Терников Андрей

Гедонический анализ рынка электробритв

1. Введение

Решение выбрать рынок электробритв для анализа взаимосвязи цен и различных характеристик, было принято осознанно. Недавно столкнувшись с таким выбором (какую же бритву стоит купить), я смотрел на различные характеристики бритв и отзывы покупателей. В более простом случае (без использования семантического анализа текстов отзывов в интернете) необходимо было сначала определиться, какие же характеристики следует принимать как основные и каким образом влияет марка производителя (его репутация) на цену, установленную на рынке на определенный товар.

Следует принять во внимание, что ценность подобного рода исследований заключается в анализе первичных максимально открытых данных, исследование характеристик товара и влияния брендов на установление цены (наценка). Ведь не всегда можно обнаружить качественный товар за огромным изобилием марок и, наоборот, очень легко приобрести товар известного производителя, который впоследствии не оправдает возложенных на него ожиданий.

Целью гедонического анализа выступает рассмотрение структуры цены от различных видимых (время работы, наличие определенных функций) и скрытых (влияние бренда) характеристик товара. Выявление переоцененных и недооцененных рынком товаров за счет влияния бренда. Кроме того, построение индекса цен позволит оценить покупательную способность относительно предыдущих периодов времени, и решить, какой товар можно было бы купить за те же деньги, например, полгода назад.

Основные задачи исследователя состоят в качественном сборе и анализе данных (особенно первичном). Это важно с позиции принятия решения о дальнейшей реализации регрессионных моделей и спецификации значимых (наиболее влиятельных) переменных.

В данном исследовании я постараюсь ответить на вопросы, связанные с осмысленностью установления цен на рынке электробритв и определить основные характеристики, имеющие преобладающее влияние на установление той или иной цены на определенную группу товаров.

2. Данные

Проведем оценку переменных, предварительно раздробив категориальную переменную «power» на три переменные:

Variable	Count	Mean	p25	p50	p75	SD	CV	Min	Max
price_aug	202	3001.059	825	1754.5	3916	2963.017	.9873238	292	13524
price_jan	225	4174	1155	2390	5490	4429.964	1.061323	420	22890
available	225	13.80444	1	12	24	11.4824	.8317901	1	33
shave	225	.4088889	0	0	1	.4927248	1.205034	0	1
power_1	225	.0933333	0	0	0	.2915476	3.123724	0	1
power_2	225	.6	0	1	1	.4909903	.8183171	0	1
power_3	225	.3066667	0	0	1	.4621379	1.506971	0	1
head	225	2.502222	2	3	3	.9118488	.3644156	1	5
float_head	225	.4888889	0	0	1	.5009911	1.024754	0	1
way_shave	225	.44	0	0	1	.4974937	1.130668	0	1
movable_unit	225	.36	0	0	1	.4810702	1.336306	0	1
adj_unit	225	.0977778	0	0	0	.2976762	3.044416	0	1
time_work	223	39.20628	30	45	50	17.12494	.4367908	0	100
time_charge	222	265.3153	60	60	480	265.9834	1.002518	0	1440
display	225	.1422222	0	0	0	.3500567	2.461336	0	1
fast_charge	225	.3111111	0	0	1	.4639804	1.491365	0	1
waterproof	225	.5288889	0	1	1	.5002777	.9459032	0	1
ind_charge	225	.7955556	1	1	1	.4041943	.5080654	0	1
ind_charge_d	225	.1822222	0	0	0	.3868883	2.123167	0	1
ind_dcharge	225	.2666667	0	0	1	.4432026	1.66201	0	1
ind_fcharge	225	.1688889	0	0	0	.3754891	2.223291	0	1
ind_rcharge	225	.0711111	0	0	0	.2575834	3.622267	0	1
ind_blades	225	.1288889	0	0	0	.335824	2.605531	0	1
ind_clean	225	.1066667	0	0	0	.3093773	2.900412	0	1
charge_dev	225	.1111111	0	0	0	.3149704	2.834734	0	1
nozzle	225	.0933333	0	0	0	.2915476	3.123724	0	1
trimmer	225	.7822222	1	1	1	.4136558	.5288213	0	1
charge_stand	225	.1511111	0	0	0	.3589557	2.375442	0	1
voltage_sw	225	.4355556	0	0	1	.4969351	1.140922	0	1
protect_cap	225	.4355556	0	0	1	.4969351	1.140922	0	1
travel_case	225	.5288889	0	1	1	.5002777	.9459032	0	1
clean_brush	225	.64	0	1	1	.4810702	.7516722	0	1
travel_lock	225	.2266667	0	0	0	.4196087	1.851215	0	1
grease_oil	225	.0888889	0	0	0	.2852178	3.208701	0	1
slip_handle	225	.2888889	0	0	1	.4542568	1.572427	0	1
info	225	.3555556	0	0	1	.4797486	1.349293	0	1
braun	225	.2088889	0	0	0	.4074212	1.950421	0	1
philips	225	.1822222	0	0	0	.3868883	2.123167	0	1
panasonic	225	.1244444	0	0	0	.3308239	2.658406	0	1
remington	225	.0711111	0	0	0	.2575834	3.622267	0	1
berdsk	225	.0533333	0	0	0	.2251983	4.222469	0	1
vitek	225	.0488889	0	0	0	.2161165	4.420565	0	1
ladomir	225	.0488889	0	0	0	.2161165	4.420565	0	1
other	225	.2622222	0	0	1	.4408235	1.681107	0	1

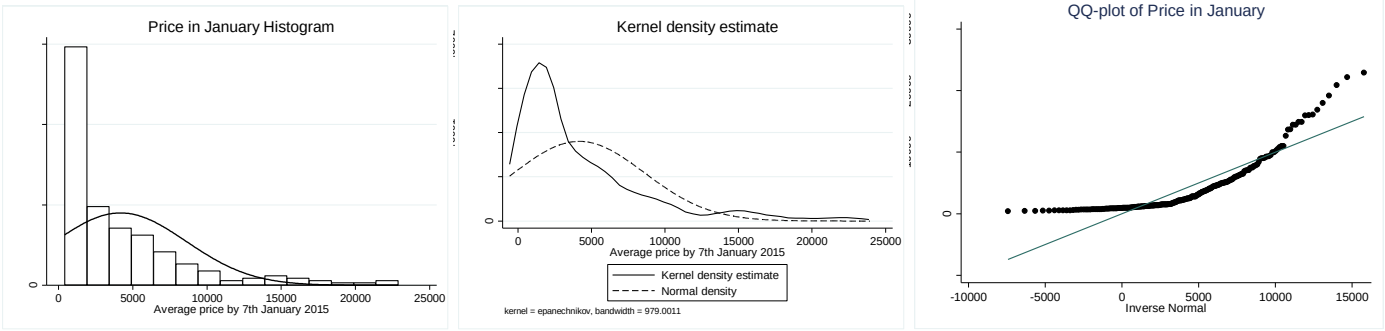
Характер переменных довольно разнороден (у большинства переменных высокий коэффициент вариации). Кроме того, почти у всех количественных переменных наблюдается правосторонняя асимметрия (среднее больше медианы), что говорит нам о неком смещении данных относительно нормального распределения. В дополнение к вышесказанному категориальные переменные не вырождены, но достаточно разнородны (процентное распределение для них характеризуется средним значением, поскольку они принимают значения 0 и 1).
Если смотреть по показателю вариации, то все без исключения переменные нехороши с точки зрения разброса данных.

Поскольку в исследование предполагается исследовать силу марок и их влияние на цену, с необходимостью следует привести описательные статистики январской цены в зависимости от определенного бренда.

price_jan	Count	Mean	p25	p50	p75	SD	CV	Min	Max
Braun	47	6150.17	3576	5100	8390	3677.828	.5980043	999	16900
Philips	41	7967.683	3690	5993	10190	5851.01	.7343427	2160	22890
Panasonic	28	6219.286	2495	5110	7826.5	4533.453	.7289347	1056	19200
Remington	16	3073.063	2250	2799.5	3893	1443.887	.4698527	1094	6892
Бердск	12	1183.25	930	1235	1475	286.1078	.2417982	700	1500
VITEK	11	1309.182	1095	1281	1441	386.8247	.2954706	719	2240
Ладомир	11	746.3636	520	720	920	275.1099	.3686003	470	1300
Other	59	1072.847	730	960	1360	427.0765	.3980776	420	2500
Total	225	4174	1155	2390	5490	4429.964	1.061323	420	22890

Теперь, если смотреть с точки зрения показателя вариации, выделяются 2 марки с приемлемым разбросом цен. Но, огрубляя, можно отметить, что марки (кроме TOP-3 по количеству в выборке) довольно сконцентрированы в определенном ценовом диапазоне. С обыденной точки зрения у «топовых» марок и выбор моделей больше и ценовой сегмент более обширный.

