сФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ

ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

Национальный исследовательский университет

«Высшая школа экономики»

Факультет экономических наук

Проектная работа по дисциплине "Эконометрика 2"

Тема: «Оценивание влияния различных факторов на цену одного грамма чая с использованием квантильной регрессии»

ФИО	% Вклада	Выполняемые задачи:
Гаврин Глеб Юрьевич	33.333...%	Анализ признаков, отбор переменных AIC, гетероскедастичность (Бройш-Паган, график), отбор выбросов (DFFITS, DFBETA), нормальность остатков, бутстрап.
Каленбет Ульяна Денисовна	33.333...%	Парсинг, создание переменных, функциональная форма (тест Бокса-Кокса), тест Чоу, предсказание на новом объекте, квантильная регрессия, тест на адекватность.
Шалунова Анна Игоревна	33.333...%	Создание переменных, гетероскедастичность (график, Уайт, Глейзер), мультколлинеарность (Матрица корреляций, VIF).

Руководители:

Вакуленко Елена Сергеевна

Погорелова Полина Вячеславовна

Москва, 2025

1. *Введение*

Данное исследование ставит перед собой задачу описать модель ценообразования одного грамма чая в зависимости от ряда различных факторов, а также рассмотреть имеют ли они различные эффекты влияния на разные ценовые категории чая. Чтобы достичь этой цели, в данной работе использовались линейная и полулогарифмическая регрессионные модели, а также квантильная регрессия. Расчеты производились с помощью языка программирования python с использованием как и собственно написанных функций, так и функций из ряда библиотек.

2. *Работа с данными*

2.1 Исходные данные

В данной работе использовались данные, предоставленные в открытый доступ интернет-магазином https://teaco.ru/. Сбор данных проводился с помощью собственноручно написанного парсера на базе библиотеки Selenium. Данные были взяты из каталога /catalog/chay/, а также из карточек товаров и были преобразованы в таблицу со столбцами: «Страница», «Название», «Цена за грамм (₽/г)», «Описание», «Регион», «Ссылка», «Аккордеон: ЦВЕТ, АРОМАТ И ВКУС», «Аккордеон: СОСТАВ», «Аккордеон: КОЛИЧЕСТВО В УПАКОВКЕ», «Аккордеон: ВЕС ПАКЕТИКА (Г)».

2.2 Отбор признаков

Далее таблица была преобразована для дальнейшей работы. Были созданы переменные:

1. «Количество ингредиентов» — счетная переменная, количество элементов столбца «Аккордеон: СОСТАВ», разделенных по знаку запятой.
2. «Россия», «Китай», «Шри-Ланка», «Индия», «ЮАР», «Тайвань» — бинарные переменные, отвечающие за страну, принимающие значение 1, если чай содержал эту страну в столбце «Регион» и 0 иначе. В качестве значения по умолчанию была взята запись «Нет Данных» из того же столбца.
3. «Купаж» — бинарная переменная, принимающая значение 1, если в столбце «Описание» есть упоминание слова «купаж» и 0 иначе.
4. «Содержит ароматизатор» — бинарная переменная, принимающая значение 1, если в столбце «Аккордеон: СОСТАВ» есть «ароматизатор» и 0 иначе.
5. «Непакетированный» — бинарная переменная, принимающая значение 1, если чай не пакетированный и 0 иначе, с проверкой по столбцу «Аккордеон: КОЛИЧЕСТВО В УПАКОВКЕ».
6. «Прозрачный» и «Непрозрачный» — бинарные переменные, отвечающие за прозрачность и непрозрачность соответственно, взятые из столбца «Аккордеон: ЦВЕТ, АРОМАТ И ВКУС». Эти переменные были обе добавлены в таблицу, так как

анализ датасета выявил, что есть чаи, для которых не указано ни прозрачноть, ни непрозрачность, так что обе перменные принимали значение ноль.

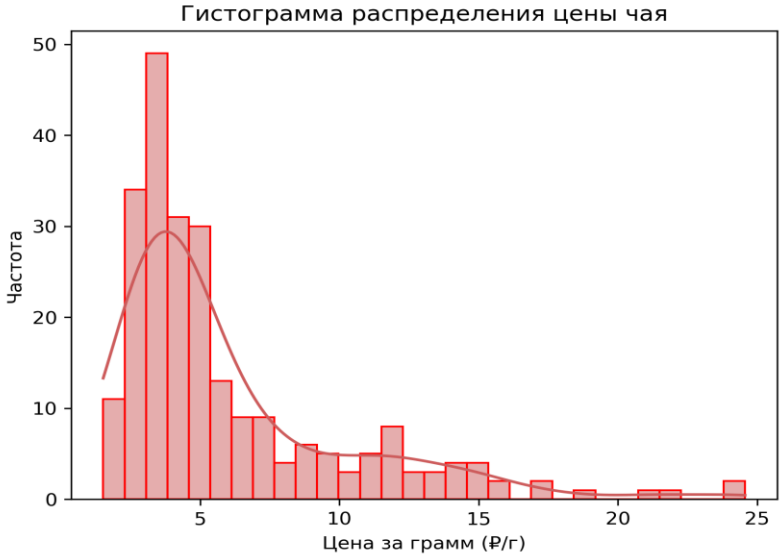
7. «Интенсивный» и «Неинтенсивный» — бинарные переменные, отвечающие за интенсивность и не интенсивность соответственно, взятые из столбца «Аккордеон: ЦВЕТ, АРОМАТ И ВКУС». По аналогии с прозрачностью были обе добавлены в переменные, так как некоторые чаи не содержали информацию о интенсивности.
8. «Терпкость» — переменная, отвечающая за разную степень терпкости, для её создания была прописана шкала, определяющая значение переменной в зависимости от содержания столбца «Аккордеон: ЦВЕТ, АРОМАТ И ВКУС». Для «незначительной терпкости» было взято значение переменной 1, для «средней» — 2, для «значительной» — 3, для «терпкой» — 4 и для «высокой» — 5. Ноль был взят в качестве значения по умолчанию, если столбец не содержал ключевые слова, приписывающие переменной определенное значение от 1—5, перечисленные ранее.
9. «Кислотность» — переменная, отвечающая за разную степень кислотности чая, для её создания также была прописана шкала, определяющая значение переменной в зависимости от содержания столбца «Аккордеон: ЦВЕТ, АРОМАТ И ВКУС». Всего было выявлено три уровня, с нулём в качестве значения по умолчанию. Для «умеренной кислотности» — 1, для «средней» — 2 и для «высокой» — 3.
10. «Продолжительное послевкусие», «Кислота», «Горечь», «Сладкость», «Фрукты», «Цветы», «Травы», «Острота» и «Орехи» — бинарные переменные, принимающие значение 1, если в столбце «Аккордеон: ЦВЕТ, АРОМАТ И ВКУС» встречаются слова, входящие в эти категории (подробнее см. в коде), и 0 иначе.
11. «Интенсивность послевкусия» — переменная, принимающая значения от 1 до 6, с 1 отвечающей за «слабую интенсивность», 2 — за «умеренную», 3 — за «выраженную», 4 — за «яркую», 5 — за «интенсивную» и 6 — за «мощную». Данные были взяты из столбца «Аккордеон: ЦВЕТ, АРОМАТ И ВКУС». За значение по умолчанию взята 2.

