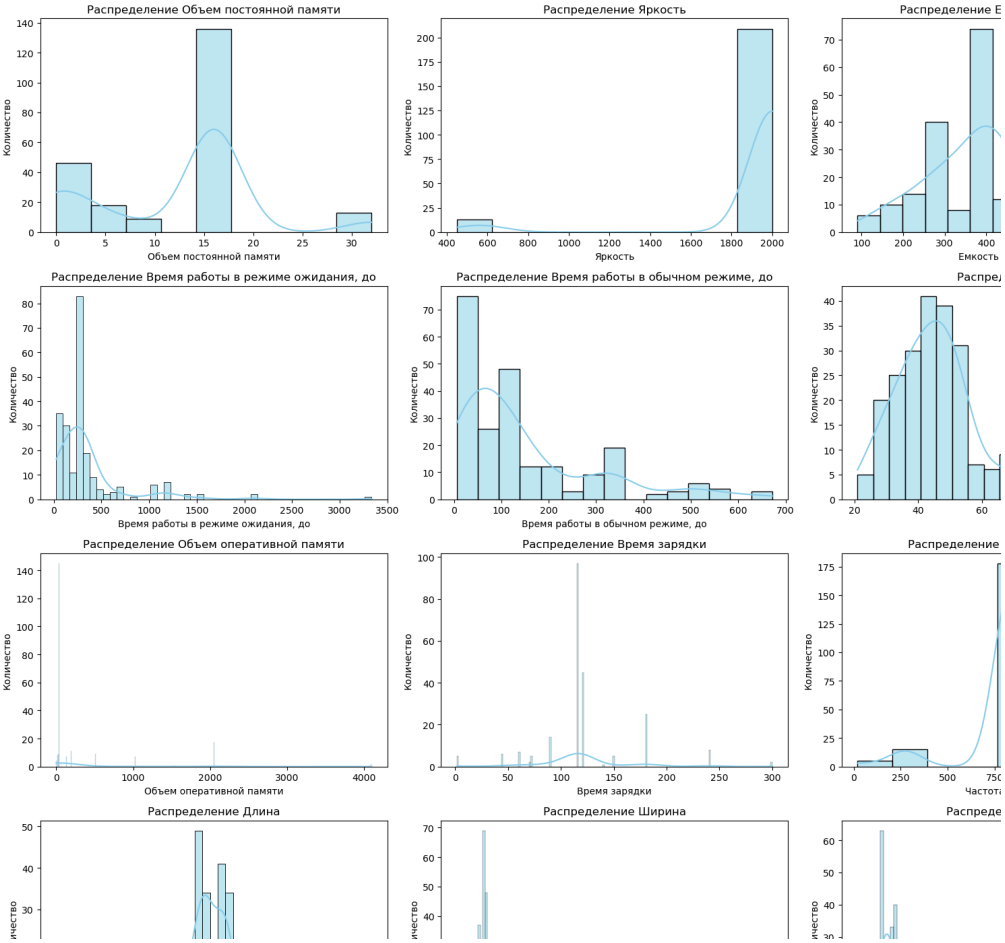
Титульный лист

Гедонистическая ценовой функции для смарт-часов

Проект подготовили
Алексанян Алек - обработка данных; построение, анализ эконометрических моделей; прогноз модели для предложенного товара **(34%)**
Гаджиагаев Максим - построение квантильной регрессии и интерпретация полученных результатов **(33%)**
Литасов Александр - парсинг данных, анализ описательных статистик и графический анализ переменных **(33%)**

Анализ описательных статистик и описание графиков

1. Распределение количественных признаков



- Распределение объема постоянной памяти: Большинство значений сосредоточено около 5 единиц. Вид распределения показывает, что есть и другие значимые пики около 15 и 25 единиц.

- Распределение Яркости: Пик находится в районе 400 единиц яркости, что может быть типичным стандартом яркости экрана. Распределение достаточно сильно скошено вправо, что указывает на наличие устройств с высокими значениями яркости.

- Распределение Емкости аккумулятора: Большая часть значений сосредоточена около 400 единиц емкости. Помимо основного пика, существуют меньшие пики около 200 и 600 единиц, что может отражать наличие разных версий устройств с различной емкостью батарей.

2. Кластеризация с помощью KMeans

- Кластер 0:

а) Наименьшая частота процессора и небольшой объем оперативной памяти.