Однако почему же возникает такая разнородность в цене по всей выборке в целом? Для этого следует проверить январскую цену на нормальность распределения. Сначала графически (для наглядности частотная гистограмма, график ядерной плотности и «кваниль-квантиль»):



Все три графических метода показывают, что распределение цены довольно далеко от нормального. Убедимся в этом и при помощи формальных тестов:

Shapiro-Francia W' test for normal data

Variable	Obs	W'	V'	z	Prob>z
price_jan	225	0.76577	42.179	7.797	0.00001

Shapiro-Wilk W test for normal data

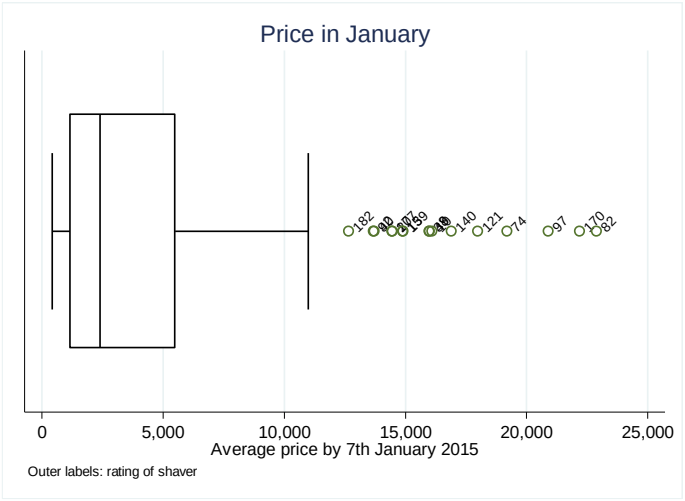
Variable	Obs	W	V	z	Prob>z
price_jan	225	0.76595	38.699	8.461	0.00000

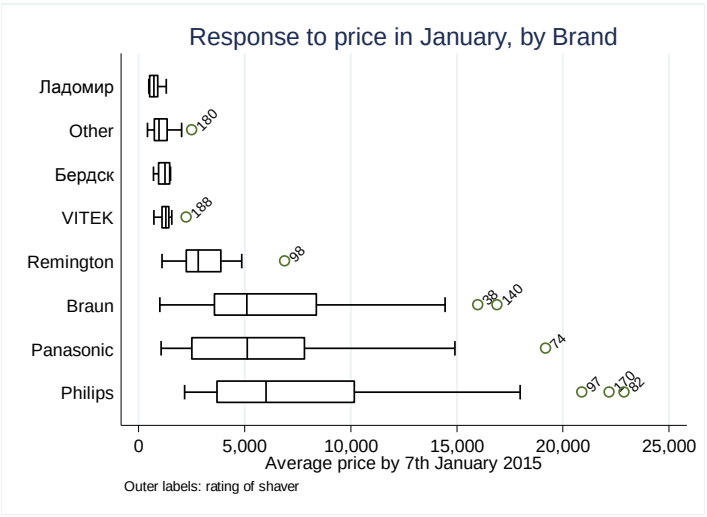
Skewness/Kurtosis tests for Normality

Variable	Obs	Pr (Skewness)	Pr (Kurtosis)	adj chi2 (2)	Prob>chi2
price_jan	225	0.0000	0.0000	70.70	0.0000

Здесь все три теста позволяют отвергнуть нулевую гипотезу о нормальности распределения.

Получается, что что-то может влиять на распределение нашей зависимой переменной. Тогда следует проверить распределение на выбросы (в общем, и отдельно по брендам):





Если смотреть в целом на диаграмму по январской цене, то 75-й персентиль ящика лежит ниже 12000 рублей за единицу товара. Обращая внимание на разделение по брендам, заметим, что практически для каждого бренда существуют «выбивающиеся» из основной массы товары с высокой ценой, но их относительно немного. Кроме того, у «топовых» брендов значительно выше как медианная цена, так и размах вариации, что подтверждает предшествующие рассуждения.

По-хорошему для исследования необходимо оставить в данных максимально возможное учтенное количество моделей, чтобы проанализировать рынок с минимальными потерями данных. Подойдем к решению этого вопроса более формально. Для этого разберемся с коррелированностью зависимой переменной и других количественных переменных (предполагается, что коррелированность (даже если она есть) будет оставаться примерно на том же уровне, что с выбросами, что без них).

Создадим переменную «price_jan2», в которой уберем все цены больше, чем 12000 руб. за единицу товара. Ниже представлены фрагменты корреляционных матриц для наблюдений «с» выбросами и «без» них:

	price_jan	price_jan2
price_aug	0.980***	0.971***
available	0.103	0.0861
head	0.404***	0.376***
time_work	0.304***	0.269***
time_charge	-0.475***	-0.528***

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Видно, что ни характер, ни направление связи не изменились (очень незначительно), что позволяет нам на данном этапе сохранить в данных те модели, чья цена выше 12000 рублей.

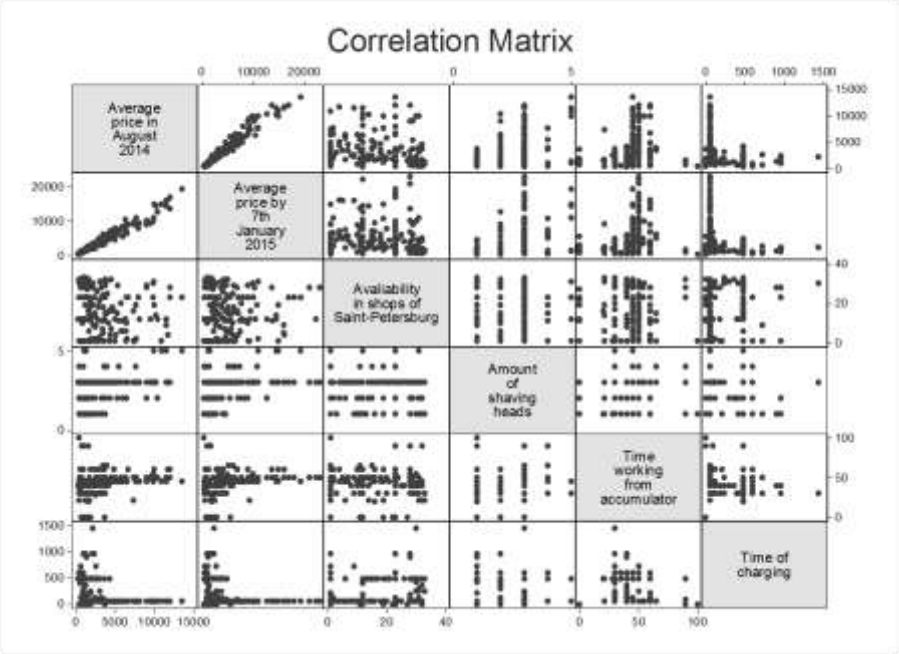
Теперь следует привести полную таблицу парных корреляций между количественными переменными:

	price_aug	price_jan	available	head	time_work	time_charge
price_aug	1					
price_jan	0.980***	1				
available	0.0957	0.103	1			
head	0.411***	0.404***	-0.0153	1		
time_work	0.306***	0.304***	0.150*	0.0823	1	
time_charge	-0.477***	-0.475***	0.0519	-0.337***	-0.0227	1

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

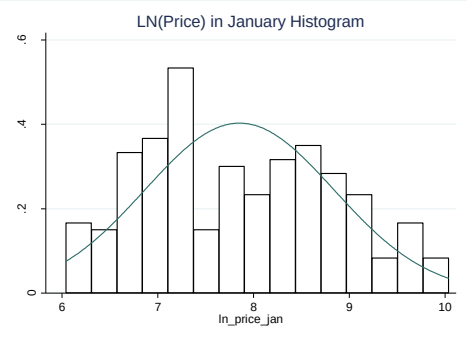
Не считая сильной положительной коррелированности между ценами января и августа (что довольно естественно), существенно сильных корреляций между переменными не наблюдается (как среди значимых, так и среди незначимых). Поэтому нам нет смысла (предварительно) выкидывать из будущей модели переменные, опасаясь за то, что они сильно взаимосвязаны и объясняют друг друга непосредственно.

