных авиакомпаний за период с 2018 по 2022 годы. Содержание данного датасета представлено на рисунке 2.1

Изображение выглядит как текст, снимок экрана, линия, число

Автоматически созданное описание

**Рисунок 2.1 — Часть датасета**

На данном рисунке изображен фрагмент датасета, представленного в виде таблицы. На рисунке отображены первые 10 строк таблицы.

Датасет содержит следующие ключевые поля:

1. “airline{old:[” — ] new:[" — ]}авиакомпания, осуществляющая перелет по маршруту.
2. {old:[“flight” — ] new:["flight" — ]}номер рейса.
3. “source\_city{old:[” — ] new:[" — ]}аэропорт вылета.
4. “departure\_time{old:[” —] new:[" —]} время вылета.
5. “stops” — количество плановых промежуточных посадок в полете.
6. “arrival\_time” — время прибытия.
7. “destination\_city” — пункт назначения.
8. “class” — класс обслуживания данного билета.
9. {old:[“duration”] new:["duration"]} — длительность перелета.
10. “days\_left” — разность (в днях) между датой бронирования и датой перелета.
11. {old:[“price”] new:["price"]} — стоимость билета.

Ознакомившись с данным датасетом, можно сказать, что он представляет из себя крайне ценный с точки зрения информации объект, где указаны различные аспекты исследуемых данных. Однако перед тем, как построить на основе этой информации регрессию и выполнить практический этап, необходимо провести анализ качества данных для исключения возможных ошибок, могущих повлиять на конечный результат. Анализ качества данных включал в себя следующие этапы:

1. Проверка полноты данных.

Проверка наличия пропущенных значений в ключевых полях. В данном датасете пропущенных значений обнаружено не было.

1. Проверка на дубликаты.

Выявление и удаление дублирующих записей для предотвращения искажений в анализе данных. В данном датасете дубликатов обнаружено не было.

1. Анализ целостности данных.

Проверка на наличие несоответствий в данных и присутствие логических ошибок. При проверке данного датасета была подтверждена целостность данных.

После подтверждения пригодности данных необходимо подготовить информацию к использованию в построении квантильной регрессии. Для этого проведем предобработку. Предобработка данных включала в себя несколько шагов, а именно:

1. Выборка данных для построения квантильной регрессии.

{old:[В рамках предобработки данных было установлено, что для выполнения практического этапа необходимо выбрать два столбца данных для отображения зависимости между ними в виде квантильной регрессии. Регрессия будет поострена в двумерной плоскости, где переменная X — “] new:[В рамках предобработки данных было установлено, что для выполнения практического этапа необходимо выбрать два столбца данных для отображения зависимости между ними в виде квантильной регрессии. Регрессия будет поострена в двумерной плоскости, где переменная X — "]}duration{old:[” — ] new:[" — ]}{old:[длительность полета, а переменная Y — “] new:[длительность полета, а переменная Y — "]}price{old:[” — стоимость перелета. Таким образом, в построенной регрессии будет отображена зависимость стоимости билета от длительности перелета.] new:[" — стоимость перелета. Таким образом, в построенной регрессии будет отображена зависимость стоимости билета от длительности перелета.]}

1. Преобразование поля “price” в числовой формат.

{old:[В рамках предобработки данных необходимо привести все числа к одному формату для последующего использования в построении квантильной регрессии на известных данных. Для этого поле “] new:[В рамках предобработки данных необходимо привести все числа к одному формату для последующего использования в построении квантильной регрессии на известных данных. Для этого поле "]}price{old:[”, означающее стоимость билета, будет преобразовано в числовой формат.] new:[", означающее стоимость билета, будет преобразовано в числовой формат.]}

1. Преобразование поля “duration” в числовой формат.

{old:[В рамках предобработки данных необходимо привести все числа к одному формату для последующего использования в построении квантильной регрессии на известных данных. Для этого поле “ ] new:[В рамках предобработки данных необходимо привести все числа к одному формату для последующего использования в построении квантильной регрессии на известных данных. Для этого поле " ]}duration{old:[”, означающее длительность перелета, будет преобразовано в числовой формат.] new:[", означающее длительность перелета, будет преобразовано в числовой формат.]}

1. Обнаружение и устранение выбросов.

