Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет «Высшая школа экономики»

Факультет экономических наук

Отчет
по проектной работе

Оценивание гедонистической ценовой функции для шоколадных плиток с помощью моделей множественной регрессии

Выполнили студенты:

ФИО	**% вклада**	**Вклад**
Белоцерковец А.Д. БЭАД223	33.(3)%	Парсинг данных и обработка датасета, прогнозирование цены на товар, отсутствующий в выборке, оформление отчета
Кирщин И.А. БЭАД223	33.(3)%	Обработка данных и графический анализ переменных, построение модели квантильной регрессии, интерпретация полученных результатов, оформление отчета
Мирзоева А.Б. БЭАД223	33.(3)%	Проведение тестов на проверку выполнения ТГМ на данных, оценка модели линейной регрессии, интерпретация полученных результатов, оформление отчета

2025

Содержание

1 EDA

1.1 Сбор данных

Для анализа и дальнейшего построения моделей нами были выбраны данные, включающие основные характеристики и цены шоколада с сайта магазина «Магнит» с локализацией в Москве. Сбор данных был реализован по следующему алгоритму:

1. **Извлечение данных:** Для каждой товарной карточки извлекались 4 секции:
   - Основные характеристики (вес, тип упаковки, содержание какао)
   - Ценовое предложение (текущая/старая цена, скидка, рейтинг)
   - Пищевая ценность (БЖУ, калорийность)
   - Состав продукта
2. **Преобразование данных:**
   - Числовые поля (цены, проценты, БЖУ) очищались от нечисловых символов и конвертировались в числовой формат.
   - Текстовые поля (название, состав) нормализовывались: удалялись лишние пробелы, спецсимволы
   - Категориальные признаки (бренд, тип шоколада) унифицировались через словарь замен
3. **Структурирование:** Данные агрегировались в табличный формат с унифицированными колонками:

- Приоритетные поля: название, цена, артикул, ссылка
- Дополнительные атрибуты: рейтинг, отзывы, качественные характеристики

В результате была получена выборка из 317 товаров, содержащая 20+ характеристик товара, включая название, текущую цену, бренд, производителя, рейтинг, количество калорий, вес, тип упаковки, тип шоколада, процент содержания какао, состав. Интеграция всех этапов в единый конвейер обеспечила воспроизводимость и возможность масштабирования на другие категории товаров.

1.2 Формирование выборки

Из данных, полученных в результате парсинга, были отобраны шоколадные плитки весом 75-100 г. Шоколадные плитки весом больше 100 г. не включались в выборку, так как они в подавляющем большинстве являются увеличенными версиями стандартных шоколадных плиток, уже включенных в выборку, и отличаются лишь весом (например, «Milka» любит выпускать свои шоколадные плитки не только порциями по 90 г., но и порциями по 260 г.).

Получившиеся данные представляют собой пространственную выборку шоколадных плиток весом 75-100г. с прилавков магазина «Магнит», содержащую 168 наблюдений. Каждая шоколадная плитка характеризуется следующим набором признаков:

- **Цена.** Целевая переменная. Количественная переменная, равная текущей цене шоколадной плитки в интернет-магазине «Магнит» на 4 мая.
- **Вес.** Количественная переменная, равная весу шоколадной плитки в граммах.
- **Российский бренд.** Дамми-переменная на то, является ли бренд шоколада российским. 1 – если бренд является российским, 0 – если бренд не является российским («Milka», «Schogetten», «Ritter Sport», «Premiere of taste», «Merci», «Kinder»).
- **Тип шоколада.** Категориальная переменная, принимающая 4 значения: Белый, Молочный, Темный, Горький.
- **Калорийность.** Количественная переменная, равная количеству ккал / 100 г. в шоколадной плитке.
- **Наличие орехов в начинке.** Дамми-переменная на наличие орехов в начинке. 1 – если в начинке содержатся орехи, 0 – иначе.
- **Наличие ягод или сухофруктов в начинке.** Дамми-переменная на наличие ягод или сухофруктов в начинке. 1 – если в начинке содержатся ягоды или сухофрукты, 0 – иначе.
- **Наличие снеков в начинке.** Дамми-переменная на наличие снеков в начинке. 1 – если в начинке содержатся снеки (печенье/хлопья/вафли и т.д.), 0 – иначе.
- **Наличие десертной начинки.** Дамми-переменная на наличие десертной начинки. 1 – если в шоколаде содержится десертная начинка (йогурт/тирамису/помадка/крем-брюле и т.д.), 0 – иначе.
- **Наличие экзотической начинки.** Дамми-переменная на наличие различных необычных компонентов в начинке. 1 – если в начинке есть не совсем привычные в шоколаде компоненты (лаванда/роза/мёд/матча/кунжут и т.д.), 0 – иначе.
- **Доля содержания какао.** Количественная переменная, равная доле содержания какао в шоколадной плитке.
- **Упаковка.** Категориальная переменная, принимающая 3 значения: флоупак, картонная упаковка, бумажная упаковка.
- **Рейтинг.** Количественная переменная, равная рейтингу товара в интернет-магазине «Магнит».

- **Наличие пищевых добавок.** Дамми-переменная на наличие пищевых добавок в составе шоколада. 1 – если в шоколаде содержатся пищевые добавки, маркированные буквой E, 0 – иначе.
- **Отсутствие сахара.** Дамми-переменная на отсутствие сахара в составе шоколада. 1 – если в составе шоколада отсутствует сахар, 0 – иначе.
- **Преобладание сахара.** Дамми-переменная на то, что сахар в составе шоколада идет первым. 1 – если в составе шоколада сахар идет первым, 0 – иначе.
- **Пальмовое масло.** Дамми-переменная на содержание пальмового масла в составе шоколада. 1 – если в составе шоколада содержится пальмовое масло, 0 – иначе.

1.3 Обработка данных и графический анализ переменных

Теперь перейдем подробнее к самим переменным.