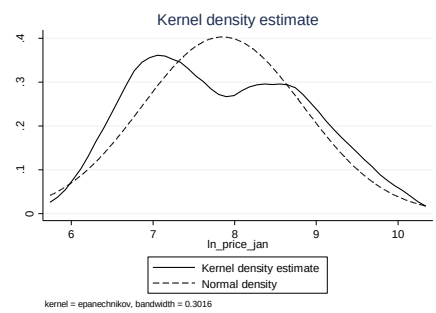
Эти данные можно представить также в корреляционной матрице (график попарных корреляций):

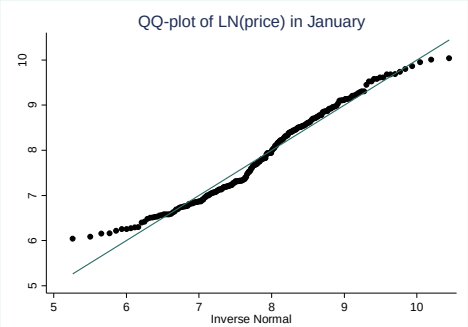


Ячейки матрицы для нашей зависимой переменной (средняя цена в январе) указывают на то, что спецификация функциональной формы в первоначальных данных требует доработки.

С первого взгляда напрашивается линеаризация (с помощью логарифма) нашей зависимой переменной. Изменение цены на определенное число процентов за счет изменения независимой переменной (при прочих равных) на 1 пункт. Эта интерпретация экономически осмысленна: осталось проверить, улучшилась ли ситуация с нормальностью распределения, выбросами и парными корреляциями (взаимосвязи). Проведем такой же анализ для новой переменной «ln_price_jan», равной логарифму январской цены:







Теперь распределение стало ближе напоминать нормальное (но все еще таковым не является). Интересно, что более явно проявилась бимодальность распределения, что может объясняться наличием в выборке товаров двух преобладающих категорий: бюджетные электробритвы (их цена ниже средней) и электробритвы премиум-класса (их цена выше средней).

Формальные тесты снова отвергают гипотезу о нормальности распределения:

Shapiro-Francia W' test for normal data

Variable	Obs	W'	V'	z	Prob>z
ln_price_jan	225	0.97228	4.993	3.350	0.00040

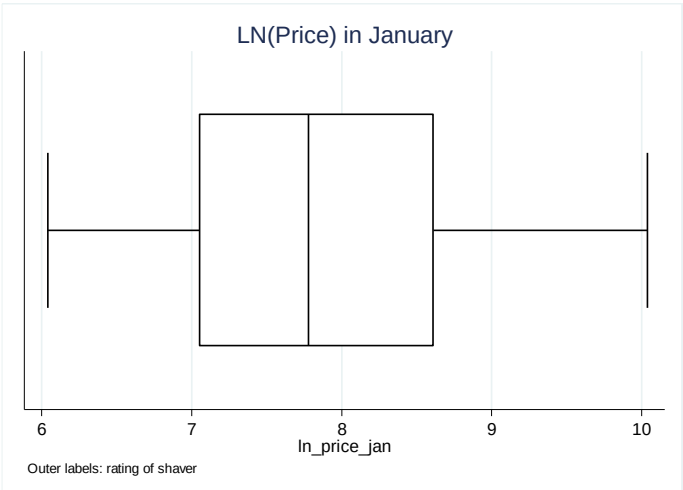
Shapiro-Wilk W test for normal data

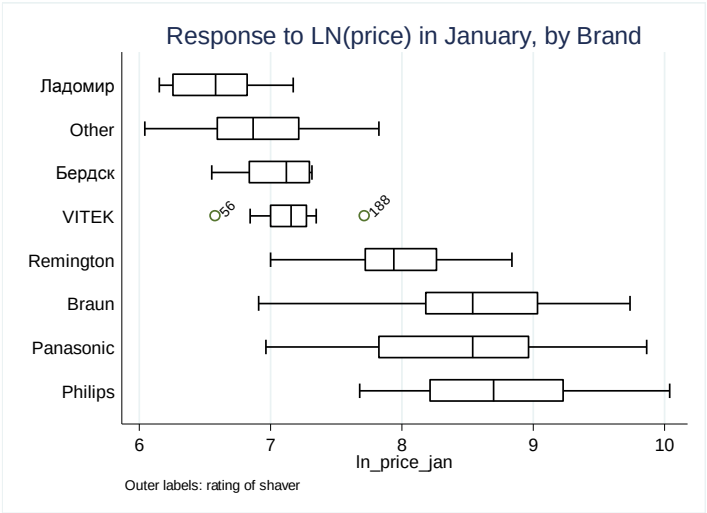
Variable	Obs	W	V	z	Prob>z
ln_price_jan	225	0.96878	5.161	3.799	0.00007

Skewness/Kurtosis tests for Normality

Variable	Obs	Pr (Skewness)	Pr (Kurtosis)	adj chi2 (2)	Prob>chi2
ln_price_jan	225	0.1625	0.0000	26.75	0.0000

Однако с добавлением логарифма заметно улучшилась ситуация с выбросами (в целом их не осталось):

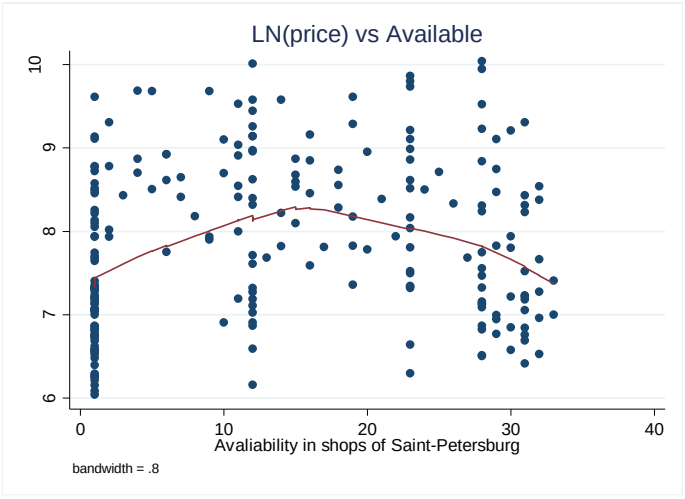


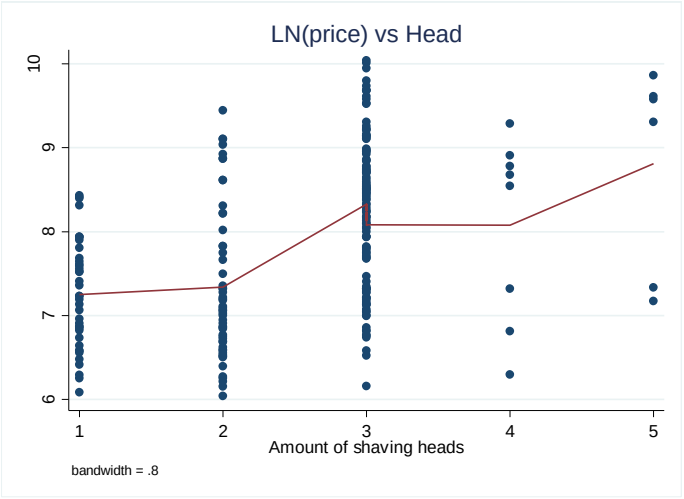


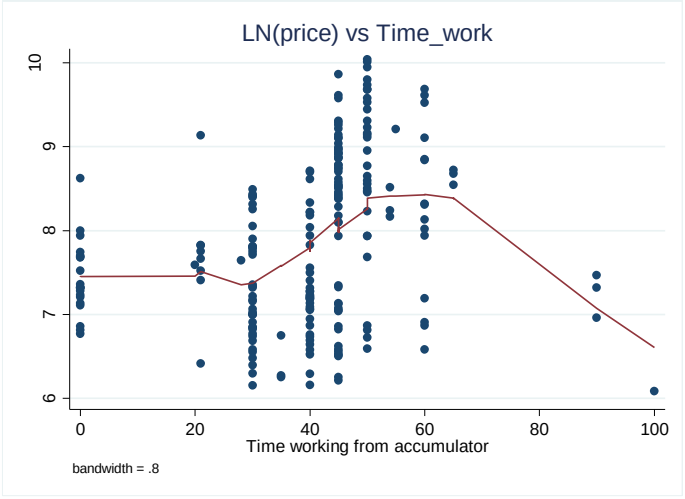
Корреляции также практически не изменили своего влияния за счет взятия логарифма зависимой переменной:

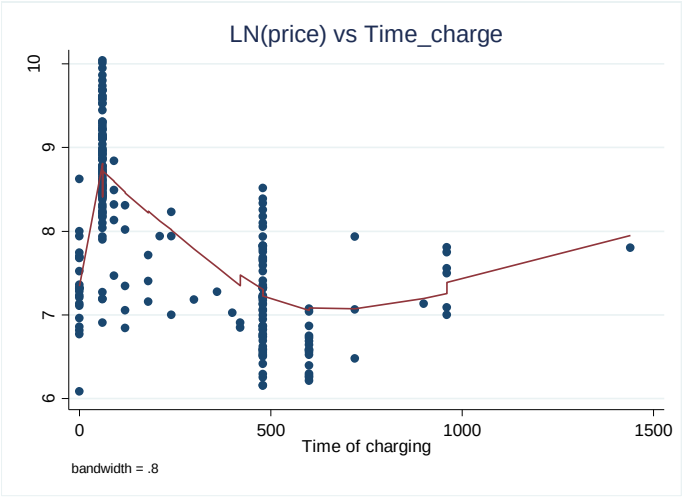
	ln_price_jan
available	0.129
head	0.436***
time_work	0.274***
time_charge	-0.575***

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$









Спецификация зависимой переменной совсем немного улучшила ситуацию, но не стоит делать преждевременных выводов.

Переходя к построению базовой модели, необходимо отметить, что несмотря на то, что распределение все-таки оказалось не совсем близко к нормальному, удалось максимально приблизить его к нормальному. Также можно проводить дальнейший анализ вследствие достаточного количества наблюдений (>100), при котором можно говорить о близости распределения к нормальному (однако оставим процедуру более тщательной проверки до момента построения моделей). С введением логарифма цены удалось избежать выбросов, что, несомненно, сохраняет наши данные в первоначальном составе. Наконец не было выявлено хоть сколько-нибудь сильно коррелированных между собой количественных переменных, что поможет в дальнейшем объяснять зависимую переменную непосредственно через взятые характеристики. Но осталась проблема со спецификацией функциональной формы переменных, которую попытаемся решить уже после построения регрессионных моделей.

3. Оценка базовой модели

3.1. Базовая модель

Построим базовую модель:

$$\text{ln\_price\_jan} = \beta_0 + \beta_1 \text{head} + \beta_2 \text{time\_charge} + \beta_3 \text{voltage\_sw} + \beta_4 \text{display} + \beta_5 \text{power\_2} + \beta_6 \text{power\_3}$$

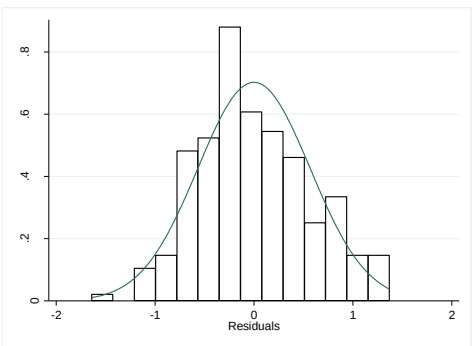
	ln_price_jan	
head	0.163***	(3.53)
time_charge	-0.00161***	(-6.26)
voltage_sw	0.773***	(8.65)
display	0.546***	(4.11)
power_2	1.006***	(8.31)
power_3	1.000***	(7.78)
_cons	6.556***	(43.83)
N	222	
adj. R^2	0.659	
AIC	391.9	
BIC	415.8	

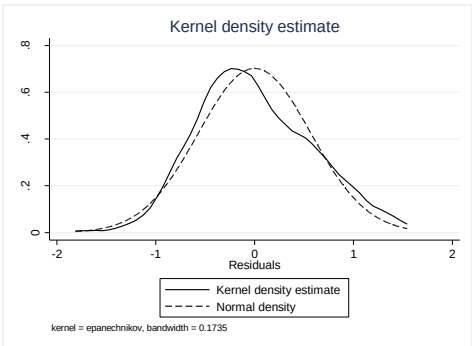
t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

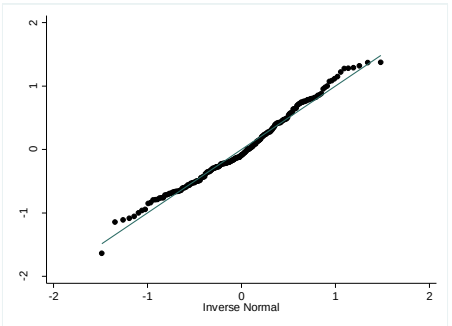
Анализ VIF показал, что в базовой модели мультиколлинеарности нет (все VIF<5).

Variable	VIF	1/VIF
power_3	3.48	0.287491
power_2	3.47	0.288594
time_charge	1.47	0.679572
head	1.23	0.81029
display	1.23	0.815324
voltage_sw	1.17	0.853594
Mean VIF	2.01	

Проверим на нормальность распределения остатков регрессии:







Формальные тесты. Распределение, скорее всего, отличается от нормального, но не слишком сильно (p-value<0.1).

Shapiro-Francia W' test for normal data

Variable	Obs	W'	V'	z	Prob>z
e	222	0.98632	2.435	1.852	0.03199

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
e	222	0.98533	2.397	2.022	0.02156

Skewness/Kurtosis tests for Normality

Variable	Obs	Pr (Skewness)	Pr (Kurtosis)	adj chi2 (2)	Prob>chi2
e	222	0.0828	0.4404	3.64	0.1623

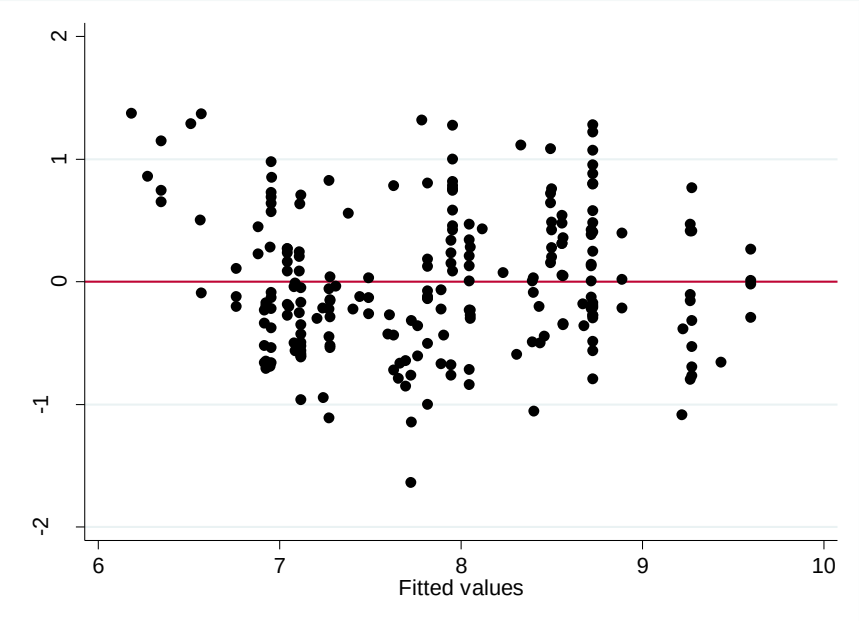
Выбросы. Поиск выбросов проводился на основании трех параметров: Studentized residuals, расстояние Кука и показатель DFITS. Полученный лист переменных не вошел в регрессию, поскольку каждая модель из этого списка имеет пропущенное значение по регрессору «time_charge» (т.е. их удаление не сместит оценки коэффициентов регрессии):

rate	brand	model	price_jan
107	Bellissima	BS2 100	2014
201	Mayer&Boch	10028	910
224	Sterlingg	10660	420