После создание вышеперечисленных переменных был произведен их отбор. Для этого была проанализирована их представимость в датасете. Были удалены такие переменные, как: «Фрукты», «Цветы», «Травы», «Острота», «Орехи», а также «Кислота», по причиние того, что она по смыслу дублирует переменную «Кислотность». Эти признаки были убраны, так как их доля в датасете мала из-за чего они с большей вероятностью окажутся незначимыми в итоговой модели, а также могут привести к снижению её обобщающей способности.

2.3 Анализ описательных статистик и графический анализ переменных

В нашем наборе данных нет пропусков, поэтому их не пришлось никак обрабатывать. Мы провели анализ итогового набора данных, рассмотрев каждый фактор.

Цена за грамм чая (Р/г): По описательным статистикам заметно, что цены большей части наблюдаемых экземпляров чая лежат в диапазоне 1 - 6 Р/г (Приложение 1, рис.1). При этом имеются выбросы в правом хвосте, 1 грамм самого дорогого чая стоит больше 24 рублей.



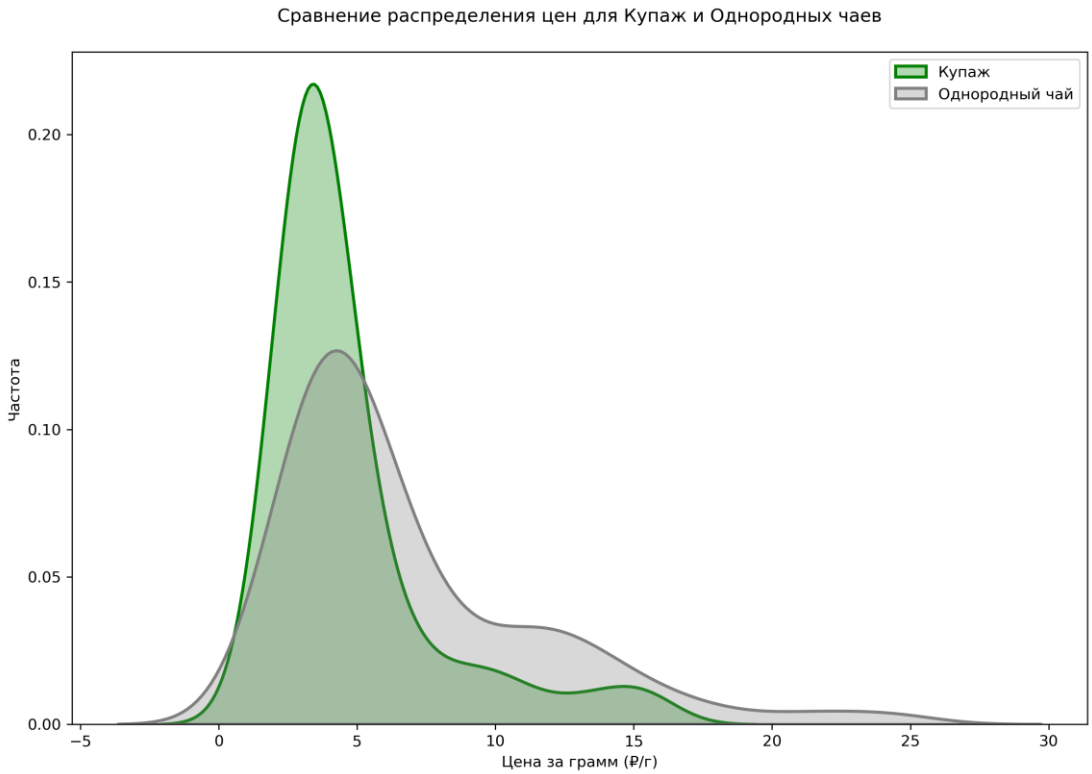
Гистограмма подтверждает предположение о выбросах

Заметна асимметрия распределения цены, поэтому есть стимул оценить модель с логарифмированной ценой.

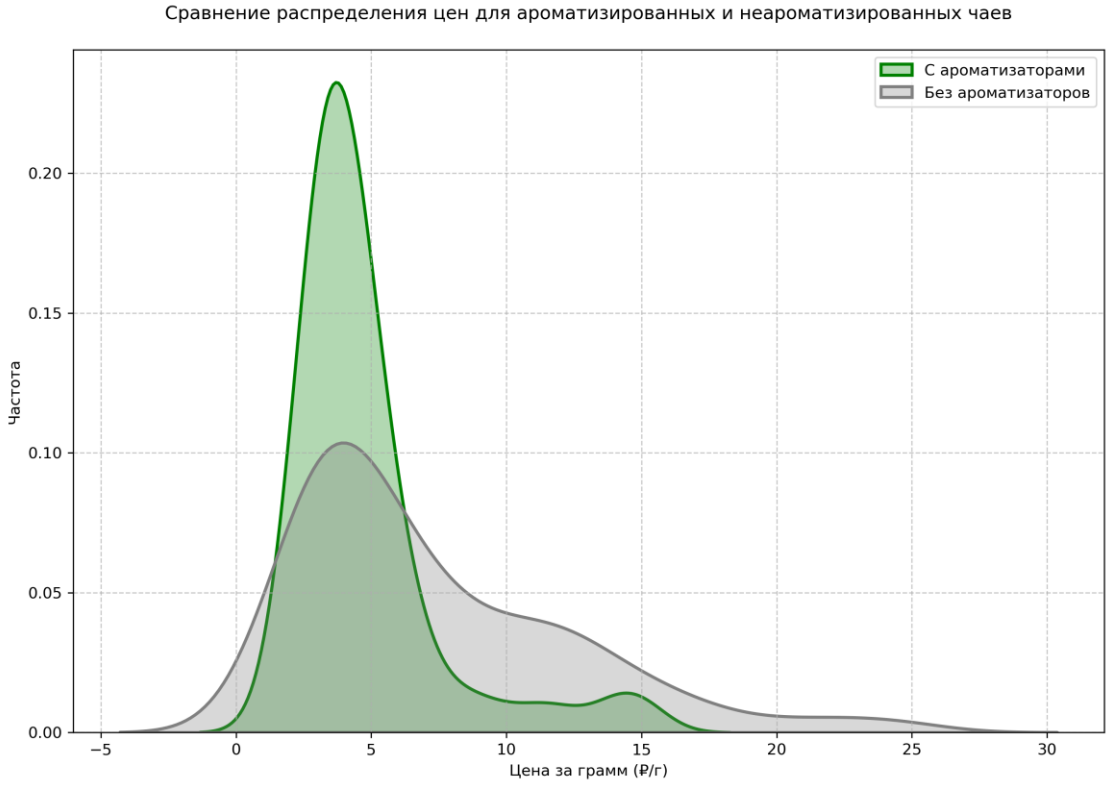
Количество ингредиентов: В выборке представлены чаи с разным количество ингредиентов, от 1 до 12. Анализ распределения цен чаев с различным количеством ингредиентов (см. график в приложениях) натолкнул на мысль, что в среднем моночаи (1 ингредиент в составе) стоят дороже. В дальнейшем мы проверим гипотезу о равенстве распределений этих двух групп чаев.