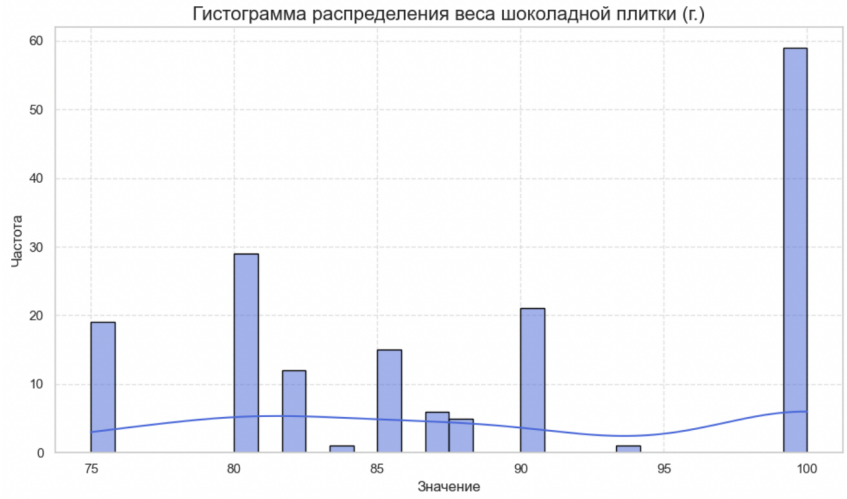
- **Цена.** Целевая переменная. Количественная переменная, равная текущей цене шоколадной плитки в интернет-магазине «Магнит» на 4 мая.

Описательные статистики:

	mean	std	min	25%	50%	75%	max
price	135.33	56.64	44.99	89.49	129.99	169.99	359.99

Выделяются 3 явных ценовых кластера, что, в том числе, можно увидеть по плотности, оцененной методом ядерной оценки.

- **Вес.**



Описательные статистики:

	mean	std	min	25%	50%	75%	max
weight	88.89	9.20	75.00	80.00	88.00	100.00	100.00

В выборке преобладают шоколадные плитки весом 100 г. Достаточно много плиток весом 80 г. Примерно равное количество плиток весом 75 г. и 90 г.

- **Российский бренд.**

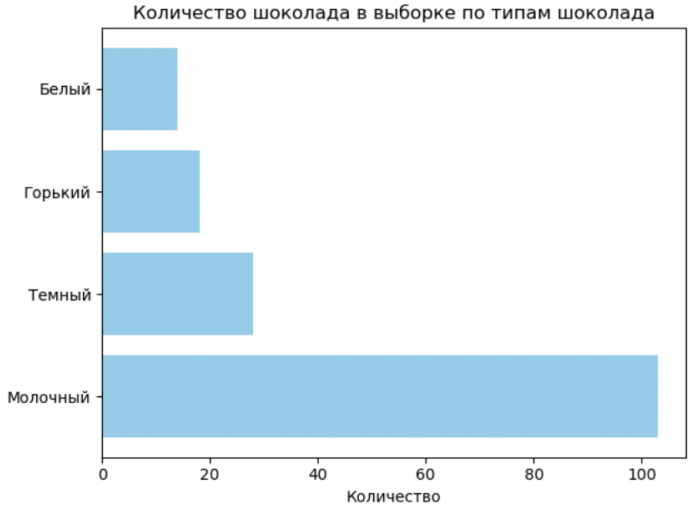
  Так как данная переменная является дамми-переменной, разумной описательной статистикой для нее будет среднее выборочное значение, т.е. доля. В имеющейся выборке 67.26% шоколадных плиток были выпущены под российскими брендами.

- **Тип шоколада.**

  Для 5 наблюдений значение данной переменной было пропущено:

	name	type
9	Шоколад молочный и белый пористый Россия щедрая душа 75г	NaN
16	Шоколад Трилогия Schogetten 100г	NaN
37	Шоколад Мороженое Schogetten 100г	NaN
156	Набор Маша и медведь шоколадный молочный Монетный двор 75г	NaN
157	Шоколад молочный набор Развивайка 75г	NaN

В 3 из данных шоколадов действительно смешаны несколько типов шоколада, поэтому однозначно отнести их к одному из типов не получится. Оставшиеся 2 выпускаются как подарочный набор, что также может затруднять определение типа шоколада, а еще данные наблюдения могут выбиваться из общей методологии. Так как таких наблюдений всего 5, удалим их из выборки.



- **Калорийность.**

Описательные статистики:

	mean	std	min	25%	50%	75%	max
kcal	537.43	37.88	261	523.5	540.5	557.25	670

Видно, что присутствуют 2 наблюдения со сравнительно низкой калорийностью, и 1 со сравнительно высокой. Значения калорийности для всего остального шоколада в выборке лежат в интервале от 470 ккал/100 г. до 593 ккал/100 г.

- **Начинка.**

  Для учитывания начинки шоколада мы выделили 5 категорий начинок:

  1. Орехи
  2. Ягоды и сухофрукты
  3. Снеки (печенье, хлопья, вафли и т.д.)
  4. Десерт (йогурт, помадка, тирамису и т.д.)
  5. Экзотическая начинка (лаванда, роза, мёд, матча, кунжут и т.д.)

  Далее для каждой из 5 категорий была создана дамми-переменная, после чего при помощи подключения к API ChatGPT был произведен парсинг названий каждой из шоколадных плиток и проставление единиц для дамми-переменных, соответствующих встретившимся категориям начинок.

  В результате:

  – 44.17% всех шоколадных плиток в выборке содержат орехи в начинке.
  – 14.11% всех шоколадных плиток в выборке содержат ягоды или сухофрукты в начинке.
  – 13.5% всех шоколадных плиток в выборке содержат снеки в начинке.
  – 22.7% всех шоколадных плиток в выборке содержат десертную начинку.
  – 16.56% всех шоколадных плиток в выборке содержат нестандартную начинку.

- **Доля содержания какао.**

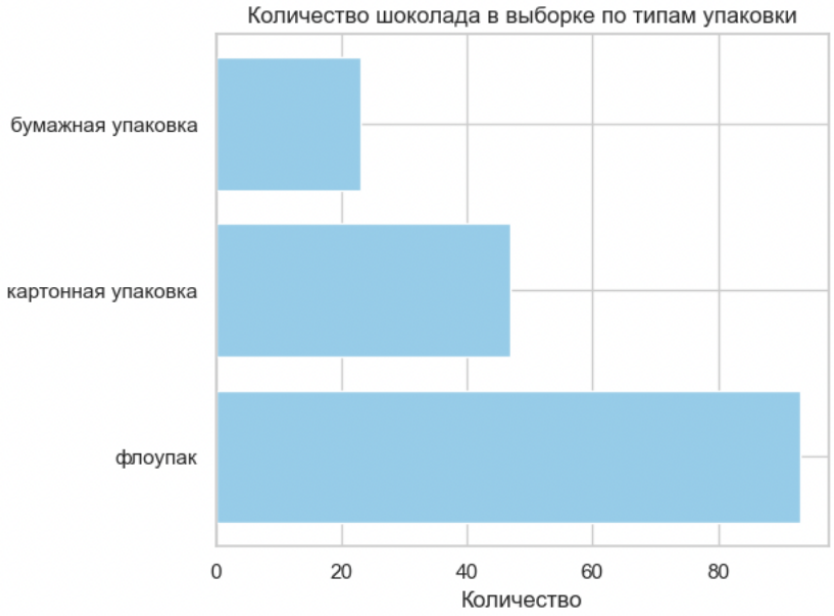
  Значение данной переменной было пропущено для 22 наблюдений. Для некоторых шоколадок из списка удалось найти процент содержания какао на сторонних ресурсах. Для остальных же мы определили процент содержания какао в соответствии с ГОСТом для каждого из типов шоколада: белый – 20%, молочный – 25%, темный – 40%, горький – 55%.