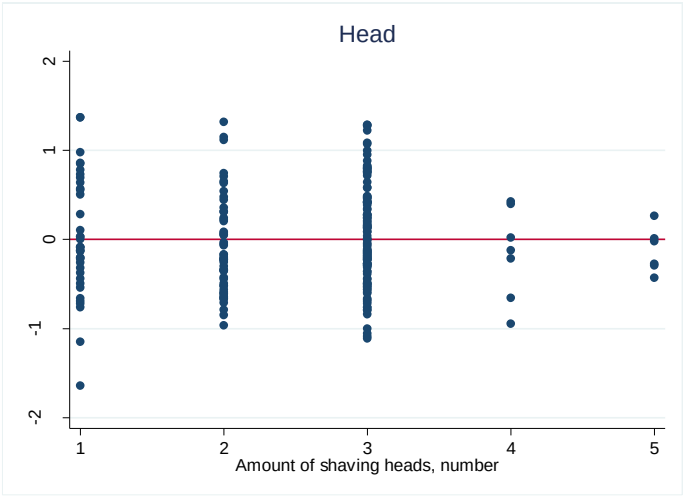
3.2. Графическая диагностика модели

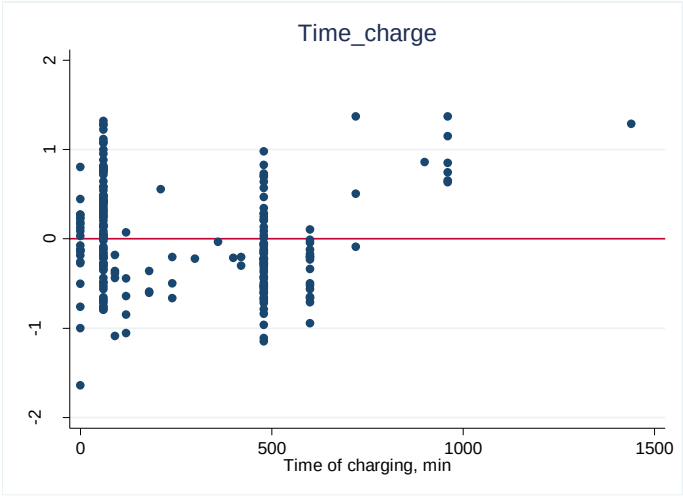
Обратимся теперь к графическому представлению остатков регрессии:

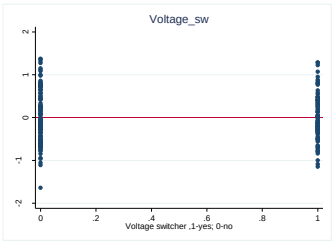
По всей выборке:

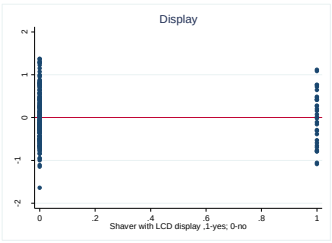


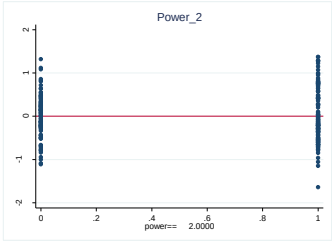
В зависимости от регрессора

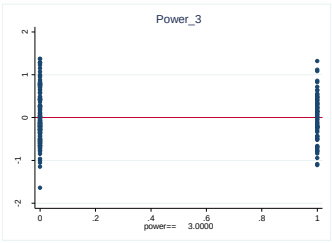












С первого взгляда, несомненно, присутствуют как ошибки в функциональной форме, так и гетероскедастичность остатков модели (скачкообразное распределение дисперсии). Однако все же стоит применить формальные тесты для выявления такого рода неточностей модели.

3.3. Эконометрические тесты

Прежде чем перейти к проверке модели на гетероскедастичность проверим нашу регрессию на наличие ошибок в функциональной форме.

Linktest:

ln_price_jan	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_hat	-0.83754	0.913195	-0.92	0.36	-2.63731	0.962238
_hatsq	0.11581	0.057477	2.01	0.045	0.002531	0.229089
_cons	7.21346	3.599442	2	0.046	0.11948	14.30744

Ramsey test:

Ramsey RESET test using powers of the fitted values of ln_price_jan Ho: model has no omitted variables
F(3, 212) = 15.59 Prob > F = 0.0000

Коэффициент перед квадратом предсказанного значения «ln_price_jan» в Linktest значим, т.е. гипотеза о том, что квадрат предсказанных значений незначим, отвергается. Кроме того, по Ramsey test эта гипотеза усиливается добавлением куба и четвертой степени предсказанных значений. Таким образом, ошибка спецификации выявлена.

Теперь можно проверить на гетероскедастичность остатков:

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity Ho: Constant variance Variables: fitted values of ln_price_jan
chi2(1) = 1.17 Prob > chi2 = 0.2797

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity Ho: Constant variance Variables: head time_charge voltage_sw display power_2 power_3
chi2(6) = 10.88 Prob > chi2 = 0.0922

White's test for Ho: homoskedasticity against Ha: unrestricted heteroskedasticity
chi2(20) = 52.90 Prob > chi2 = 0.0001

Гетероскедастичности по тестам Бреуша-Пагана нет, а по тесту Уайта – есть. Мы бы могли грубо пренебречь тестом Уайта по отношению к тесту Бреуша-Пагана, если бы в нашей модели отсутствовали ошибки спецификации функциональной формы. Однако в нашем случае это не так, поэтому следует попробовать либо преобразовать, либо добавить переменные.

Преобразование.

Попробуем преобразовать нашу базовую модель. Например, для того, чтобы полностью убрать неточности в функциональной форме (формально по двум тестам) сгенерируем новую переменную, равную отношению времени работы от аккумулятора и временем зарядки от него (условный КПД электробритвы и назовем ее «time»). Кроме того, удалим из регрессии все регрессоры, кроме «display»:

	ln_price_jan	
time	2.084***	(13.74)
display	0.437***	(3.79)
_cons	6.987***	(107.10)
N	199	
adj. R^2	0.654	
AIC	361.0	
BIC	370.8	

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Проверим на спецификацию функциональной формы переменных в новой модели:

Linktest:

ln_price_jan	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_hat	3.386292	1.910628	1.77	0.078	-0.38174	7.15432
_hatsq	-0.1485	0.118855	-1.25	0.213	-0.3829	0.085901
_cons	-9.48558	7.603077	-1.25	0.214	-24.4799	5.508766

Ramsey test:

Ramsey RESET test using powers of the fitted values of ln_price_jan Ho: model has no omitted variables
F(3, 193) = 1.35 Prob > F = 0.2597

Тесты не выявили ошибок в функциональной форме. Однако в данной модели мы не учли электробритвы, работающие от сети (переменная «power_1», ≈10% от выборки, поскольку в графе «time_charge» подразумевается время зарядки аккумулятора, а для бритв, работающих от сети – это значение равно нулю; похожие соображения ведутся и по поводу переменной «time_work», где помимо работы от аккумулятора подразумевается и время работы от батарейки: но никак не от сети), поскольку переменная «time» полностью исключает их из рассмотрения. В погоне за лучшей спецификацией модели можно забыть также и о гетероскедастичности (в двух случаях из трёх наблюдается отвержение гипотезы о гомоскедастичности):

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity Ho: Constant variance Variables: fitted values of ln_price_jan
chi2(1) = 0.56 Prob > chi2 = 0.4542

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity Ho: Constant variance Variables: time display
chi2(2) = 7.09 Prob > chi2 = 0.0289

White's test for Ho: homoskedasticity against Ha: unrestricted heteroskedasticity
chi2(4) = 16.84 Prob > chi2 = 0.0021

Еще один путь борьбы с проблемой функциональной формы – добавление недостающих регрессоров. Например, добавим в регрессию «водонепроницаемость»:

	ln_price_jan	
head	0.164***	(3.88)
time_charge	-0.00137***	(-5.09)
voltage_sw	0.695***	(7.67)
display	0.508***	(4.09)
waterproof	0.394***	(3.98)
power_2	0.690***	(4.96)
power_3	0.840***	(6.28)
_cons	6.560***	(47.34)
N	222	
adj. R^2	0.683	
AIC	377.0	
BIC	404.2	

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Проверим на функциональную форму:

Linktest:

ln_price_jan	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_hat	-0.47552	0.863142	-0.55	0.582	-2.17665	1.225611
_hatsq	0.093061	0.054366	1.71	0.088	-0.01409	0.200208
_cons	5.785912	3.398523	1.7	0.09	-0.91208	12.48391

Ramsey test:

Ramsey RESET test using powers of the fitted values of ln_price_jan Ho: model has no omitted variables
F(3, 211) = 12.88 Prob > F = 0.0000

По Linktest'у мы смогли отвергнуть ошибку в спецификации, а по тесту Ramsey – нет. Однако, несмотря на то, что формально проблемы со спецификацией существуют, один из тестов не выявил серьезной ошибки, поэтому будем считать, что ошибки в нашей спецификации не слишком серьезные.