б) Часы имеют больший размер и чуть большую емкость аккумулятора, чем другие

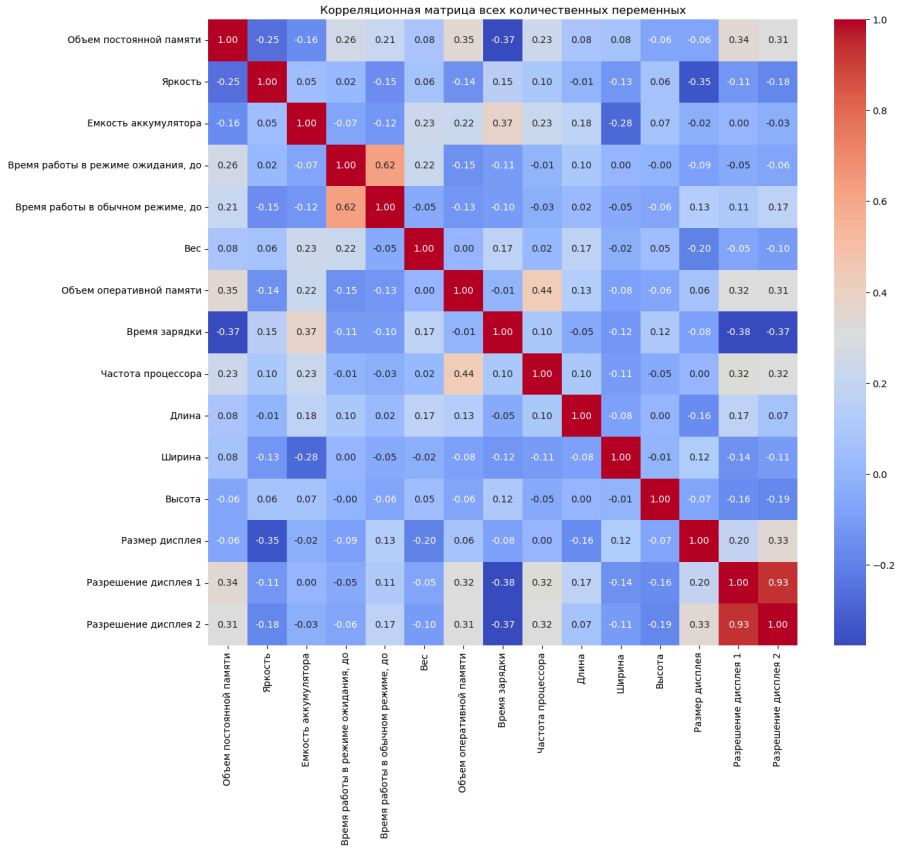
- Кластер 1:

а) Большой объем оперативной памяти, высокое разрешение дисплея говорят о том, что часы из первого кластера более мощные

- Кластер 2:

а) Повышенное время работы в режиме ожидания, сниженный объем оперативной памяти, большая память говорят о том, что часы менее мощные, но более долгоработающие

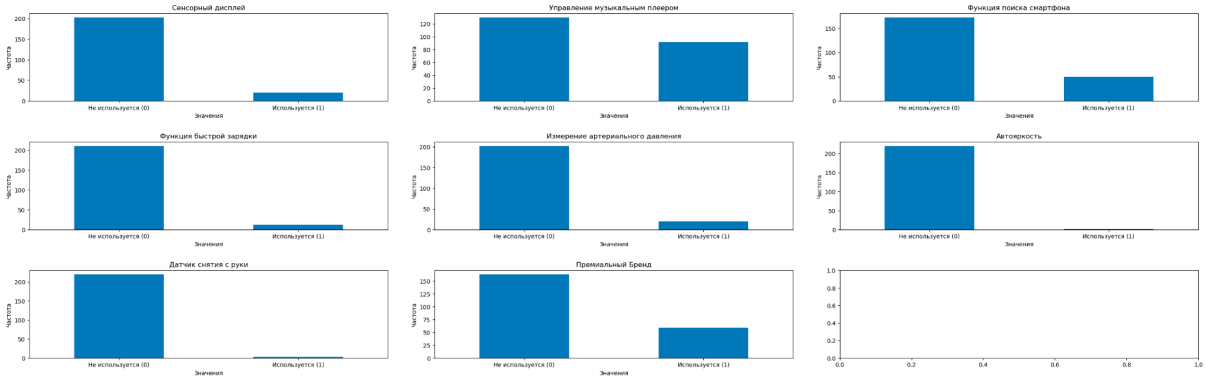
3. Корреляция



Сильные корреляции:

- Объем постоянной памяти и объем оперативной памяти (0.35): Это означает, что устройства с большим объемом постоянной памяти часто оснащаются и большим объемом оперативной памяти.
- Время работы в обычном режиме и время работы в режиме ожидания (0.62): Указывает на то, что устройства с более длительным временем работы в одном из режимов склонны к более длительной работе и в другом.
- Разрешение дисплея 1 и разрешение дисплея 2 (0.93): Предполагает, что данные два параметра фактически отражают одну и ту же характеристику разных способов измерения или отображения.