Страна-производитель: В анализируемых данных представлены чаи из 6 стран: России, Китая, Тайваня, Шри-Ланки, Индии и ЮАР. Большинство чаев из выборки произведены в России. Также заметим, что чай из Тайваня редко встречается, но при этом в большинстве случаев очень дорогой относительно других стран (средняя цена около 11.5 против 6.04 по всей выборке). Китайский чай встречается чаще, и при этом в среднем тоже существенно дороже чая из остальных стран (средняя цена около 9.44 против 6.04 по всей выборке) (Приложение 1, рис.2).

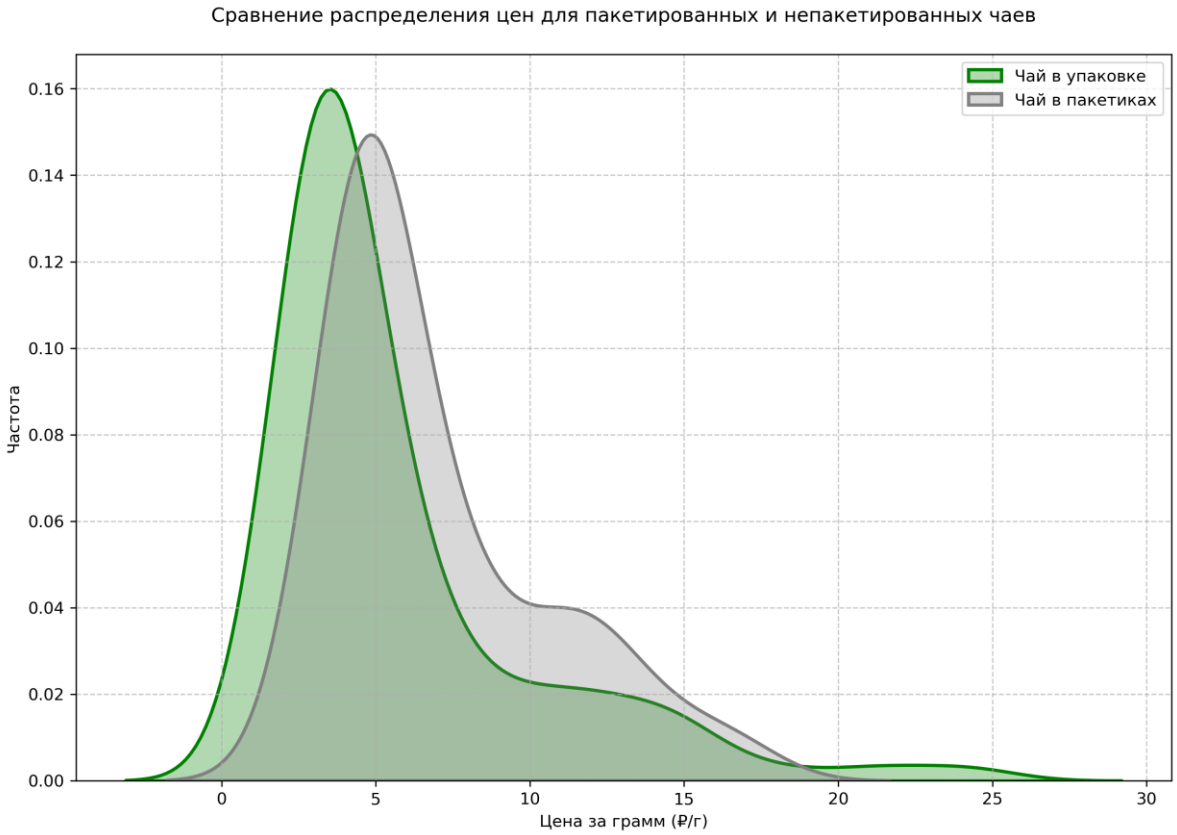
Купаж: Купаж – это чай, составленный из двух или более ингредиентов. В ~40% случаев чай является купажом. Визуально, распределения цен моночаев и купажей схоже. Чтобы проверить это предположение, мы провели тест Манна-Уитни на равенство распределений (H0: распределение цен моночаев и купажей одинаково, H1: распределение цен моночаев и купажей различается). Исходя из теста, распределения цен различаются статистически значимо (p-value ~= 0), предположение о схожести распределений оказалось ошибочным. Для проверки гипотез о распределениях нами был выбран тест Манна-Уитни, так как распределение цен не нормальное. Это подтверждено тестом Колмогорова-Смирнова (H0: распределение цен чая соответствует нормальному распределению, H1: распределение цен чая не является нормальным), p-value ~= 0.



Наличие ароматизатора: По признаку наличия ароматизатора выборка делится в равных пропорциях. Средние распределений очень близки, при этом распределение цен чаев без ароматизаторов имеет более тяжелые хвосты. Тест Манна-Уитни на равенство распределений (H0: распределение цен ароматизированных и неароматизированных чаев одинаково, H1: распределение цен ароматизированных и неароматизированных чаев различается) показал, что распределения различаются статистически значимо (p-value = 0.0003), как мы и предполагали, исходя из гистограмм.