Выявление аномальных значений или выбросов, которые могут существенно искажать итоговый результат регресии. Использовались статистические методы и визуализация для обнаружения таких значений, после чего они были обработаны соответствующим образом (например, заменены на средние значения или удалены).

Для подготовки данных к моделированию были выполнены следующие шаги:

1. Проверка соответствия форматов данных.

В качестве финального этапа обработки данных необходимо удостовериться в корректном заполнении датасета данными единого формата, что необходимо для корректной работы программы. Данная проверка будет осуществлена для подтверждения соответствия форм данных. При проверке данного датасета было подтверждено соответствие форм данных, используемых для построения квантильной регрессии в настоящей работе, что позволяет перейти к следующему этапу, заключающемуся в формировании обучающей и тестовой сборок.

1. Формирование обучающей и тестовой выборок.

Разделение данных на обучающую и тестовую выборки для оценки качества модели путем выделения данные за последний год для тестирования модели.

После всех этапов предобработки данные стали готовы для использования их в построении квантильной регрессии с помощью языка программирования R.

## **2.2 Разработка модели квантильной регрессии**

В данном разделе представлена разработка программы на языке программирования R для разработки модели квантильной регрессии с использованием статистических данных о стоимости пассажирских авиаперевозок.

На основе полученных теоретических данных напишем программу на языке программирования R. Рассмотрим функционал и механизм действия созданной программы

Для выполнения задачи построения квантильной регрессии с помощью язык программирования R используются следующие библиотеки:

1. {old:[library] new:[Library]}(quantreg) — библиотека для реализации квантильной регрессии.
2. {old:[library] new:[Library]}(dplyr) — библиотека для обработки таблицы данных.
3. {old:[library] new:[Library]}(ggplot2) — библиотека для визуализации данных и построения графиков.

После загрузки библиотек необходимо подгрузить данные в программу, для этого воспользуемся специализированной функцией read.csv.

Проведя успешную загрузку данных, необходимо приступить к предобработке. Первое действие в рамках данного блока предусматривает преобразование переменной price в числовой формат с помощью функции as.numeric. Второе действие – преобразование данных о продолжительности полета в числовой формат. Данная операция производится при помощи поэтапного перевода часов и минут в числовой формат с помощью функции as.numeric с применением необходимых коэффициентов для корректного перевода.

После завершения предобработки необходимо перейти к построению регрессии. Первых шаг — преобразование категориальных переменных — таких, как stops, class и source\_city в факторы для учета их в итоговой регрессии.

Завершив предыдущий этап, необходимо приступить к выполнению квантильной регрессии. В модели будут использованы три квантили – 25-я, 50-я (медианная) и 75-я.

Финальный этап выполнения программы – получение предсказанных значений и вывод квантильной регрессии на экран.

В результате работы созданной программы получим итоговый результат квантильной регрессии в виде графика, представленного на рисунке 2.2

Изображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

**Рисунок 2.2 — Квантильная регрессия**

Данный график демонстрирует, как меняется цена на авиабилет в зависимости от длительности полета для разных квантилей распределения цен. Изучив график, можно прийти к выводу о том, что все модели показывают прямую зависимость стоимости перелета от длительности полета, что соответствует действительности. Исходя из этого, можно сказать, что поставленная задача была успешно выполнена.

Проанализировав полученную регрессию, выделим ее основные параметры. Квантильная регрессия (Рисунок 2.2) включает следующие компоненты:

1. Свободный член (Intercept).

Свободный член в регрессионной модели — это константа, которая представляет значение зависимой переменной, когда все независимые переменные равны нулю.

1. Коэффициент при duration.

Коэффициент при переменной "duration" в регрессионной модели указывает на то, как изменение этой переменной влияет на зависимую переменную. Если коэффициент при "duration" положителен, это означает, что с увеличением значения "duration", зависимая переменная также увеличивается. Если коэффициент отрицателен, это означает, что увеличение "duration" связано с уменьшением зависимой переменной.

1. Псевдо R-квадрат.