4. Анализ категориальных признаков



- Почти все смарт-часы оснащены функциями, связанными с мониторингом здоровья (например, измерение артериального давления) и повышением удобства использования (например, быстрая зарядка и управление музыкальным плеером).
- Значительное количество смарт-часов связано с премиальными брендами

Построение эконометрических моделей

Для подбора правильной функциональной формы модели было построено 3 модели: линейная, логарифмическая, линейная в логарифмах. Для выявления лучшей модели сначала было сравнение Полулогарифмическую модель и линейную модель в логарифмах можно сравнить по нормированному $R2$, так как в этих моделях одинаковые зависимые переменные ($\ln Y$).

У полулогарфической модели нормированный R2 больше. Значит линейную в логарифмах модель можно более не рассматривать. Сравнение линейной и полулогарифмической модели происходило через метод Зарембки (частный случай преобразования Бокса-Кокса). Тест показал, что качесто подгонки у моделей разное, и поэтому был выбрана полулогарифмическая модель, так как у нее было меньшее значение RSS.

Для проверки наличия остатков использовалось 2 следующих критерия: DFFITS и стьюдентизированные остатки. В результате выявилось 25 выбросов. Сравнив результаты моделей с выбросами и без выбросов по R2, было решено построить модель на основе данных без выбросов, так как нормированный R2 оказался сильно выше

Для проверки спецификации модели использовался тест Рамсея с 1 вспомогательным регрессоров. В результате получили,

F_obs = 0.777

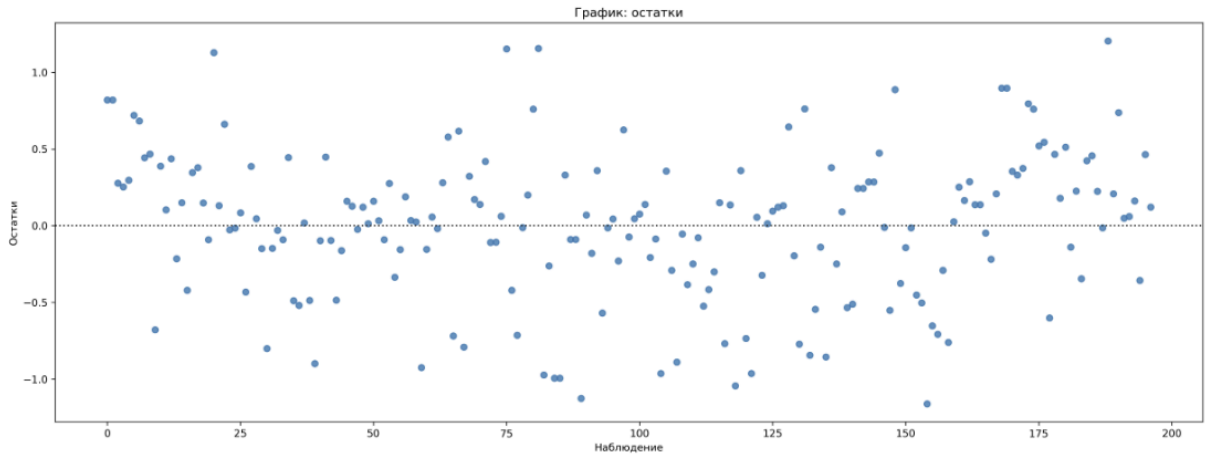
p_value = 0.379

Нулевая гипотеза не отвергается, что говорит о том, что спецификация модели является правильной, то есть еще переменные не нужно добавлять.