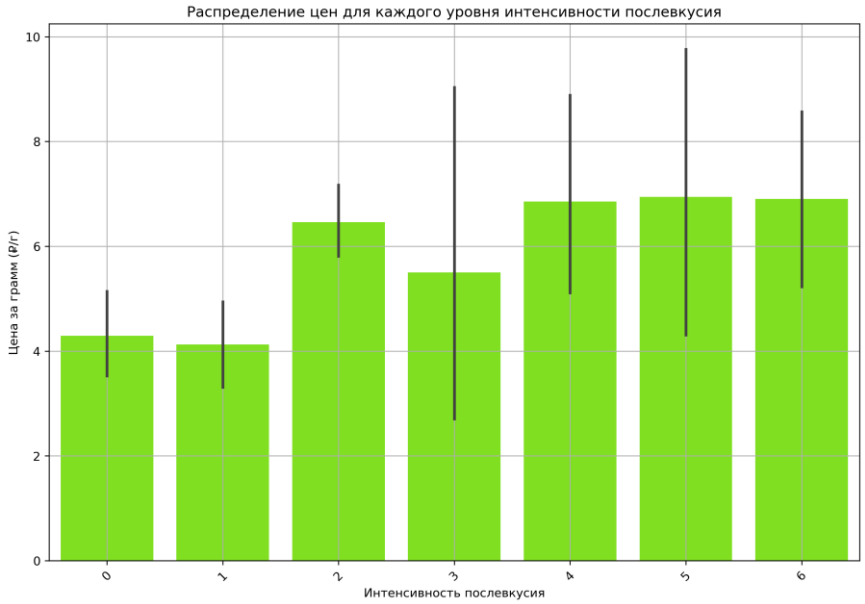
Непакетированный: Чаи в упаковке преобладают в выборке, при этом, в среднем чай в пакетиках стоит дороже, так как вид распределений схож, но распределение пакетированных чаев смещено вправо.



Терпкость: Чаи в выборке проранжированы исходя из уровня терпкости (от 0 до 5). Заметим (Приложение 1, рис.3), что чаи с уровнем терпкости 1 в среднем дороже. Это может быть связано с тем, что низкую терпкость предпочитает наибольшее количество людей.

Кислотность: По признаку кислотности чаи проранжированы от 0 до 3 (от отсутствия до сильной кислотности). Анализ цены по уровням кислотности дает схожую картину (Приложение 1, рис.4). Наибольшая средняя цена у чаев с умеренной кислотность. При этом таких чаев больше всего в выборке. Это может быть связано с тем что умеренную кислотность предпочитает большинство людей, следовательно производители, максимизирующие продажи, концентрируются на производстве таких чаёв.

Интенсивность послевкусия: Этот признак так же порядковый (от 0 до 6). Анализ цены по категориям показывает, что цены чаев с низкой или отсутствующей интенсивностью обычно ниже. Проверка этого предположения с помощью теста Манна-Уитни (H0: распределение цен чаев из группы 1 (низкая и отсутствующая интенсивность) и из группы 2 (средняя интенсивность и выше) одинаково, H1: распределение цен чаев между этими группами различается) показала, что распределения цен в этих группах действительно различаются на всех уровнях значимости (p-value ~= 0).



Горечь: В ~90% чаев из выборки отсутствует горечь. Интуитивно, факт наличия горечи должен отрицательно влиять на цену. В нашей выборке если и есть влияние горечи, то оно незначительно (Приложение 1, рис.5).

Сладость: В среднем сладкие чаи чуть дороже, но их мало в выборке (~15%).

3. Базовая модель

3.1 Модель с отобранными по AIC переменными

После получения готового датасета и первичного анализ переменных сначала была построена модель, использующая все переменные. Для неё, чтобы проверить ранее выдвинутое предположение о необходимости логарифмирования таргета для лучшей предсказательной и объясняющей способности модели, был проведен тест Бокса-Кокса.

H_0 : качество подгонки линейной и полулогарифмической моделей одинаковое
H_1 : модель с меньшей RSS лучше

Сравнение модели со всеми линейными переменными с моделью с логарифмированным таргетом показало, что полулогарифмическая модель лучше описывает данные. Логарифимирование правой части модели с переменными не производилось, так как большинство переменных бинарные и их преобразование не имеет смысла.

Далее был проделан отбор фичей по AIC, в результате которого в модели осталось 9 переменных: «Россия», «Китай», «Тайвань», «Содержит ароматизатор», «Непакетированный», «Прозрачный», «Непрозрачный», «Интенсивный», «Продолжительное послевкусие» и константа.

Для модели с отобранными по AIC переменными были проведенные тесты, а именно: тесты Уайта, Глейзера и Бройша-Пагана, а также визуальный по графику «Остатки-Прогнозы» (Приложение 2, рис.1), которые выявили гетероскедастичность. При этом проверялись следующие гипотезы:

Глейзер:

- H_0: Дисперсия ошибок постоянна (гомоскедастичность).

$$\mathrm{Var}(u_i|X) = \sigma^2 \quad \forall i$$

- H_1: Дисперсия ошибок зависит от объясняющих переменных (гетероскедастичность).

$$\mathrm{Var}(u_i|X) = \sigma_i^2, \quad \sigma_i^2 = f(X_i)$$

Уайт:

- H_0: Остатки гомоскедастичны (дисперсия ошибок постоянна).

$$\mathrm{Var}(u_i|X) = \sigma^2 \quad \forall i$$

- H_1: Остатки гетероскедастичны (дисперсия ошибок непостоянна и зависит от X).

$$\mathrm{Var}(u_i|X) \neq \sigma^2 \quad \text{для некоторых } i$$

Бройш-Паган:

- H_0: Гомоскедастичность (дисперсия ошибок не зависит от X).

$$\mathrm{Var}(u_i|X) = \sigma^2 \quad \forall i$$

- H_1: Гетероскедастичность (дисперсия ошибок линейно зависит от X)

$$\mathrm{Var}(u_i|X) = \sigma^2 + \alpha_1 X_{i1} + \cdots + \alpha_k X_{ik}$$

Модель оценили с ковариационной матрицей HC3 для борьбы с выявленной гетероскедастичностью. Её R-squared adjusted оказался равен приблизительно 0.285.

$$\begin{aligned}\widehat{\text{Цена за грамм}}\ (₽/\text{г}) = {} & 1.720^{***} - 0.149 \cdot \text{Россия}^{*} + 0.309 \cdot \text{Китай}^{**} + 0.506 \cdot \text{Тайвань} - \\ & - 0.142 \cdot \text{Содержит ароматизатор}^{*} - 0.382 \cdot \text{Непакетированный}^{***} + \\ & + 0.382 \cdot \text{Прозрачный}^{***} + 0.276 \cdot \text{Непрозрачный} - 0.231 \cdot \text{Интенсивный}^{***} + \\ & + 0.169 \cdot \text{Продолжительное послевкусие}^{**}\end{aligned}$$

3.2 Итоговая модель

Оценку R-squared adjusted модели путем перебора переменных в полулогарифмической модели получилось немного увеличить. Для этой модели также все проведенные тестыа именно: тесты Уайта, Глейзера и Бройша-Пагана, а также визуальный по графику «Остатки-Прогнозы», показали наличие гетероскедастичности (Приложение 2, рис.2). Для получения более устойчивых оценок она была оценена аналогичным методом, как и предыдущая модель, OLS с робастными стандартными ошибками (cov_type='HC3'). Её R-squared adjusted оказался равен приблизительно 0.288.