Теперь перейдем к проверке на гетероскедастичность улучшенной базовой модели:

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity Ho: Constant variance Variables: fitted values of ln_price_jan
chi2(1) = 0.39 Prob > chi2 = 0.5335

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity Ho: Constant variance Variables: head time_charge voltage_sw display waterproof power_2 power_3

chi2(7) = 9.25 Prob > chi2 = 0.2355

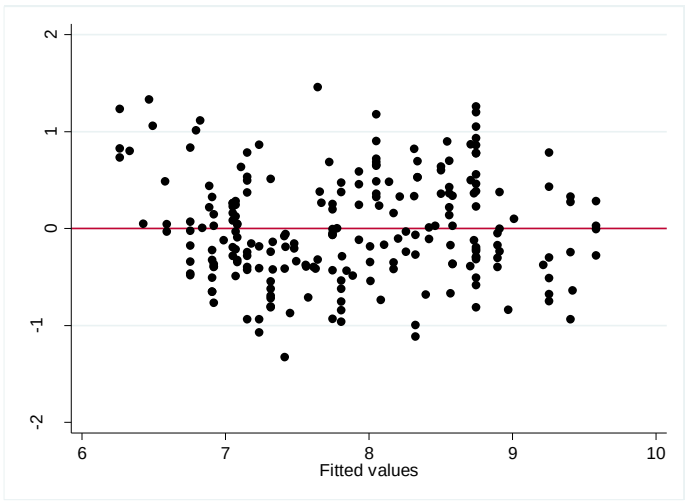
White's test for Ho: homoskedasticity against Ha: unrestricted heteroskedasticity
chi2(27) = 59.32 Prob > chi2 = 0.0003

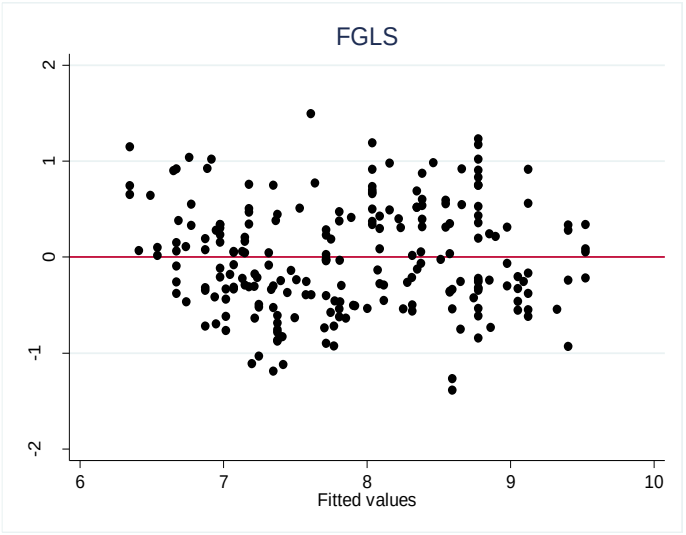
Гетероскедастичности по тестам Бреуша-Пагана нет, а по тесту Уайта – есть. Можно остановиться на этом и сказать, что тест Бреуша-Пагана не смог выявить гетероскедастичности и нам этого вполне достаточно, но для того, чтобы формально уменьшить гетероскедастичность – воспользуемся взвешенной регрессией FGLS (но этот подход лишь повысит эффективность оценок коэффициентов регрессии и уменьшит точность нашей модели):

	ln_price_jan	
head	0.200***	(4.77)
time_charge	-0.00109***	(-6.44)
voltage_sw	0.740***	(9.86)
display	0.348**	(2.74)
waterproof	0.505***	(5.10)
power_2	0.622***	(5.17)
power_3	0.899***	(8.00)
_cons	6.376***	(46.87)
N	222	
adj. R^2	0.700	
AIC	331.4	
BIC	358.6	

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Однако графически различий между первоначальной и FGLS моделями почти не наблюдается:





4. Сравнение альтернативных спецификаций

4.1. Обоснование и оценка

Перейдем к проблеме пропущенных переменных. Ведь если пропустить важный регрессор, оценки коэффициентов будут смещенными (стандартные ошибки будут занижены). Или наоборот, если добавить «антирегрессор», который будет лишним в модели, то, скорее всего оценки коэффициентов будут несмещенными, но иметь завышенные стандартные ошибки. Поэтому следует обратить на это внимание и постараться избежать.

В базовой регрессии довольно вероятен пропуск переменных (трудно сказать насколько важных, но пропуск очевиден). Доля объясненной дисперсии около 0.66 при относительно высоких информационных критериях (т.е. теряем больше информации). Однако сильное смещение оценок маловероятно (если пользоваться поправкой robust для оценок, то, невзирая на существующую гетероскедастичность (по тесту Уайта), мы можем быть уверены не только в несмещенности и состоятельности, но и эффективности полученных оценок).

Разберемся подробнее с этими проблемами и рассмотрим несколько возможных спецификаций базовой модели:

1) В первой спецификации вернемся к первоначальной зависимой переменной (без логарифма), а остальные переменные оставим такими же. С содержательной точки зрения теперь мы не будем искать процентного изменения цены при изменении на единицу любого из факторов (при прочих равных), а количественное выражение этого изменения (в рублях). По сути нельзя сравнивать модели с разными зависимыми переменными. Однако удобно представить эти модели рядом [(1) – новая модель]:

	(1) price_jan	Базовая ln_price_jan
head	729.1***	0.163***
	(3.87)	(3.53)
time_charge	-5.672***	-0.00161***
	(-6.72)	(-6.26)
voltage_sw	2700.7***	0.773***
	(6.10)	(8.65)
display	3245.4***	0.546***
	(3.41)	(4.11)
power_2	4405.4***	1.006***
	(7.63)	(8.31)
power_3	3896.8***	1.000***
	(7.65)	(7.78)
_cons	-1596.6*	6.556***
	(-2.59)	(43.83)
N	222	222
adj. R^2	0.516	0.659
AIC	4205.4	391.9
BIC	4229.2	415.8

t statistics in parentheses * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

2) В следующем случае воспользуемся базовой моделью и заменим в ней переменную «time_charge» (время зарядки от аккумулятора, мин.) на сгенерированную переменную «work_charge» (отношение «time_work», времени работы от аккумулятора (мин.), к «time_charge»). В процессе генерации переменной пришлось столкнуться с тем, что, во-первых, было задано бесконечно малое приближение для переменных равных нулю (для того, чтобы деление было нетривиально осмысленным: если уж выключить резко сетевую бритву, то она наверняка еще секунду без питания поработает на остаточном ходу), и, во-вторых, было найдено и исключено два выброса, уходящих своими значениями далеко в бесконечность (модели, работающие от батарейки, и не имеющие времени зарядки). Сама по себе новая переменная с содержательной точки зрения может обозначать КПД бритвы (сколько проработала относительно того, сколько заряжалась) [(2)].
3) В заключительной спецификации переменная «work_charge» была заменена на «ln_work_char», то есть логарифм от «work_charge», что содержательной точки зрения может интерпретироваться как эластичность цены по КПД, то есть при изменении КПД на 1% цена изменится на столько процентов, сколько показывает коэффициент перед КПД. Кроме того, добавлена переменная «waterproof», водонепроницаемость, которая по смыслу скорее ближе к бритвам, работающим не от сети [(3)].