Описательные статистики:

	mean	std	min	25%	50%	75%	max
cocoa	36.4	15.8	5	25	30	42	85

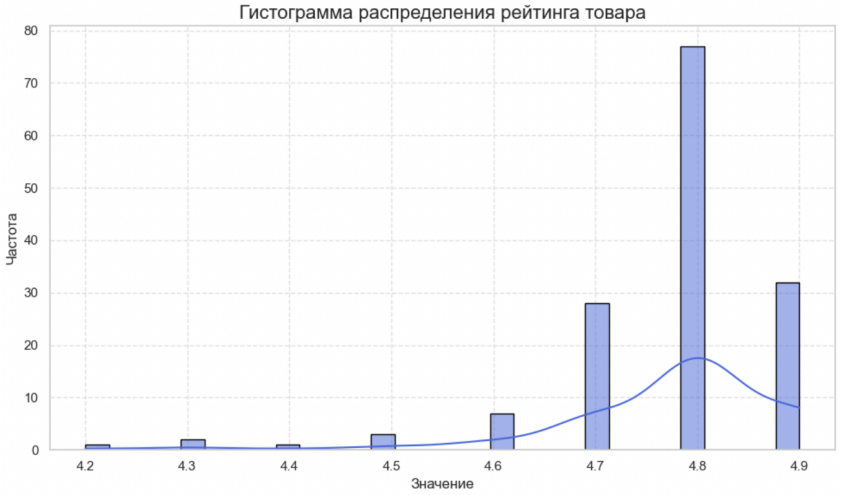
Видно, что в выборке преобладает молочный шоколад. Также присутствует тяжелый правый хвост распределения – это темный и горький шоколад.

- **Упаковка.**



- **Рейтинг.**

Для 12 наблюдений из выборки значение данного признака было пропущено. Посмотрим на гистограмму распределения рейтинга товара:



Видно, что самым часто встречающимся значением с большим отрывом является 4.8. Заполним им пропущенные значения.

Описательные статистики:

	mean	std	min	25%	50%	75%	max
rating	4.77	0.12	4.2	4.7	4.8	4.8	4.9

Видно, что почти половина выборки имеет один и тот же рейтинг 4.8. Есть группы товаров с рейтингом чуть ниже и чуть выше. Также есть единичные наблюдения с совсем низкими рейтингами.

- **Наличие пищевых добавок.**

  44.17% шоколадных плиток в выборке содержат в своем составе пищевые добавки, маркированные буквой Е (Е163, Е319, Е322, Е330, Е331, Е471, Е476, Е503, Е524).

- **Отсутствие сахара.**

  2.45% шоколадок в выборке не содержат в своем составе сахар.

- **Преобладание сахара.**

  У 84.66% шоколадок из выборки в составе преобладает сахар (идет первым в составе).

- **Пальмовое масло.**

  19.02% шоколадок в выборке содержат в своем составе пальмовое масло.

Важно упомянуть про учитывание премии за бренд. Нами было принято решение не моделировать бренд путем проставления дамми-переменных на его название либо же путем группировки брендов по «люксовости». На наш взгляд, специфика шоколада как товара заключается в том, что шоколад выбирается потребителем не из-за бренда, так как шоколад из магазина «Магнит» не принято считать символом статуса, а из-за вкуса и чистоты состава. Поэтому мы решили сфокусироваться именно на отборе признаков, отражающих качество и вкус шоколада, созданных на основе информации о составе.

2 Построение эконометрической модели.

Для построения эконометрической модели ценообразования товара необходимо произвести анализ, состоящий из нескольких этапов.

Для того, чтобы оценки, полученные с помощью метода наименьших квадратов, были несмещенными, а также эффективными в классе всех несмещенных линейных оценок, необходимо, чтобы были выполнены следующие предпосылки теоремы Гаусса-Маркова:

1. Модель линейна по параметрам и корректно специфицирована
2. Наблюдения $\{(x_i, y_i),\ i = 1, \ldots, n\}$ независимы и одинаково распределены. В случае детерминированных регрессоров требуется, чтобы матрица X имела полный ранг.
3. Математическое ожидание случайных ошибок равно нулю
$$\mathbb{E}[\varepsilon_i] = 0.$$
4. Дисперсия случайной ошибки одинакова для всех наблюдений
$$\mathrm{var}(\varepsilon_i) = \sigma^2.$$
5. Случайные ошибки, относящиеся к разным наблюдениям, взаимно независимы.

Важной, но необязательной предпосылкой является также предположение о нормальности остатков. Данная предпосылка позволяет тестировать гипотезы о параметрах модели и строить доверительные интервалы, поэтому ее она будет проверена далее в исследовании.

Последовательно будем проверять каждую из обозначенных предпосылок.

2.1 Проверка наличия выбросов в данных.

Данный этап важен для оценивания параметров модели с помощью метода наименьших квадратов, так как данный метод неустойчив к выбросам.

2.1.1 Методология.

Для проверки на наличие выбросов были использованы следующие критерии:

Стьюдентизированный остаток — это остаток, деленный на свое стандартное отклонение при условии исключения данного наблюдения, т.е.

$$e'_i = \frac{e_i}{S(i)\sqrt{1-h_i}}$$

где:

- e_i — остаток для конкретного наблюдения, полученный по уравнению регрессии, построенному с учетом всех наблюдений;
- $S(i)$ — стандартное отклонение остатков, полученное по уравнению регрессии, построенному по тому же набору наблюдений, но без учета наблюдения i;
- h_i — это диагональный элемент матрицы проектора $X(X'X)^{-1}X'$.

Решающее правило:

Если $e'_i \notin [-t_{crit}, t_{crit}]$, $t_{crit} = t_{1-\alpha,n-k}$, где n - количество наблюдений, k - количество регрессоров с учетом константы $\Rightarrow$ выброс

DFFITS

$$DFFIT_i = \hat{Y}_i - \widehat{Y_{i(i)}},$$

где $\hat{Y}_i$ и $\widehat{Y_{i(i)}}$ – это предсказанные по модели значения с учетом и без учета наблюдения i.

$$DFFITS_i = e'_i\sqrt{\frac{h_i}{1-h_i}},$$

где e'_i – стьюдентизированный остаток.
Если $DFFITS_i > 2\sqrt{\frac{k}{n}}$, то i-е наблюдение может быть выбросом,
где k – это число регрессоров, а n – это число наблюдений.