Далее мы проверили все предпосылки теоремы Гаусса-Маркова:

1) Систематическая ошибка

Проверка наличия систематической ошибки происходило с помощью график остатков

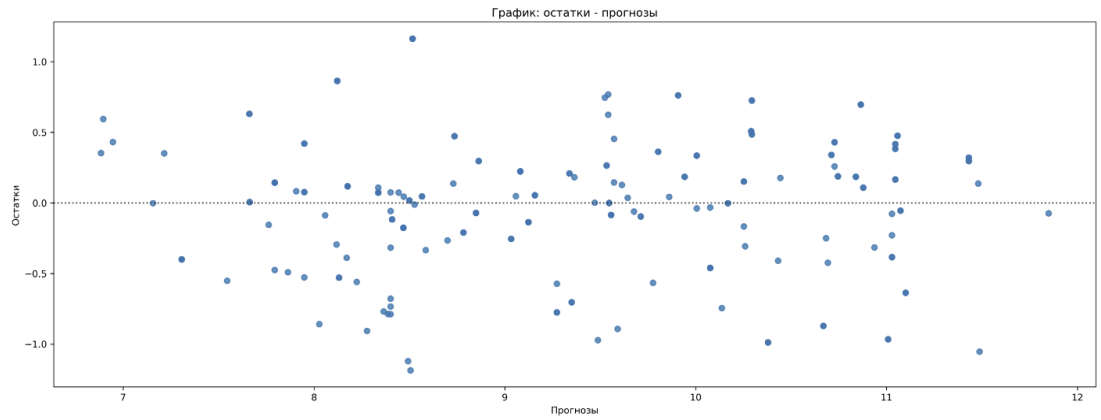


Вывод: Систематической ошибки не наблюдается, так как значения колеблются вокруг нуля без смещения, значит первая предпосылка ТГМ не нарушена

2) Гомоскедастичность

Проверим трёмя способами: с помощью графика остатки-прогнозы, тест Бройша-Пагана, тест Уайта

а) График остатки-прогнозы:



Вывод: Если приглядется, то можно увидеть некоторую взаимосвязь. Чем больше прогнозное значение, тем шире разброс остатков. Кажется есть гетероскедастичность. Причем заметим, что что остатки в среднем равны нулю, значит функциональная форма правильная

б) Тест Бройша-Пагана: При проведении данного теста, было выявлено, что Pvalue=0, что говорит, что есть гетерескедастичность.

в) Тест Уайта: Pvalue при данном тесте так же равен, что финально доказывает, что в данных наблюдается гетерескедастичность.

Далее была попытка избавиться от гетерескедастичности двумя различными способами: доступный обобщённый МНК и использование робастных оценок в форме Уайта. Так мы получили следующий резульатат: Построены 2 модели, которые решают проблему гетероскедастичности. Далее будем использовать модель с робастными оценками в форме Уайта, так как они показывают сильно лучше результат по AIC и BIC

3) Мультиколлинеарность

Для того, чтобы понять, есть ли мультиколлинеарность, предлагаю сначала просто посмотреть на логичность знаков оценённых весов у каждого признака и на корреляционную матрицу факторов. Интуиция подсказывает, что все коэффициенты должны быть положительны. Тем не менее, модель показывает некоторые отрицательные значения (см. код в Python). Поэтому проверим мультиколлинеарность далее с помощью корреляции (см.выше). Как видим, не наблюдается высокая корреляция. Проверим более точно наличие мультиколлинеарности с помощью *VIF*. Для этого оценим несколько моделей, беря по очереди в качестве объясняемой переменной один из количественных переменных. Далее будем считать $Rj2$ этих моделей и с помощью него считать $VIF(Xj)$ по следующей формуле. Если $VIF(Xj) > 10$, то можно будет говорить о наличии мультиколлинеарности. В результате получили, что все количественные переменные, кроме Разрешение_дисплея_2 не показывают мультиколлинеарность. Ввиду того, что мультиколлинеарности почтит не обнаружено, то нет причин применять МГК, так как данный метод в основном используется для борьбы с мультиколлинеарностью. Кроме того, МГК может использоваться просто для уменьшения числа переменных.

4) Эндогенность

Для начала посмотрим на то, какая переменная имеет наивысшую корреляцию с ценой часов, что может говорить о двусторонней связи между ценой и данным параметром.