$$\begin{aligned}\widehat{\text{Цена за грамм}}\ (₽/\text{г}) = {} & 1.709^{***} + 0.023 \cdot \text{Количество ингредиентов} - 0.179 \cdot \text{Россия}^{*} + \\ & + 0.322 \cdot \text{Китай}^{**} + 0.535 \cdot \text{Тайвань} - 0.120 \cdot \text{Купаж} - \\ & - 0.174 \cdot \text{Содержит ароматизатор}^{**} - 0.394 \cdot \text{Непакетированный}^{***} + \\ & + 0.386 \cdot \text{Прозрачный}^{***} + 0.283 \cdot \text{Непрозрачный} - 0.231 \cdot \text{Интенсивный}^{***} \\ & + 0.176 \cdot \text{Продолжительное послевкусие}^{**}\end{aligned}$$

Эту модель после оценки зафиксировали как итоговую и провели более широкий ряд тестов. Первым из которых был тест на адекватность модели, который показал, что модель адекватна. Для этого проверили нулевую гипотезу о том, что все коэффициенты, кроме

коэффициента перед константой равны нулю против альтернативной о том, что хотя бы один коэффициент не перед константой не равен нулю.

$$H_0 : \beta_2 = \cdots = \beta_m = 0$$
$$H_1 : \exists\, i : \quad \beta_i \neq 0 \text{ (модель адекватна)}$$

	Классич. ошибки	HC3 ошибки	Разница (%)
const	0.1277	0.1158	-9.3145
Количество ингредиентов	0.0185	0.0161	-13.1162
Россия	0.0996	0.089	-10.6349
Китай	0.1121	0.1362	21.4719
Тайвань	0.2164	0.3484	60.9488
Купаж	0.0816	0.0886	8.4972
Содержит ароматизатор	0.0887	0.0783	-11.6521
Непакетированный	0.0948	0.0901	-5.0244
Прозрачный	0.1204	0.1216	1.0224
Непрозрачный	0.161	0.1866	15.9259
Интенсивный	0.0768	0.0857	11.6282
Продолжительное послевкусие	0.0769	0.0787	2.3363

В модели с HC3 стандартные ошибки для регрессоров «Китай» и «Тайвань» серьезно выросли в сравнении с исходной моделью (на 21 и 60% соответственно), что говорит о локальной гетероскедастичности. Причем регрессоры «Россия» и «Содержит ароматизатор» оказались значимы в модели с HC3, что подтверждает устойчивость результатов модели с HC3.

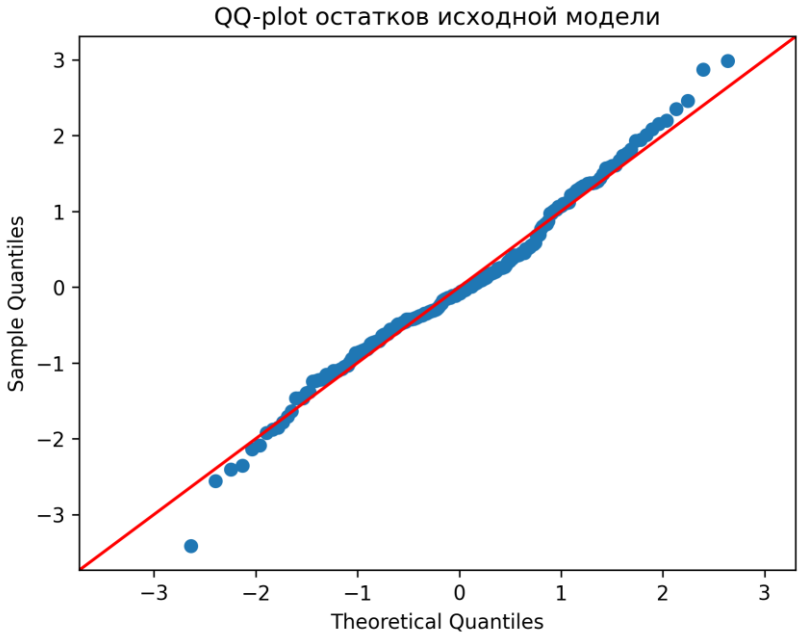
После анализа гетероскедастичности было проведено сравнение модели с выбранными признаки и этой же моделью, но из которой убрали незначимые переменные. В модели без них снизился R-squared adjusted.

Были проведены тесты на необходимость делить выборку на две подвыборки на пакетрированный и непакетрированный чай и было выявлено, что модели статистически не различаются и делить выборку не имеет смысла. В отличии от деления по стране: модель для России оказалась отличной от модели для всех остальных стран. Так R-squared adjusted для России оказался высоким — 0.329, в то время как для остальных стран он сильно снизился по сравнению с итоговой моделью — до 0.178. Было решено оставить итоговую модель общей по всем странам.

Мультиколлинеарности в итоговой модели не было обнаружено по итогам анализа матрицы корреляций (нет высоких корреляций), также как и по VIF, VIF больше 10 не наблюдалось (Приложение 2, рис.3).

3.3 Нормальность остатков

Для понимания, насколько адекватны оценки доверительных интервалов коэффициентов, был проведен анализ нормальности случайных ошибок, а точнее остатков, так как только они могут быть вычислены по данным. Для проверки остатков на нормальность был построен Quantile-Quantile plot, демонстрирующий, насколько распределение остатков модели соответствует нормальному распределению. Красная линия в данном случае соответствует нормальному распределению. Quantile-Quantile plot был построен на исходной модели (без HC3), так как корректировка на гетероскедастичность может исказить выводы.