С помощью данных методов были выделены 12 аномальных наблюдений на уровне значимости $\alpha = 0.05$, которые были удалены из выборки.

Ввиду того, что все шоколадки без добавления сахара были классифицированы как выбросы, после удаления аномальных наблюдений из выборки на месте переменной «no_sugar» остался нулевой столбец, поэтому также была удалена и данная переменная.

2.2 Проверка гипотезы о нормальности остатков.

2.2.1 Методология.

Для проверки гипотезы о нормальности распределения остатков были использованы следующие критерии:

Тест Харке-Бера
$H_0 : S = 0, K = 3.$
$H_1 : S \neq 0, K \neq 3.$
Тест выглядит следующим образом:

$$JB = n\left(\frac{S^2}{6} + \frac{(K-3)^2}{24}\right),$$

где $S = \frac{\sum e_i^3}{n\hat{\sigma}_{ML}^3}$, $K = \frac{\sum e_i^4}{n\hat{\sigma}_{ML}^4}$, e_i - остатки модели,
n - количество наблюдений,
$\hat{\sigma}_{ML}^2 = \frac{\sum e_i^2}{n}$
$\hat{\sigma}_{ML}^2$ - оценки полученные методом максимального правдоподобия.

Используемая статистика:

$$JB = n\left(\frac{S^2}{6} + \frac{(K-3)^2}{24}\right) \sim \chi_2^2,$$

Тест Колмогорова-Смирнова:

$$D_n = \sup_x |F_n(x) - F(x)|,$$

где $F_n(x)$ – эмпирическая функция распределения выборки,
$F(x)$ – предполагаемая теоретическая функция распределения,
$\sup_x$ – супремум по всем возможным значениям x.
При $n \to \infty$ статистика $\sqrt{n}D_n$ имеет распределение Колмогорова:

$$\lim_{n\to\infty} P(\sqrt{n}D_n \leqslant z) = K(z) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2z^2}$$

Тест	**P-value**	**Вывод**
Jarque-Bera Test	0.072	Гипотеза о нормальности остатков не отвергается при $\alpha = 0.05$
Kolmogorov-Smirnov Test	0.301	Гипотеза о нормальности остатков не отвергается при $\alpha = 0.05$

Таблица 1: Результаты тестов на нормальность распределения остатков

2.3 Проверка корректной спецификации.

2.3.1 Выбор функциональной формы.

В контексте задачи выбора корректной спецификации возникает вопрос о выборе функциональной формы среди трех альтернатив.

$$\begin{aligned}\text{Price}_i &= X_i + \varepsilon_i \\ \log(\text{Price}_i) &= X_i + \varepsilon_i \\ \log(\text{Price}_i) &= \log(X)_i + \varepsilon_i\end{aligned}$$

Однако, третья спецификация не применима для нашей задачи, так как подавляющее большое количество объясняющих факторов - дамми переменные, следовательно, что делает невозможным взятие логарифма от признаков.

Сравним линейную и полулогарифмическую модели с помощью PE-теста:

Тест проверяет, остаётся ли в остатках одной функциональной формы модели информация, которую можно было бы объяснить с помощью альтернативной функциональной формы.

PE тест МакКиннона, Уайта и Дэвидсона

Шаг 1. Вычисляют

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X \qquad \ln\hat{Y} = \hat{\alpha} + \hat{\beta}X \tag{1}$$

Шаг 2. Оценивают вспомогательные регрессии

$$\ln Y = \alpha + \beta X + \delta_{LOG}\left(\hat{Y} - \exp(\widehat{\ln Y})\right) + \varepsilon \tag{2}$$

$$Y = \alpha + \beta X + \delta_{LIN}\left(\widehat{\ln Y} - \ln \hat{Y}\right) + \varepsilon \tag{3}$$

Шаг 3. Проверяют в них гипотезы

$$H_0: \quad \delta_{LOG} = 0 \quad \text{и} \quad H_0': \quad \delta_{LIN} = 0 \tag{4}$$

Спецификация	**Коэффициент**	**Станд. ошибка**	**t-статистика**	**p-значение**	**Вывод**
Логарифмическая	0.005	0.002	3.25	0.001	Отвергаем H_0 при $\alpha = 0.05$
Линейная	-0.222	0.193	-1.148	0.252	Не отвергаем H_0 при $\alpha = 0.05$

Таблица 2: Результаты тестов на выбор спецификации модели

В результате проведения теста была выбрана линейная спецификация.

2.3.2 Проверка спецификации модели.

После выбора функциональной формы необходимо проверить верную спецификацию модели. На данном этапе стоит задуматься о содержательных теоретических соображениях, которые могут иметь место при формировании цены на шоколад.

Например, рост доли какао может положительно влиять на цену до определенного порога, после которого дальнейшее увеличение доли какао делает шоколад слишком горьким для массового рынка, что может приводить к снижению цены.

Для проверки спецификации модели воспользуемся тестом Рамсея:

Теста Рамсея

1. Оцениваем коэффициенты функции регрессии (*)

$$\hat{Y} = \hat{\alpha}_1 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_1 X_k$$

2. Сохраняем столбец оцененных значений $\hat{Y}$
3. Оцениваем коэффициенты вспомогательной регрессии (**)

$$Y = \alpha + \beta_1 X_1 + \ldots + \beta_k X_k + \alpha_2 \hat{Y}^2 + \varepsilon$$

4. Тогда проверка гипотезы о правильной спецификации равносильна проверке гипотезы (коэффициенты при $\hat{Y}$):

$$H_0 : \alpha_2 = 0$$
$$H_1 : \alpha_2 \neq 0$$

5. Вычисляем значение тестовой статистики

$$\boxed{F = \frac{(RSS_R - RSS_{UR})/(m-1)}{RSS_{UR}/(n-(k+m))}}$$

где RSS_R - это сумма квадратов остатков модели (*),
а RSS_{UR} - это сумма квадратов остатков модели (**)

В результате проведения теста был получен p-value = 0.495, в связи с чем нулевая гипотеза о правильной спецификации модели не отвергается на уровне значимости $\alpha = 0.05$.

2.4 Проверка мультиколлинеарности.