Так, мы находим, что объём постоянной памяти имеет самую большую корреляцию с ценой (0.37). Возможно, логика заключается в том, что большой объем памяти и делает цену такой высокой. С другой же стороны, завышенная цена позволяет производителю производить смарт часы с большей памятью. Поэтому эндогенность данной регрессора данной переменной может быть обоснована. Далее посмотрим на корреляции объёма постоянной памяти с другими переменными и видим, что есть сильная связь между объёма постоянной памяти и объёма оперативной памяти (корреляция ~ 0.35). Для понимания, можно ли использовать объём оперативной памяти в качестве регрессора посчитает его корреляции с ценой и получаем -0.04. Как видим, оперативная память слабо коррелируют с ценой. В связи с этим, для IV мы можем в качестве инструмента объём оперативной памяти. Для дальнейшего сравнения моделей, сравним OLS без инструментальной переменной, а затем IV с инструментальной переменной.

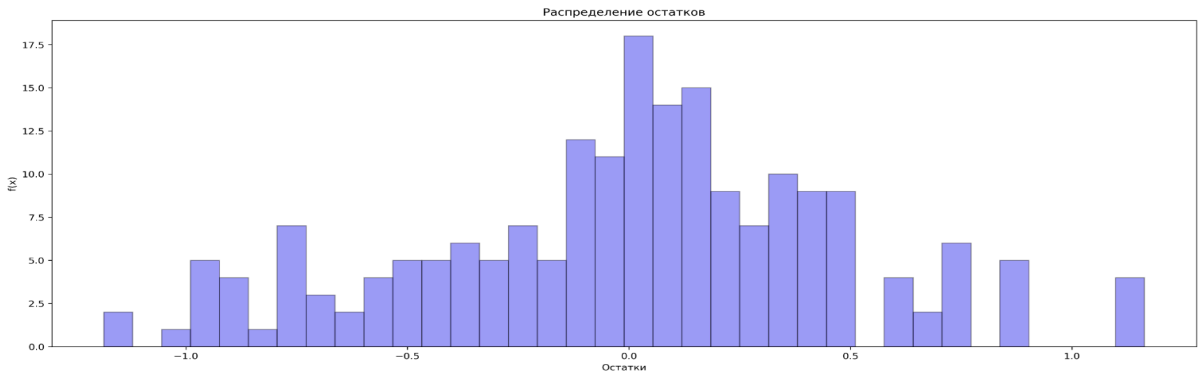
Далее необходимо проверить релевантность и валидность выбранного инструмента, для чего использовался Тест Хаусмана. Тест показал, что Pvalue=0, что говорит о том, что значит мы принимаем гипотезу о невалидности инструмента

Теперь проверим релевантность. Посчитаем F-статистику первого шага. Если она будет больше 10, то инструмент сильный (сильно коррелирует с X). F_statistics = 9.701

К сожалению, наш выбранный инструмент оказались невалидным и слабо коррелирующим с X. Это наталкивает на вывод, что в модели нет эндогенности. Сравнивая результаты моделей IV и нашей модели OLS, приходим к выводу, что IV модель показывает результат даже хуже, чем модель без инструментальных переменных. Оставляем старую модель, и считаем что наш X - экзогенный

Нормальность распределения остатков

Для начала посмотрим на гистограмму распределения остатков:



В целом, распределение похоже на нормальное. Проверим гипотезу на нормальность с помощью теста Харке-Бера. Тест показал следующие результаты, что статистика меньше критического значения, значит нулевая гипотеза не отвергается, значит остатки распределены нормально

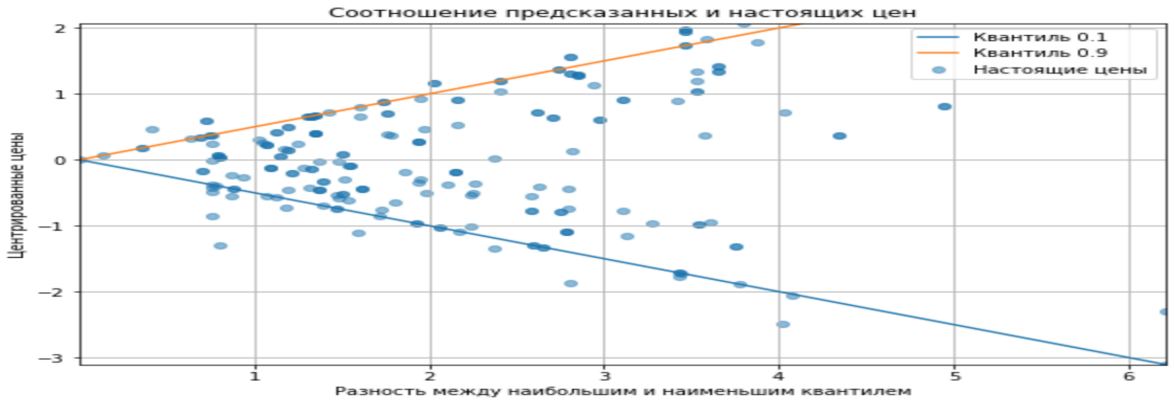
Таким образом, мы получили финальную модель и доказали, почему выбрали именно ее на каждом шаге. Теперь посмотрим на значимость коэффициентов модели и самой модели в целом. Ниже вывод только значимых переменных: Объем_постоянной_памяти, Сенсорный_дисплей, Управление_музыкальным_плеером, Функция_поиска_смартфона, Длина,Ширина, Время_работы_в_режиме_ожидания,_до, Функция_быстрой_зарядки, Высота Измерение_артериального_давления, Датчик_снятия_с_руки, Размер_дисплея, Разрешение_дисплея_2 и константа