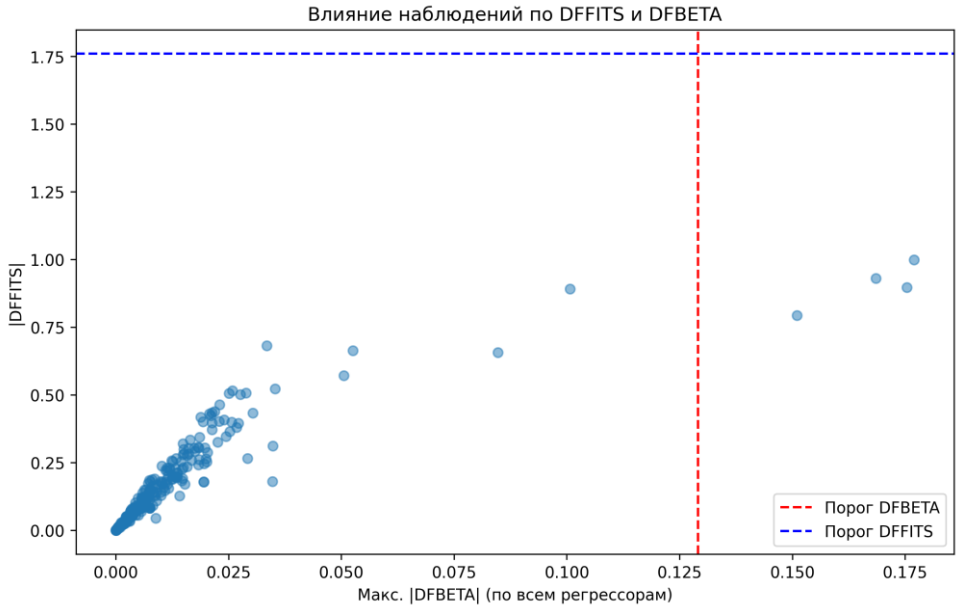


По графику видно, что отклонения есть, но они не очень сильные. Точки выше графика в правом хвосте и ниже графика в левом говорят о правосторонней асимметрии. Был проведен тест Шапиро-Уилка на нормальность остатков (H0: остатки регрессии распределены нормально, H1: остатки регрессии не распределены нормально). P-value теста = 0.033. Исходя из теста гипотеза о нормальности остатков отвергается, но при этом p-value не сильно меньше 0.05. Комплексный анализ Quantile-Quantile plot и теста Шапиро-Уилка дает стимул сделать вывод, что распределение остатков близко к нормальному, то есть допущение о нормальности не должно исказить результатов оценивания. Для того, чтобы удостовериться в корректности выводов, были построены доверительные интервалы для оценок коэффициентов с помощью парного бутстрапа. Разница ширины доверительных интервалов из модели и доверительных интервалов, построенных бутсрапом, меньше 10% для каждого коэффициента, пересечение этих доверительных интервалов > 93% для каждого коэффициента. Это подтверждает устойчивость оценок.

3.4 Анализ наличия выбросов

Анализ был проведен с помощью DFFITS и DFBETA. DFBETA показывает, насколько изменится каждый коэффициент регрессии, если исключить конкретное наблюдение. Принято считать, что если $DFBETA > 2/\sqrt{n}$, где n – число наблюдений, то наблюдение сильно влияет на коэффициент. DFFITS показывает, насколько сильно изменится предсказание, если исключить конкретное наблюдение. Большое значение DFFITS говорит о том, что наблюдение искажает предсказания. Считается, что у таких наблюдений $DFFITS > 2/\sqrt{k/n}$, k – число регрессоров. Мы провели комплексный анализ на наличие выбросов исходя из значений DFBETA и DFFITS. Анализ проводился на обычной модели (без HC3), так как целью было выявление влиятельных наблюдений. DFBETA каждого наблюдения вычислялось как максимальное по модулю значение из оценок этого наблюдения с каждым из регрессоров. DFFITS каждого наблюдения вычислялось как модуль из DFFITS этого наблюдения (так как предсказанное значение при удалении наблюдения может измениться как в большую, так и в меньшую сторону). По графику

ниже можно сделать вывод, что влияние на коэффициенты есть, но нет влияния на предсказания, так как нет точек, превышающих порог по DFFITS.



Наблюдений, серьезно влияющих на предсказания, нет, а наблюдений, значимо влияющих на оценки коэффициентов, всего 4. Все эти наблюдения значимо влияют на оценку коэффициента регрессора «Китай». Доверительный интервал для коэффициента перед этим регрессором, оцененный бутстрапом, не включает 0, что говорит о значимости регрессора и устойчивости оценки коэффициента. Поэтому было принято решение оставить данные наблюдения, их влияние незначительно.

3.5 Тестирование модели

Далее было сделано предсказание с помощью итоговой модели для нашего собственного товара с указанными ниже характеристиками:

'const': [1], 'Количество ингредиентов': [5], 'Россия': [1], 'Китай': [0], 'Тайвань': [0], 'Купаж': [0], 'Содержит ароматизатор': [1], 'Непакетированный': [0], 'Прозрачный': [1], 'Непрозрачный': [0], 'Интенсивный': [0], 'Продолжительное послевкусие': [1].

Оказалось, что, согласно модели, 1 грамм чая с подобными характеристиками должен стоить 7.65*₽*. Данный результат является чуть более высоким, чем средний фактический на рассматриваемой выборке (~6.01), а также входит в доверительный 95% интервал [6.31, 9.28] для предсказания.

4. Квантильная регрессия

Дальнейший исследовательский интерес нашей команды был связан с определением того, различным ли образом влияют на цену 1 грамма чая отобранные нами признаки. Чтобы с этим разобраться, мы прибегли к модели квантильной регрессии. Мы рассматривали 3 ценовые сегмента: дешевый (q = 0.1), средний (q = 0.5) и дорогой (q = 0.9). Для каждого из них были посчитаны оценки коэффициентов и применен тест Вальда для сравнения этих коэффициентов (H0: коэффициенты при признаке равны для разных квантилей, H1: коэффициенты при признаке не равны для разных квантилей).

Было выяснено, что значимыми признаками для дорогого сегмента являются «const», «Китай», «Содержит ароматизатор», «Непакетированный» и «Продолжительное послевкусие», а итоговая модель имеет вид**:**

$$\begin{aligned}\widehat{\text{Цена за грамм}}\ (₽/\text{г}) = 2.385 &+ 0.013 \cdot \text{Количество ингредиентов} - 0.217 \cdot \text{Россия} + 0.617 \cdot \text{Китай} + \\ &+ 0.760 \cdot \text{Тайвань} - 0.105 \cdot \text{Купаж} - 0.347 \cdot \text{Содержит ароматизатор} - \\ &- 0.517 \cdot \text{Непакетированный} + 0.370 \cdot \text{Прозрачный} + 0.422 \cdot \text{Непрозрачный} \\ &- 0.135 \cdot \text{Интенсивный} + 0.277 \cdot \text{Продолжительное послевкусие}\end{aligned}$$

Для среднего ценового сегмента оказались значимыми «const», «Россия», «Непакетированный», «Прозрачный», «Непрозрачный» и «Продолжительное послевкусие», а модель выглядит:

$$\begin{aligned}\widehat{\text{Цена за грамм}}\ (₽/\text{г}) = 1.530 &+ 0.020 \cdot \text{Количество ингредиентов} - 0.253 \cdot \text{Россия} + 0.338 \cdot \text{Китай} + \\ &+ 0.655 \cdot \text{Тайвань} - 0.047 \cdot \text{Купаж} - 0.090 \cdot \text{Содержит ароматизатор} - \\ &- 0.332 \cdot \text{Непакетированный} + 0.473 \cdot \text{Прозрачный} + 0.515 \cdot \text{Непрозрачный} \\ &- 0.253 \cdot \text{Интенсивный} + 0.159 \cdot \text{Продолжительное послевкусие}\end{aligned}$$

Для дешевого сегмента оказались значимыми «const», «Количество ингредиентов», «Россия», «Купаж», «Непакетированный», а итоговая модель:

$$\begin{aligned}\widehat{\text{Цена за грамм}}\ (₽/\text{г}) = 1.310 &+ 0.041 \cdot \text{Количество ингредиентов} - 0.182 \cdot \text{Россия} - 0.277 \cdot \text{Китай} + \\ &+ 0.263 \cdot \text{Тайвань} - 0.188 \cdot \text{Купаж} - 0.007 \cdot \text{Содержит ароматизатор} - \\ &- 0.429 \cdot \text{Непакетированный} + 0.179 \cdot \text{Прозрачный} + 0.101 \cdot \text{Непрозрачный} \\ &- 0.057 \cdot \text{Интенсивный} + 0.041 \cdot \text{Продолжительное послевкусие}\end{aligned}$$

Была выявлена разница в значимости для одного признака: «Китай» на дешевом и дорогом сегменте (p-value 0.0019) и на среднем и дешевом (p-value 0.0413).