Проверим наличие мультиколлинеарности в модели при помощи VIF:

VIF – variance inflation factor, превышает 10

$$\mathbf{VIF}(X_j) = \frac{1}{1 - R_j^2}$$

где R_j^2 – коэффициент множественной детерминации регрессора X_j на все остальные регрессоры.

```
VIF для регрессора is_sugar_first: 4.5501
VIF для регрессора palm: 1.5976
VIF для регрессора weight: 2.2540
VIF для регрессора rus_brand: 2.1960
VIF для регрессора rating: 1.2902
VIF для регрессора kcal: 1.4832
VIF для регрессора cocoa: 8.5778
VIF для регрессора filling_NUTS: 1.2524
VIF для регрессора filling_BERRIES: 1.1367
VIF для регрессора filling_SNACKS: 1.4455
VIF для регрессора filling_DESSERT: 1.3709
VIF для регрессора filling_EXOTIC: 1.2674
VIF для регрессора food_additives: 1.6195
VIF для регрессора type_Горький: 8.8548
VIF для регрессора type_Молочный: 3.9269
VIF для регрессора type_Темный: 6.1622
VIF для регрессора package_Бумага: 1.8718
VIF для регрессора package_Флоупак: 2.1641
```

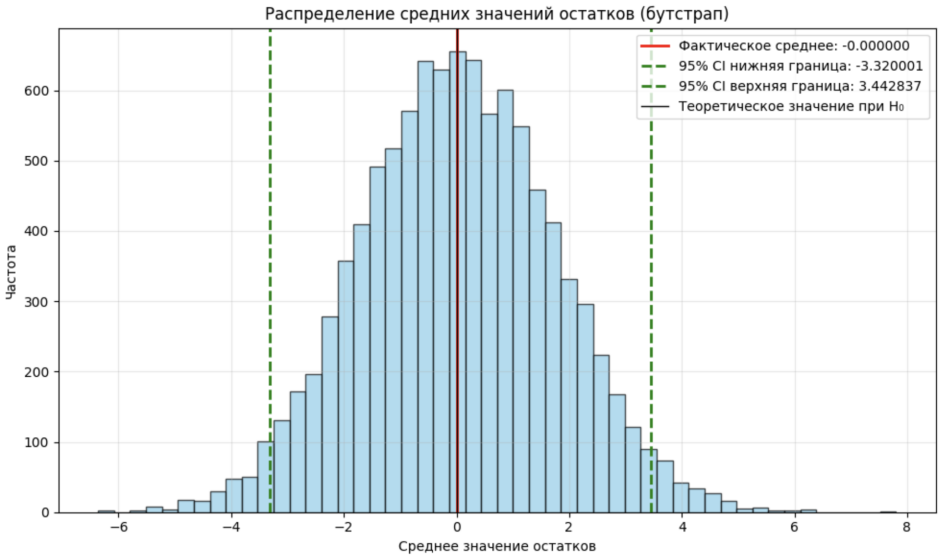
В результате расчета VIF не было выявлено факторов с VIF > 10, поэтому проблема квазимультиколлинеарности в данной задаче отсутствует.

2.5 Проверка гипотезы о несистематичности случайной ошибки.

С помощью процедуры бутстрапирования построим выборочное распределение среднего значения остатков модели и протестируем гипотезу о равенстве математического ожидания ошибки нулю с помощью t-статистики.

$$H_0 : E(\varepsilon) = 0 \quad H_1 : E(\varepsilon) \neq 0$$

В результате проверки гипотезы был получен p-value = 1.00, поэтому нулевая гипотеза о несистематичности случайной ошибки не отвергается на уровне значимости $\alpha = 0.05$.



2.6 Проверка на гетероскедастичность.

Проверим гипотезу о гомоскедастичности ошибки при помощи визуальной оценки графика «Прогнозы - остатки» и тестов Бройша-Пагана и Уайта:

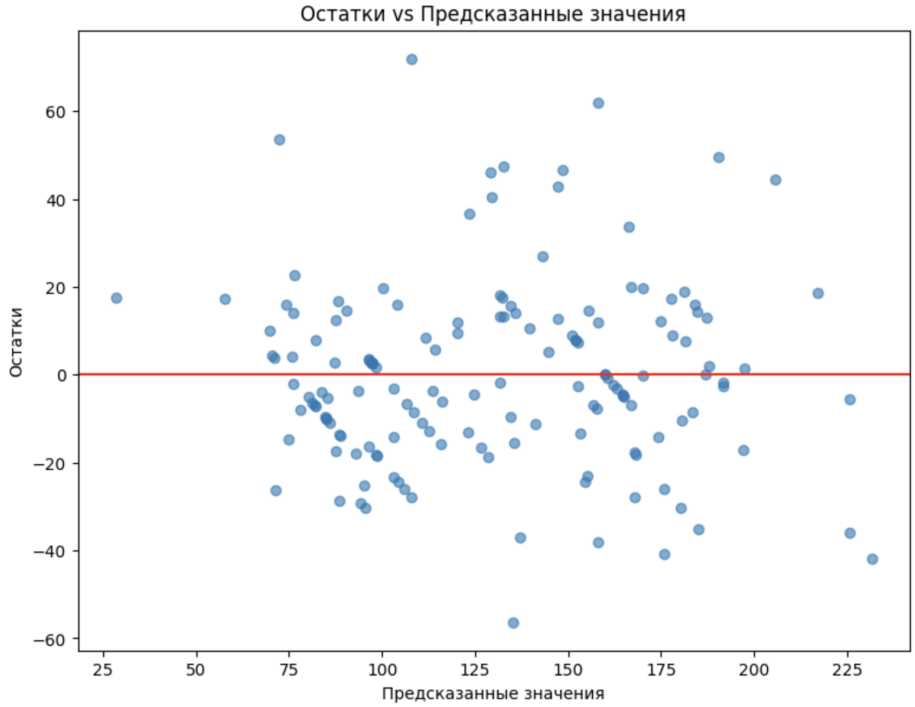


График «Прогнозы - остатки» напоминает бесформенное облако точек, в связи с чем нет оснований полагать, что имеет место гетероскедастичность.