Теперь финально покажем результат теста на адекватность модели в целом. F-test показывает, что Pvalue =0, значит H0 отвергается и модель адекватна

Квантильная регрессия.

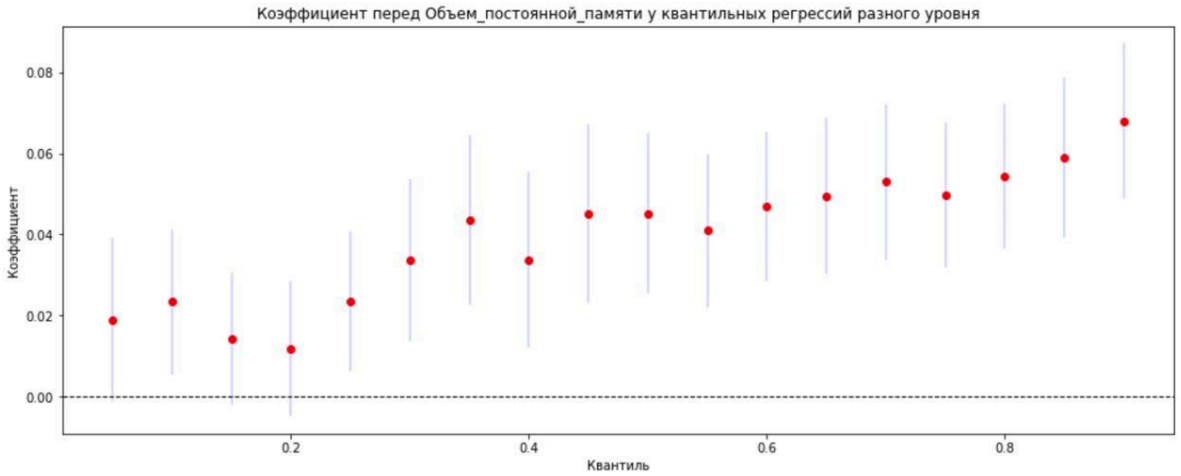
Для реализации квантильной регрессии мы опредилили количественные переменные. В нашем датасете получилось 15 количественных переменных, с которыми мы можем дальше работать.

Для более наглядного представления, получанных нами данных по умным часам, мы воспользовались центрированием наших значений. Ниже вы можете видеть график рассеивания реальных цен смарт часов. На этом графике представлено соотношение предсказанных и реальных значений, сравниваемых с разницей между наибольшим и наименьшим квантилями. Синяя линия (квантиль 0.1) показывает, что при малых разностях между квантилями предсказанные значения занижены по сравнению с реальными, а оранжевая линия (квантиль 0.9) показывает, что при увеличении разницы между квантилями предсказания становятся более точными и начинают превышать реальные значения. Это может указывать на различное поведение модели в зависимости от уровня изменчивости данных.



Анализ зависимостей цен от факторов.

Далее приведен пример графика, где представлено соотношение коэффициентов при выбранных нами регрессорах с квантилями, исходя из которых можно понять, как какие-либо параметры влияют на определенные сегменты смарт часов (дешевые, средние, дорогие).



Исходя из данного графика, можно сделать вывод о том, что "Объём постоянной памяти" в телефоне сильнее влияет на более дорогой сегмет смарт часов, чем на дешевый. То есть богатая часть населения готова больше платить деньги за большой объёем памяти на умных часах.

Предсказание для своего товара

Товар, который мы использовали для предсказания можно найти в коде в Python. Ниже представлен вывод результатов модели

Предсказание = 28316.5

CI = [3425.49, 233982.12]