4.2. Сравнение спецификаций.

Теперь сначала сравним полученные модели (зависимая переменная одинаковая):

	(2) ln_price_jan	(3) ln_price_jan	Базовая ln_price_jan
head	0.100*	0.112**	0.163***
	(2.34)	(3.00)	(3.53)
voltage_sw	0.621***	0.607***	0.773***
	(8.02)	(7.86)	(8.65)
display	0.429***	0.390***	0.546***
	(3.80)	(3.43)	(4.11)
power_2	1.470***	0.965***	1.006***
	(10.65)	(7.24)	(8.31)
power_3	1.474***	1.074***	1.000***
	(10.84)	(8.79)	(7.78)
work_charge	1.613***		
	(10.54)		
ln_work_char		0.417***	
		(9.91)	
waterproof		0.304***	
		(3.37)	
time_charge			-0.00161***
			(-6.26)

_cons	5.193***	6.754***	6.556***
	(33.03)	(54.82)	(43.83)
N	220	220	222
adj. R^2	0.741	0.742	0.659
AIC	325.8	326.2	391.9
BIC	349.5	353.3	415.8

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

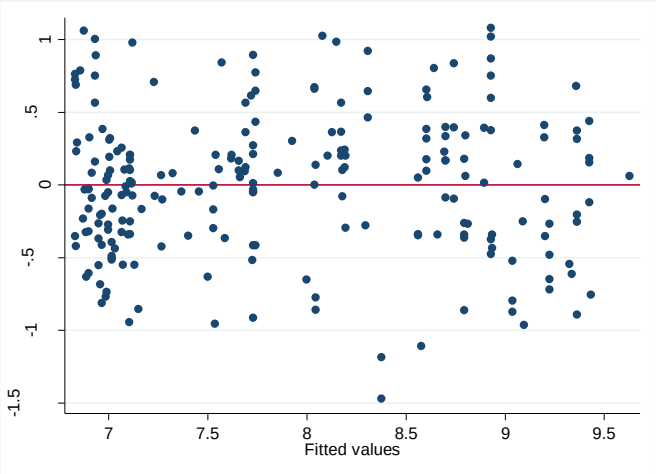
Эти модели мы можем сравнить и сказать, что 2 новые полученные спецификации примерно равны между собой и довольно значительно лучше базовой модели как по R^2-adj, так и по информационным критериям AIC и BIC (ниже, чем в базовой – значит, мы упустили меньше информации на выходе).

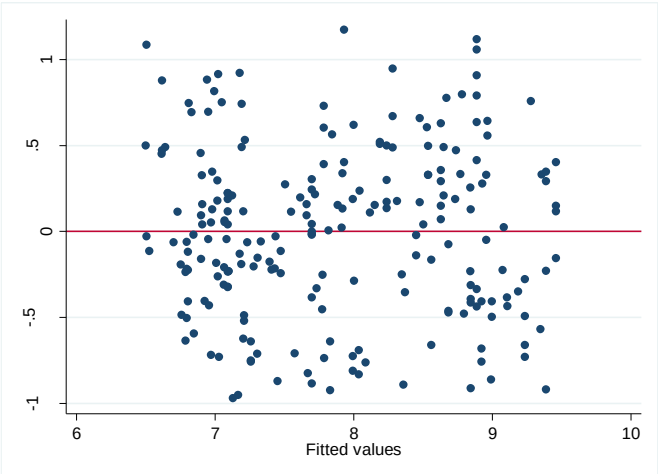
Однако не все так прекрасно получилось в этих моделях, поскольку переменные «power_2» и «power_3» сильно обратно коррелированны между собой (-0.8145 на 5% уровне значимости).
Проверим обе модели на мультиколлинеарность:

(2)		(3)	
Variable	VIF	Variable	VIF
power_3	4.04	power_2	4.73
power_2	4.04	power_3	3.87
work_charge	1.85	ln_work_char	1.87
head	1.26	waterproof	1.66
display	1.24	display	1.26
voltage_sw	1.23	voltage_sw	1.26
		head	1.25
Mean VIF	2.28	Mean VIF	2.27

Все показатели VIF для обеих моделей меньше 5, что говорит нам о том, что мультиколлинеарности не было выявлено.

Графическое распределение остатков регрессий показывает чуть лучшую дисперсию в модели (3) по сравнению с (2) (более равномерное распределение относительно нуля):





Обратимся тогда, например, к третьей модели и проинтерпретируем, полученные в ней коэффициенты: все коэффициенты оказались значимыми на 5%-ом уровне. Также интересно, что все коэффициенты имеют положительный знак (т.е. с увеличением показателя при коэффициенте на 1 пункт при прочих

равных условиях цена увеличивается на определенное количество процентов). Например, при увеличении бритвенных головок («head») на 1 единицу цена увеличится на $(e^{0.112}-1)\cdot 100\% = 11.85\%$ (при прочих равных условиях). Также довольно интересной получилась константа модели (если обнулить все коэффициенты, то минимальная цена установится на уровне 857.5 рублей). В этом случае из выборки пропадают еще 7 моделей стоимостью меньшей этой. Интерпретацией коэффициента КПД («ln_work_char») является эластичность: то есть при увеличении КПД на 1% цена увеличится (примерно) на 0.417 процента.

5. Оценка недооцененности (переоцененности) моделей продукта.

Рассчитав предсказанные значения логарифма цены по модели, и, найдя разницы между фактическим и предсказанным значениями (после оценки регрессии без константы) можно вывести таблицу наиболее переоцененных и недооцененных моделей:

rate	brand	model	price_jan	Цена по модели
133	Irit	IR-3020	473	1247.02
219	Ладомир	B820	500	1296.265
71	Braun	MobileShave M-60	999	2518.319
55	Braun	Contour 360 Series 3	4770	11956.2
65	Panasonic	ES-RT33	2790	6947.249
51	Saturn	ST-HC7395	1750	4269.329
163	Бердск	3364	910	2203.032
169	Saturn	ST-HC8018	720	1722.219
204	Remington	R6150	3399	8041.919
154	Sakura	SA-5401	1349	3099.162
.....				
195	Remington	F3800	1800	748.9334
33	Panasonic	ES-RW30	2500	1035.405
121	Philips	S9151	17990	7259.186
3	Braun	CruZer5 Body	2797	1122.155
62	Philips	PT 723	3290	1309.778
88	Philips	RQ 1185	10190	3956.864
97	Philips	S9521	20890	7259.186
46	Braun	150s-1 Series 1	1980	669.4771
170	Philips	S9511	22190	7259.186
128	Braun	5070cc Series 5	8990	2785.899

Для производителей эта информация полезна тем, что, посмотрев на характеристики и, посмотрев на цену, они могут прикинуть, что данная модель электробритвы себя (не)окупает (в рамках данной модели) и как стоит уделять внимание тем характеристикам, которые указаны в модели, чтобы контролировать цену. Для магазинов эта информация поможет узнать, на какие характеристики обращать внимание, и какой товар стоит выставлять ближе, а на какой товар делать специальные акции.

6. Оценка силы марок.

В качестве бренда-эталона возьмем марку «Braun». Построим регрессию цены относительно брендов.

	price_jan	
philips	1817.5	(1.70)
panasonic	69.12	(0.07)
remington	-3077.1***	(-4.76)
berdsk	-4966.9***	(-9.09)
vitek	-4841.0***	(-8.77)
ladomir	-5403.8***	(-9.89)
other	-5077.3***	(-9.34)
_cons	6150.2***	(11.38)
N	225	
adj. R^2	0.391	
AIC	4312.9	
BIC	4340.2	

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Теперь проверим гипотезу о значимости бренда:

H0: philips=panasonic=remington=berdsk=vitek=ladomir=other=0

```
( 1)  philips - panasonic = 0
( 2)  philips - remington = 0
( 3)  philips - berdsk = 0
( 4)  philips - vitek = 0
( 5)  philips - ladomir = 0
( 6)  philips - other = 0
( 7)  philips = 0

       F(  7,   217) =   33.33
            Prob > F =    0.0000
```

Менее чем на 0.0001 уровне значимости мы отвергаем нулевую гипотезу о том, что марки неразличимы с позиции влияния на январскую цену. Значит, бренд каким-то образом влияет на цену.