Тест Уайта

$$H_0 : \sigma_i^2 = \sigma^2 \ \forall i$$
$$H_1 : \exists i,j : \sigma_i^2 \neq \sigma_j^2$$
1. Оценка регрессии $Y = \alpha + \beta_1 X_1 + \cdots + \beta_3 X_k + \varepsilon \Rightarrow e_i$
2. $e^2 = \alpha_1 + \sum_{l=2}^{k} \alpha_l X_l + \sum_{l=2}^{k} \beta_{l2} X_l^2 + \sum_{l,j=2}^{k} \gamma_{lj} X_l X_j + u \Rightarrow R^2$

$$nR^2 \overset{H_0}{\sim} \chi^2_{m-1}$$

Тест Бройша - Пагана

$$Y = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$
$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n}$$
$$\frac{e^2}{\hat{\sigma}^2} = \gamma_0 + \gamma_1 Z_1 + \cdots + \gamma_p Z_p + u \Rightarrow ESS$$
$$ESS \overset{h_0}{\sim} \chi_p^2$$

Тест	Статистика	p-value	Вывод
Тест Бройша-Пагана	14.73	0.68	Гипотеза о гомоскедастичности не отвергается при $\alpha = 0.05$
Тест Уайта	150.918	0.418	Гипотеза о гомоскедастичности не отвергается при $\alpha = 0.05$

Таблица 3: Результаты тестов на гетероскедастичность

2.7 Оценка финальной модели.

Была оценена модель с следующей спецификацией:

$$\begin{aligned}\text{Price}_i =& \beta_0 + \beta_1 \cdot \text{package[Т.Бумага]}_i + \beta_2 \cdot \text{package[Т.Флоупак]}_i \\ &+ \beta_3 \cdot \text{type[Т.Горький]}_i + \beta_4 \cdot \text{type[Т.Молочный]}_i + \beta_5 \cdot \text{type[Т.Темный]}_i \\ &+ \beta_6 \cdot \text{weight}_i + \beta_7 \cdot \text{is\_sugar\_first}_i + \beta_8 \cdot \text{palm}_i \\ &+ \beta_9 \cdot \text{kcal}_i + \beta_{10} \cdot \text{rating}_i + \beta_{11} \cdot \text{cocoa}_i \\ &+ \beta_{12} \cdot \text{rus\_brand}_i + \beta_{13} \cdot \text{filling\_NUTS}_i + \beta_{14} \cdot \text{filling\_BERRIES}_i \\ &+ \beta_{15} \cdot \text{filling\_SNACKS}_i + \beta_{16} \cdot \text{filling\_DESSERT}_i + \beta_{17} \cdot \text{filling\_EXOTIC}_i \\ &+ \beta_{18} \cdot \text{food\_additives}_i + \varepsilon_i\end{aligned}$$

Таблица 4: Результаты регрессионного анализа (OLS, зависимая переменная: price)

Переменная	Коэффициент	Станд. ошибка	t-статистика	p-value
Intercept	-406.6894	103.243	-3.939	0.000
package[Т.Бумага]	10.3634	6.974	1.486	0.140
package[Т.Флоупак]	-18.4343	5.477	-3.366	0.001
type[Т.Горький]	26.2121	19.431	1.349	0.180
type[Т.Молочный]	-0.5631	7.605	-0.074	0.941
type[Т.Темный]	-16.9593	11.866	-1.429	0.155
weight	2.4960	0.306	8.164	0.000
is_sugar_first	40.6920	11.780	3.454	0.001
palm	-16.9872	5.802	-2.928	0.004
kcal	0.1989	0.066	3.009	0.003
rating	31.5411	18.406	1.714	0.089
cocoa	1.3670	0.363	3.766	0.000
rus_brand	-4.9402	5.739	-0.861	0.391
filling_NUTS	-4.5765	4.132	-1.107	0.270
filling_BERRIES	5.2056	5.644	0.922	0.358
filling_SNACKS	-8.6950	6.496	-1.338	0.183
filling_DESSERT	-0.5539	5.033	-0.110	0.913
filling_EXOTIC	2.0551	5.548	0.370	0.712
food_additives	-11.7014	4.679	-2.501	0.014

Число наблюдений: 151, Степени свободы: 132

R^2 оцененой модели равен 0.792

По итогам оценивания параметров модели статистически значимыми на уровне $\alpha = 0.05$ оказались следующие факторы:

- **Флоупак** ($\beta_2 = -18.4343$, p-value $= 0.001$): При прочих равных использование флоупак-упаковки снижает цену шоколада на 18.43 рубля по сравнению с базовой категорией упаковки.

Это может быть связано с восприятием флоупака как менее премиального варианта и более низкой себестоимостью такой упаковки.

- **Вес** ($\beta_6 = 2.4960$, p-value $= 0.000$): При прочих равных увеличение веса шоколада на 1 грамм приводит к увеличению цены на 2.50 рубля. Данный эффект вполне очевиден, так как самого шоколада банально становится больше.
- **Преобладание сахара** ($\beta_7 = 40.6920$, p-value $= 0.001$): При прочих равных, если сахар указан первым ингредиентом, цена шоколада выше на 40.69 рублей.
- **Пальмовое масло** ($\beta_8 = -16.9872$, p-value $= 0.004$): При прочих равных наличие пальмового масла в составе снижает цену шоколада на 16.99 рублей. Это может быть связано с более низкой себестоимостью и негативным восприятием данного компонента потребителями.
- **Калорийность** ($\beta_9 = 0.1989$, p-value $= 0.003$): При прочих равных увеличение калорийности на 1 ккал приводит к увеличению цены шоколада на 0.20 рубля.
- **Доля какао**($\beta_{11} = 1.3670$, p-value $= 0.000$): При прочих равных увеличение процента содержания какао на 1 увеличивает цену шоколадной плитки на 1.37 рубля. Высокий процент какао, например, в молочном шоколаде, часто является индикатором качества.
- **Пищевые добавки** ($\beta_{18} = -11.7014$, p-value $= 0.014$): При прочих равных наличие пищевых добавок в составе снижает цену шоколада на 11.70 рубля. Это может быть связано с более низкой себестоимостью и негативным восприятием данных компонентов потребителями.

Проверим гипотезу о значимости модели в целом:

$$H_0 : \forall j \in \{1, 2, ..., 18\} : \beta_j = 0$$

$$H_1 : \exists j \in \{1, 2, ..., 18\} : \beta_j \neq 0$$

$$F = \frac{\frac{ESS}{k-1}}{\frac{RSS}{n-k}} \sim F_{k-1,n-k}$$

где k – количество регрессоров с константой.
Как и ранее, сравниваем наблюдаемое и критическое значение статистики.
Если $F_{\text{набл.}} > F_{k,n-k}$, то H_0 отвергается.

В результате проведения теста был получен p-value $= 2.16 \times 10^{-16}$, следовательно, гипотеза о незначимости модели в целом отвергается при уровне значимости $\alpha = 0.05$.

3 Альтернативная модель.