Следующим шагом мы отобразили доверительные интервалы для коэффициентов квантильной регрессии и полулогарифмической регрессии (Приложение 3). Из них видно, что значимые коэффициенты, а также их оценки различаются для квантилей, а также оценки и доверительные интервалы для квантильной и полулогарифмический регрессии различаются.

Учитывая все выше указанные факторы, мы можем утверждать, что на разные ценовые сегменты чая влияют разные факторы.

5. Заключение

Подводя итоги исследования, наша команда пришла к некоторым интересным результатам:

Статистический анализ данных привел к выводам относительно факторов, определяющих ценообразования чая. Например, чаи со средней и выше интенсивностью послевкусия имеют статистически значимую премию в цене по сравнению с чаями, у которых послевкусие слабое или отсутствует. Ещё одним значимым результатом стало то, что моночаи (однокомпонентные) в среднем значительно дороже купажированных чаев (многокомпонентных смесей). Это может быть связано с тем, что себестоимость легко балансировать за счет композиции, а также с большей популярностью классических моночаев у потребителей.

Результаты регрессионного анализа дополнили эти выводы. Итоговая модель показала, что чаи из Китая значительно дороже, чем те чаи, у которых не указана страна производства. В то время как чаи из России дешевле, что совпало с результатами квантильной регрессии. Однако вывод о стоимости непакетированного чая разошелся с аналогичным выводом в квантильной регрессии: непакетированный чай оказался значительно дешевле пакетированного.

Также из модели видно, что люди, покупающие чай в специализированных магазинах, таких как https://teaco.ru/, могут предпочитать более качественный чай — без ароматизаторов, неинтенсивный, но с продолжительным послевкусием.

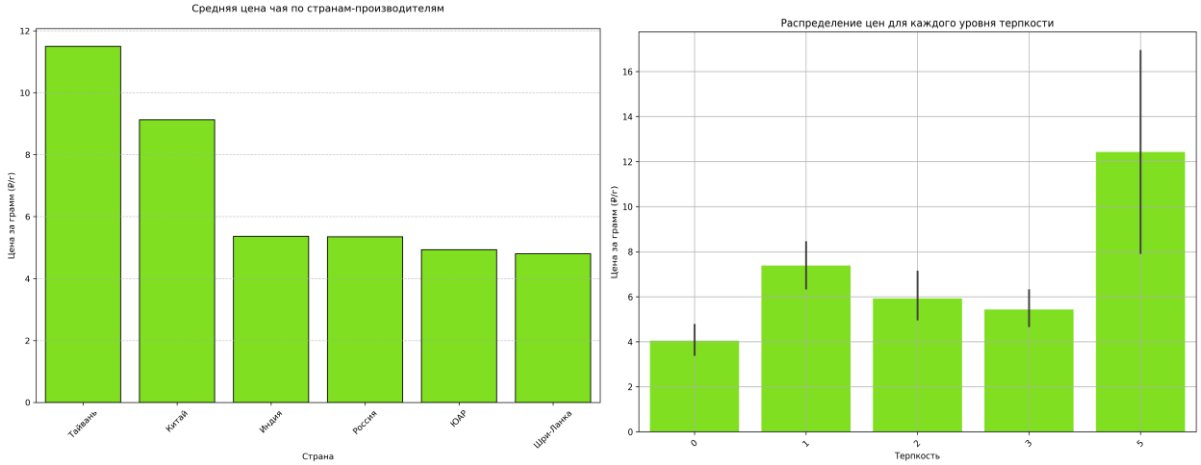
Квантильный анализ выявил, что чаи, которые производятся в России, для всех ценновых сегментов являются более дешевыми при прочих равных. Также было замечено, что непакетерованные чаи во всех сегментах стоят при прочих равных дороже, чем пакетированные, вопреки сырьевым затратам на упаковку, которые должны увеличивать цену пакетированного чая. Это интересный эффект, который открывает большое поле для исследований.

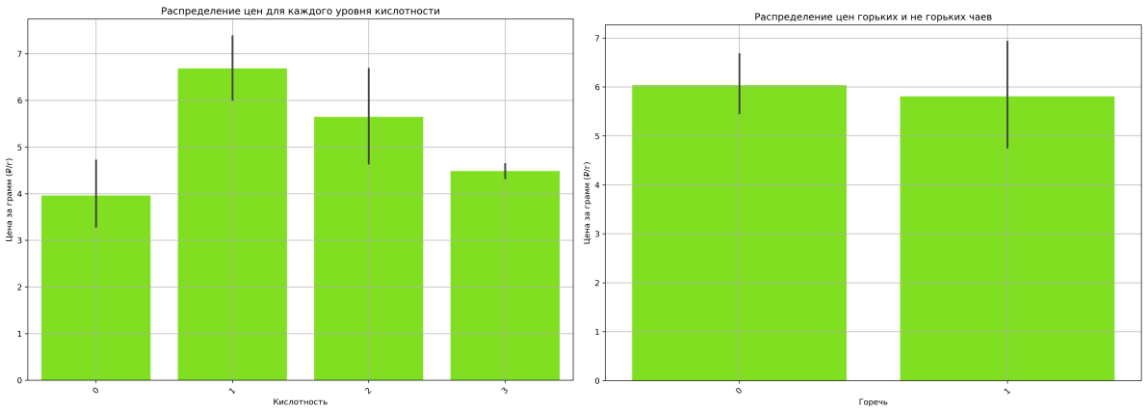
В заключение, наша команда верит, что изучение ценообразования чая крайне важная тема, как и для исследователей эконометристов, так и для потребителей. Мы надеемся, что большее количество исследовательских объединений обратит на эту проблему внимание и придет к не менее прорывным выводам.