Псевдо R-квадрат в контексте квантильной регрессии — это мера, аналогичная традиционному R-квадрату в линейной регрессии, но адаптированная для случаев, когда мы не предполагаем нормальное распределение ошибок и не используем средние значения для оценки. Псевдо R-квадрат может варьироваться от 0 до 1, где значения ближе к 1 указывают на лучшее соответствие модели данным.

Оценке подвергнем 50-й квантиль(медианный):

* {old:[Свободный ] new:[свободный ]}член (Intercept) равен 3830.43;
* {old:[Коэффициент при ] new:[коэффициент при ]}duration равен 349.79;

Исходя из этого, можем сделать вывод о том, что связь зависимой переменной с независимой является положительной.

* {old:[Псевдо ] new:[псевдо ]}R-{old:[квадрат равен 0.01722. ] new:[квадрат равен 0.01722;]}

Данное низкое значение указывает на слабую объясняющую способность модели для этой конкретной зависимости.

В ходе нашего исследования зависимости стоимости перелета от дальности мы выявили ряд важных результатов, которые подчеркивают сложность ценообразования в авиационной отрасли. Анализ показал, что существует положительная связь между расстоянием и ценой билета: чем длиннее маршрут, тем выше стоимость перелета. Это может быть обусловлено увеличением затрат на топливо, обслуживание и другие операционные расходы. Тем не менее, низкая объяснительная сила модели и малое значение псевдо R-квадрата указывает на то, что дальность полета является лишь одним из множества факторов, влияющих на ценообразование. Полученные данные могут быть полезны для авиакомпаний при разработке более эффективных стратегий ценообразования.

Подводя итоги, можно сказать, что в ходе выполнения данной курсовой работы была проведена всесторонняя исследовательская работа по изучению и применению на практике квантильной регрессии на примере данных о стоимости авиаперелетов. Квантильная регрессия, как метод, позволила более глубоко понять влияние учтенных факторов на ценообразование в авиационной отрасли, учитывая не только средние значения, но и распределение цен. Проведённый анализ показал, что стоимость авиабилетов действительно зависит от длительности полета, однако проанализировав псевдо R-квадрат можно сделать вывод о том, что данная связь не является ключевой.

Использование квантильной регрессии дало возможность выявить, как связь данных показателей влияют на разные уровни цен. Также полученные выводы могут служить основой для дальнейших исследований в области статистического анализа с применением языка R. Таким образом, работа не только углубила понимание сути функционирования квантильной регрессии и синтаксиса языка R как инструмента анализа данных, но и продемонстрировала ее практическую применимость в реальной экономической сфере.

## **ЗАКЛЮЧЕНИЕ**

Анализ и прогнозирование данных — важнейшая задача аналитика, которая в зависимости от решаемой задачи может оказать сильнейшее влияние на целую отрасль. Правильно проанализированные и спрогнозированные данные могут быть применены во всех сферах жизни общества для различных решений, способствующих развитию государства.

В ходе выполнения данной курсовой работы был проведен анализ известных данных при помощи квантильной регрессии, реализованной инструментами языка R на примере данных о стоимости авиабилетов за период 2018-2022 годов.

Задачи, решенные в данной курсовой работе:

* {old:[Выбор данных и описание данных] new:[выбор данных и описание данных]};
* {old:[Построение сценария и логики обработки данных] new:[построение сценария и логики обработки данных]};
* {old:[Построение модели квантильной регрессии на основе известных данных] new:[построение модели квантильной регрессии на основе известных данных]};
* {old:[Анализ получившейся квантильной регрессии] new:[анализ получившейся квантильной регрессии]};
* {old:[Оценка эффективности разработанной программы построения квантильной регрессии.] new:[оценка эффективности разработанной программы построения квантильной регрессии;]}

Таким образом, данная курсовая работа подтверждает высокую эффективность применения методов квантильной регрессии для глубокого анализа статистических данных. Эти методы позволяют более точно исследовать зависимости и выявлять закономерности, что делает их особенно ценными в различных сферах. В результате проведенного исследования мы получаем важные данные, которые могут быть полезны не только для дальнейших исследований в области авиаперевозок, но и для практического применения в других областях, где важен детальный анализ данных и понимание факторов, влияющих на результаты.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

## ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

1. Носков Сергей Иванович МНОГОКРИТЕРИАЛЬНЫЙ МЕТОД АНТИРОБАСТНОГО ОЦЕНИВАНИЯ ПРИ ПОСТРОЕНИИ ЛИНЕЙНОЙ РЕГРЕССИИ // Известия ТулГУ. Технические науки. 2023. №8.
2. Петров Александр Васильевич Инструменты экспериментальных исследований полиномиальной регрессии // Вестник ИрГТУ. 2013. №10 (81).
3. {old:[Angrist, Joshua D.; Pischke, Jörn-Steffen (2009). ] new:[Angrist, Joshua D.; Pischke, Jörn-Steffen (2009). ]}{old:["Quantile Regression". Mostly Harmless Econometrics: An Empiricist's Companion. ] new:["Quantile Regression". Mostly Harmless Econometrics: An Empiricist's Companion. ]}Princeton University Press.
4. Фёрстер Э., Рёнц{old:[ Б. Методы корреляционного и регрессионного анализа. Руководство для экономистов. — М.: Финансы и статистика, 1983. — 304 с. — (Библиотечка иностранных книг для экономистов и статистиков).] new:[ Б. Методы корреляционного и регрессионного анализа. Руководство для экономистов. — М.: Финансы и статистика, 1983. — 304 с. — (Библиотечка иностранных книг для экономистов и статистиков).]}

## ПРАКТИЧЕСКАЯ ЧАСТЬ

1. Flight Price Prediction 2018-2022 [Электронный ресурс] <https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction> (Дата обращения 01.09.2024).
2. Язык программирования R/Документация [Электронный ресурс] <https://www.r-project.org/other-docs.html> (Дата обращения 01.09.2024).
3. Koenker{old:[, Roger (2005). Quantile Regression. Cambridge University Press. ] new:[, Roger (2005). Quantile Regression. Cambridge University Press.]}

# ПРИЛОЖЕНИЯ

Приложение А — Код программы на языке программирования R.

## **Приложение А**

*Листинг А.1*

library(quantreg)

library(dplyr)

library(ggplot2)

data <- read.csv("C:\\Users\\<путь к файлу>\\Clean\_Dataset.csv")

data$price <- as.numeric(gsub(",", "", data$price))

convert\_duration <- function(duration\_str) {

{old:[ h <- ] new:[ h <- ]}as.numeric(sub("h.\*", "", duration\_str{old:[)) # ] new:[)) # ]}Часы

{old:[ m <- ] new:[ m <- ]}as.numeric(sub(".\*h |m", "", sub(".\*h ", "", duration\_str{old:[))) # ] new:[))) # ]}Минуты

{old:[ h[is.na(h)] <- 0 ] new:[ h[is.na(h)] <- 0]}

{old:[ m[is.na(m)] <- 0 ] new:[ m[is.na(m)] <- 0]}

return(h + m / 60)

}

data$duration <- convert\_duration(data$duration)

data$stops <- as.factor(data$stops)

data$class <- as.factor(data$class)

data$source\_city <- as.factor(data$source\_city)

data$destination\_city <- as.factor(data$destination\_city)

data$airline <- as.factor(data$airline)

model\_50 <- rq(price ~ duration, tau = 0.5, data = data)

model\_25 <- rq(price ~ duration, tau = 0.25, data = data)

model\_75 <- rq(price ~ duration, tau = 0.75, data = data)

data$pred\_50 <- predict(model\_50)

data$pred\_25 <- predict(model\_25)

data$pred\_75 <- predict(model\_75)

ggplot(data, aes(x = duration, y = price)) +

geom\_point(alpha = 0.3) +

geom\_line(aes(y = pred\_50), color = "blue") +

geom\_line(aes(y = pred\_25), color = "red") +

geom\_line(aes(y = pred\_75), color = "green") +

labs(title = "Квантильная регрессия: цена от длительности полета",

{old:[ x = "Длительность полета (в часах)",] new:[ x = "Длительность полета (в часах)",]}

{old:[ ] new:[ ]}y = "Цена (в рублях)") +

theme\_minimal()