В рамках исследования важно учитывать, что в ассортименте магазина «Магнит» шоколадные плитки сегментированы по ценовым категориям, и для каждой из этих категорий ключевые факторы ценообразования могут существенно различаться. Например, в бюджетном сегменте ключевым фактором может выступать рейтинг товара, отражающий качество шоколада в терминах соотношения цены и качества, тогда как чистота состава не играет значимой роли. В премиальном сегменте акцент смещается: на первый план выходит отсутствие в составе синтетических пищевых добавок и вредных жиров, при этом фактор соотношения цены-качества перестает иметь большое значение. Таким образом, указанные в спецификации факторы по-разному будут влиять на цену шоколада в разных ценовых диапазонах, в то время, как обычная линейная регрессия, оцененная методом МНК, способна моделировать лишь средние по всей выборке эффекты.

Анализ распределения цен на шоколад выявил три выраженных сегмента, соответствующих разным ценовым категориям. Распределение демонстрирует асимметрию с левосторонним смещением, что указывает на преобладание товаров в нижнем ценовом диапазоне. Кроме того, статистические тесты подтверждают наличие около 7% выбросов. Эти особенности — сегментация значений, асимметрия и аномальные значения — делают результаты оценивания модели при помощи МНК нерелевантными. В связи с этим для корректного моделирования целесообразно рассмотреть альтернативные методы, устойчивые к неоднородности распределения и наличию выбросов.

Альтернативным методом, подходящим для решения задачи, является квантильная регрессия. Квантильная регрессия позволяет моделировать зависимость целевой переменной от признаков для различных уровней значений целевой переменной. В нашей задаче данный метод поможет отдельно оценить влияние факторов на цену в каждом из ценовых сегментов, что позволит учесть гетерогенность влияния признаков на целевую переменную и получить информативные предельные эффекты, используя при этом те 7% выборки, что при оценивании модели методом МНК были удалены как выбросы.

Будем оценивать квантильные регрессии для квантилей 0.2, 0.5, 0.9. Результаты оценивания квантильных регрессий приведены ниже.

Оценки коэффициентов:

	Intercept	Бумага	Флоупак	Горький	Молочный	Темный	weight	sugar first
q=0.2	-479.01	10.46	-27.55	3.66	-3.17	-14.75	1.56	26.48
q=0.5	-398.14	15.48	-18.13	14.87	-6.09	-15.16	2.72	27.84
q=0.9	-546.03	15.37	-31.78	126.49	0.56	8.96	3.88	56.50

	kcal	rating	palm	rus brand	no sugar	cocoa	NUTS	BERRIES
q=0.2	0.12	74.62	-3.87	-12.95	18.42	1.29	-0.92	4.86
q=0.5	0.18	30.12	-16.00	-11.54	65.13	1.31	-0.33	7.32
q=0.9	0.06	66.05	-26.99	26.61	68.97	-0.43	-11.42	13.68

	SNACKS	DESSERT	EXOTIC	food additives
q=0.2	-7.81	-4.11	3.86	-11.36
q=0.5	-7.85	1.82	7.71	-5.61
q=0.9	-16.06	-17.01	-8.47	-25.79

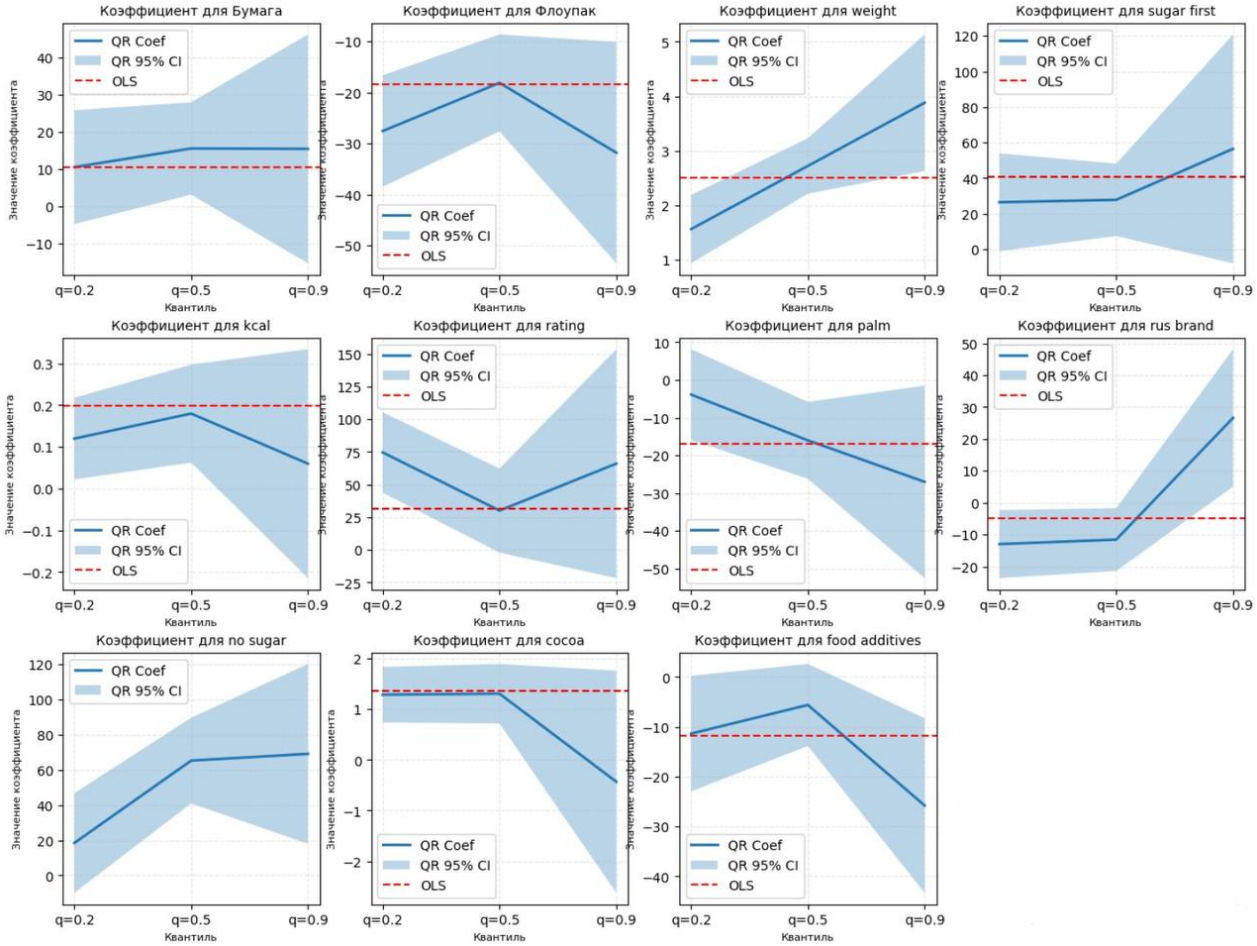
p-value коэффициентов:

	Intercept	Бумага	Флоупак	Горький	Молочный	Темный	weight	sugar first
q=0.2	0.00	0.18	0.00	0.83	0.67	0.18	0.00	0.06
q=0.5	0.00	0.02	0.00	0.37	0.36	0.13	0.00	0.01
q=0.9	0.02	0.33	0.00	0.06	0.98	0.78	0.00	0.09

	kcal	rating	palm	rus brand	no sugar	cocoa	NUTS	BERRIES
q=0.2	0.03	0.00	0.53	0.02	0.20	0.00	0.83	0.40
q=0.5	0.00	0.07	0.00	0.02	0.00	0.00	0.93	0.15
q=0.9	0.66	0.14	0.04	0.02	0.01	0.70	0.19	0.17

	SNACKS	DESSERT	EXOTIC	food additives
q=0.2	0.21	0.39	0.53	0.06
q=0.5	0.17	0.69	0.12	0.18
q=0.9	0.26	0.11	0.54	0.00

Визуализируем значения коэффициентов на рассматриваемых квантилях для факторов, значимых хотя бы на одном из квантилей на уровне значимости $\alpha = 0.05$, вместе с их 95% доверительными интервалами и их МНК-оценками из предыдущего раздела.



Проинтерпретируем значимые при $\alpha = 0.05$ значения коэффициентов:

- **Бумажная упаковка.** Значимость только на q = 0.5 (p-value = 0.02). Эффект: +15.48 рублей в среднем сегменте (q = 0.5). Это может быть связано с тем, что в среднем ценовом диапазоне бумажная упаковка добавляет премию, возможно, из-за восприятия её как более экологичной.
- **Флоупак.** Значимость на всех квантилях (p-value = 0.00). Эффект: снижение цены на 27.55 руб. (q=0.2), 18.13 руб. (q=0.5) и 31.78 руб. (q=0.9). Наибольшее влияние наблюдается в премиальном сегменте, что свидетельствует о неприемлемости такой упаковки для дорогого шоколада.
- **Вес плитки.** Значимость на всех квантилях (p-value = 0.00). Эффект: увеличение цены на 1.56 руб./г (q=0.2), 2.72 руб./г (q=0.5) и 3.88 руб./г (q=0.9). Эффект усиливается в премиальном сегменте, что может быть связано с большей себестоимостью грамма качественного шоколада.

- **Калорийность.** Значимость на q = 0.2 (p-value = 0.03) и q = 0.5 (p-value = 0.00). Эффект: +0.12 руб./ккал (q=0.2) и +0.18 руб./ккал (q=0.5). Данному эффекту трудно дать содержательную интерпретацию.
- **Пальмовое масло.** Значимость на q = 0.5 (p-value = 0.00) и q = 0.9 (p-value = 0.04). Эффект: снижение цены на 16.00 руб. (q=0.5) и 26.99 руб. (q=0.9). Наибольший негативный эффект в премиальном сегменте, где потребители избегают «дешёвых» заменителей.
- **Отсутствие сахара.** Значимость на q = 0.5 (p-value = 0.00) и q = 0.9 (p-value = 0.01). Эффект: +65.13 руб. (q=0.5) и +68.97 руб. (q=0.9). Сильная премия за «здоровый» состав в среднем и премиальном сегментах.
- **Доля какао.** Значимость на q = 0.2 (p-value = 0.00) и q = 0.5 (p-value = 0.00). Эффект: +1.29 руб. (q=0.2) и +1.31 руб. (q=0.5). Важна для бюджетного и среднего сегментов, но незначима для премиального (p=0.7).
- **Российский бренд.** Значимость на всех квантилях (p-value = 0.02). Эффект: -12.95 руб. (q=0.2), -11.54 руб. (q=0.5) и +26.61 руб. (q=0.9). В премиальном сегменте российские бренды смогли создать положительное позиционирование (например, премиум-линейка «Алёнки» или «Bucheron»).
- **Рейтинг.** Значимость на q = 0.2 (p-value = 0.00). Эффект: +7.46 руб при увеличении рейтинга товара на 0.1. В бюджетном сегменте высокий рейтинг является ключевым драйвером цены, возможно, как индикатор соотношения цены и качества.
- **Пищевые добавки.** Значимость на q = 0.9 (p-value = 0.00). Эффект: снижение цены на 25.79 руб. в премиальном сегменте, где ожидается «чистый» состав продукта.
- **Преобладание сахара.** Значимость на q = 0.5 (p-value = 0.01). Эффект: увеличение цены на 27.84 руб. Парадоксальный результат, которому трудно дать содержательную интерпретацию.

4 Построение оценки цены на собственный товар.

Особый интерес представляет прогнозирование стоимости товаров с уникальными комбинациями признаков, аналогов которых нет в ассортименте ретейлера.

Будем оценивать шоколад швейцарского бренда «Villars» (бренд отсуствует на сайте магазина «Магнит») со следующими характеристиками:

Таблица 5: Закодированные признаки для шоколада Villars

Признак	**Значение**
Вес	100
Российский бренд	0
Тип шоколада	Горький
Калорийность	540
Наличие орехов в начинке	0
Наличие ягод или сухофруктов в начинке	0
Наличие снеков в начинке	0
Наличие десертной начинки	0
Наличие экзотической начинки	0
Доля содержания какао	0.72
Упаковка	Картонная упаковка
Рейтинг	4.8 (закодируем пропущенное значение)
Наличие пищевых добавок	0
Отсутствие сахара	0
Преобладание сахара	0
Пальмовое масло	0

Таблица 6: Прогноз цены шоколада Villars

Модель	**Прогнозная цена (руб.)**
q = 0.2	196.50
q = 0.5	224.83
q = 0.9	286.94
Линейная регрессия	220.44

5 Заключение.

В рамках работы была проведена комплексная оценка гедонистической ценовой функции с использованием методов множественной и квантильной регрессии. Исследование включало сбор и предварительный анализ данных, отбор релевантных признаков, влияющих на ценообразование. Было проверено выполнение условий теоремы Гаусса-Маркова и оценена модель линейной регрессии. После были построены модели квантильной регрессии, позволяющие оценить влияние характеристик на разные части ценового распределения и выявлена неоднородность эффектов переменных в зависимости от ценового сегмента.

Применение двух подходов — классической и квантильной регрессий — позволило получить более полное представление о факторах ценообразования. Если множественная регрессия дала усредненную оценку, то квантильная регрессия выявила, как значимость параметров меняется для бюджетного, среднего и премиального сегментов.