6. Приложение

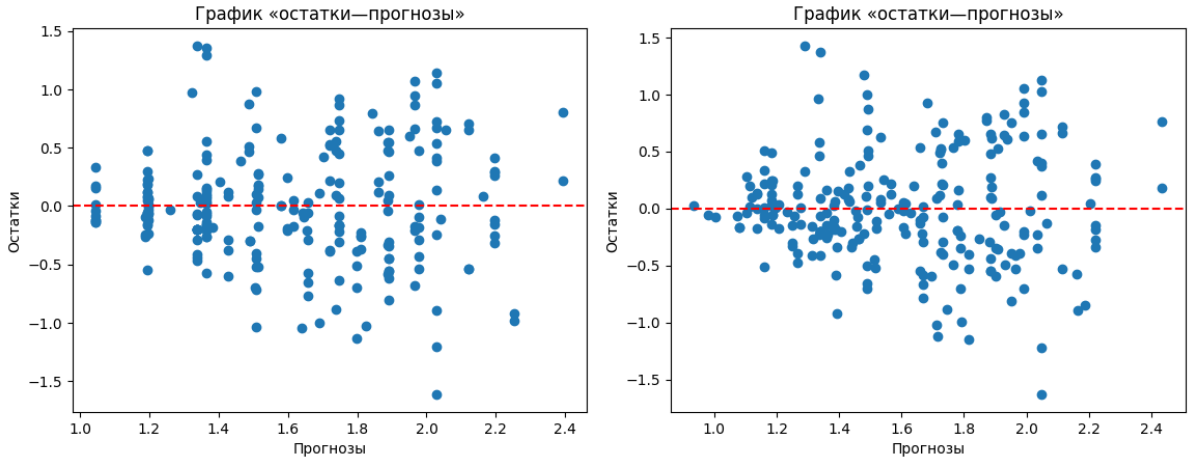
Приложение 1:

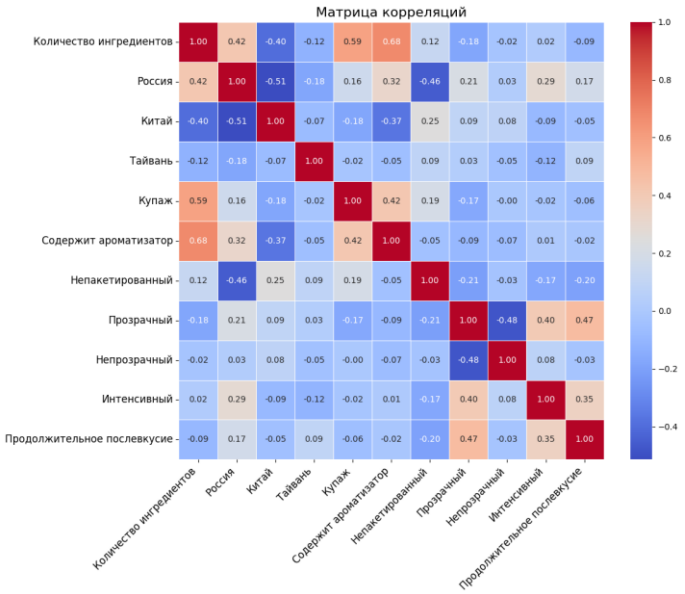
	price
count	240
mean	6.011859
std	4.223824
min	1.516
25%	3.269
50%	4.415
75%	6.995
max	24.566667



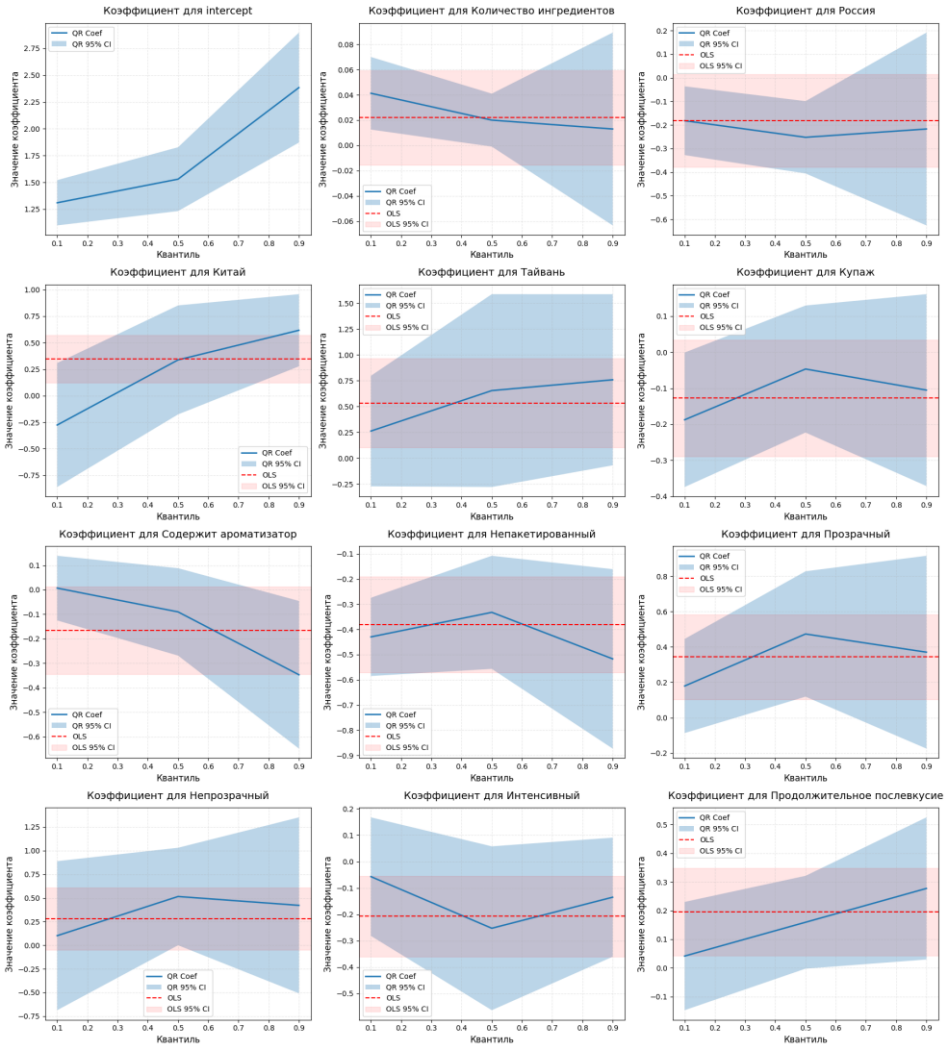


Приложение 2:





Приложение 3:


Коэффициент для intercept
Коэффициент для Количество ингредиентов
Коэффициент для Россия
Коэффициент для Китай
Коэффициент для Тайвань
Коэффициент для Купаж
Коэффициент для Содержит ароматизатор
Коэффициент для Непакетированный
Коэффициент для Прозрачный
Коэффициент для Непрозрачный
Коэффициент для Интенсивный
Коэффициент для Продолжительное послевкусие
QR Coef
QR 95% CI
OLS
OLS 95% CI
Значение коэффициента
Квантиль