Далее проверим гипотезу о том, что 3 наиболее влиятельные марки оказывают одинаковое влияние на цену:

H0: philips=panasonic=remington=0

```
( 1)  philips - panasonic = 0
( 2)  philips - remington = 0
( 3)  philips = 0

       F(  3,   217) =   14.54
            Prob > F =    0.0000
```

Аналогично и в этом случае мы отвергаем нулевую гипотезу и говорим, что и 3 наиболее влиятельные марки электробритв оказывают разное влияние на цену товара.

Рейтинг силы марок:

Variable	Coef.	[95% Conf.Int.]	
philips	1817.513	-283.851	3918.877
panasonic	69.1155	-1927.24	2065.475
remington	-3077.11	-4352.48	-1801.73
vitek	-4840.99	-5929.27	-3752.71
berdsk	-4966.92	-6043.83	-3890.01
other	-5077.32	-6148.2	-4006.44
ladomir	-5403.81	-6480.72	-4326.9

Коэффициенты показывают, что если бритва принадлежит определённой марке, то это автоматически изменяет цену на этот коэффициент. У Philips наибольший значимый (на 10% уровне значимости) положительный коэффициент, то есть характеристики бритвы объясняются за счет марки и надбавка за бренд Philips составляет 1817 рублей. У Panasonic коэффициент оказался не значимым (не выраженное влияние марки, либо его отсутствие). У остальных фирм наблюдается обратная зависимость от бренда (здесь скорее всего другая интерпретация: этот товар определенной фирмы, значит он дешевле (возможно и качественней)).

7. Расчет индексов цен.

7.1. Стабильность коэффициентов при характеристиках.

Разделим на группы цены за август и январь (group1 – январь, group2 – август).

Сравним модели по группам с базовой моделью:

	(group1) ln_price_jan	(group2) ln_price_jan	Базовая ln_price_jan
head	0.145**	0.513	0.163***
	(3.16)	(1.70)	(3.53)
time_charge	-0.00154***	-0.00154	-0.00161***
	(-5.93)	(-1.04)	(-6.26)
voltage_sw	0.776***	0.782	0.773***
	(9.00)	(0.96)	(8.65)
display	0.640***	-0.141	0.546***
	(4.44)	(-0.32)	(4.11)
power_2	0.913***	2.364***	1.006***
	(8.06)	(7.60)	(8.31)
power_3	0.912***	0	1.000***
	(7.47)	(.)	(7.78)
_cons	6.650***	4.491***	6.556***
	(46.43)	(4.70)	(43.83)
N	201	21	222

adj. R^2	0.657	0.631	0.659
AIC	339.9	50.85	391.9
BIC	363.0	56.07	415.8

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Здесь мы видим расхождения в модели по августу. Это подтверждается тестом (до попарного сравнения коэффициентов power_3. гипотеза о том, что нет различий между коэффициентами для января и августа, принималась, а после – отвергается). Это происходит из-за коррелированности последней и предыдущей (power_2.) переменных, поэтому, рассуждая вне данной модели, гипотеза о равенстве коэффициентов регрессии не отвергается (по сути ценность характеристик тогда не меняется).

```
( 1)  head1 - head2 = 0
( 2)  time_charge1 - time_charge2 = 0
( 3)  voltage_sw1 - voltage_sw2 = 0
( 4)  display1 - display2 = 0
( 5)  power_21 - power_22 = 0

       F(  5,   209) =    1.95
            Prob > F =    0.0876
```

```
( 1)  head1 - head2 = 0
( 2)  time_charge1 - time_charge2 = 0
( 3)  voltage_sw1 - voltage_sw2 = 0
( 4)  display1 - display2 = 0
( 5)  power_21 - power_22 = 0
( 6)  power_31 - o.power_32 = 0

       F(  6,   209) =    8.60
            Prob > F =    0.0000
```

7.2. Расчет простого индекса цен

Индекс получился равный: 1.3403.

7.3. Расчет индекса цен методом прямого сопоставления

Индекс получился равный: 1.2933.

Полученный индекс меньше предыдущего: возможно за счет более менее дорогих (относительно среднего) новинок повлияло на увеличение этого индекса (с учетом новых моделей).

7.4. Гедонический индекс цен.

Индекс получился равный: 1.2670.

Полученный индекс получился еще меньше за счет добавления предсказанных значений по новинкам, которые еще не были выпущены в августе (приближая тем самым цены к августовским, нормируя относительно среднего, которое за счет предсказанных значений увеличилось).

7.5. Доверительный интервал для гедонического индекса цен

Распишем гедонический индекс цен (x – цена января, y – цена августа):

$\frac{\sqrt[n]{x_1 \cdot \ldots \cdot x_n}}{\sqrt[n]{y_1 \cdot \ldots \cdot y_n}} = \sqrt[n]{\frac{x_1}{y_1} \cdot \ldots \cdot \frac{x_n}{y_n}}$, возьмем логарифм этой величины: $\frac{1}{n}\left(\ln \frac{x_1}{y_1} + \ldots + \ln \frac{x_n}{y_n} \right)$, теперь это напоминает формулу доверительного интервала для среднего значения только вместо переменной у нас стоит натуральный логарифм отношения цен. Рассчитав стандартное отклонение и границы, не забудем возвести их в степень экспоненты (поскольку брали логарифм).

Полученный интервал: Индекс $\in [1.218; 1.318]$ с доверительной вероятностью 95%.

8. Выводы

Анализ рынка электробритв показал, что, несмотря на то, что в большинстве исследований говорится об отсутствии влияния марки продаваемого товара на цену, в моем исследовании удалось выяснить некоторую обратную тенденцию. Действительно это влияние может быть не заметно с первого взгляда, но анализ показал, что бренды электробритв влияют на цену неоднородно (некоторые завышено, некоторые занижено). В силу того, что для одних производителей бренд – это способ произвести наценку на свой товар (например, Philips), другие производители используют бренд в качестве привлечения покупателя не к бренду, а к соотношению «цена-качество», относящееся к этому бренду.

Естественно существует множество и видимых характеристик, влияющих на цену товара. Цена повышается, если бритва и работает дольше и заряжается меньше, имеет более обтекаемую форму бритвенных головок (чем больше, тем лучше), водонепроницаемый корпус. Кроме того, следует отметить, что просматривается и влияние режима работы бритвы. Хотя на сегодня бритвы, работающие только от сети, представлены в меньшинстве, не следует сбрасывать их со счетов – ведь это и показатель надежности (хотя с удобством – здесь довольно спорный вопрос). Дело в том, что анализ показал, что включение такого фактора как режим работы помогает чуть лучше объяснить цену, но это не является ключевым фактором, так как в значительной степени этот показатель соотносится со временем работы. Здесь может случиться неразбериха, но дело в том, что преобладающее количество бритв на рынке работают автономно: в принципе их можно выделять и в отдельный кластер, но, если оценивать весь рынок с позиции гомогенности – не следует опираться только на режим работы электробритвы. Есть и другие переменные, которые в состоянии объяснить значительную долю дисперсии цены. Например, имеет ли бритва автоматическое переключение напряжения (отсылает потребителя к безопасности) или наличие дисплея (статусность). Эти переменные с очевидностью влияют на повышение цены не только с позиции проведенного анализа, но и с интуитивной точки зрения.

Наконец отношение цен, сложившееся на рынке на самом деле дает нам гарантию достаточно точного прогноза цены на определенный товар в последующие периоды. Также можно предположить и зависимость данных цен от характеристик, предсказав тем самым цену на новый товар определенного товаропроизводителя с заданными техническими характеристиками. Кроме того, обнаружился и интересный факт, отсылающий нас к влиянию бренда на цену. Те производители, чьи товары переоценены производителем оказываются, недооценены рынком, а товары, недооцененные, наоборот: переоценены. Этот эффект помогает сгладить цены, поскольку рыночный механизм оказывается склонен «не доверять» тем ценам, которые установились (нет объективных причин за счет характеристик к повышению или занижению цены).