ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Образовательная программа «Программная инженерия»

Практическое домашнее задание по теме:

«Прикладные методы математической статистики»

Подготовили студенты группы БПИ204:
Сидоренков Олег
Сушкова Дарья
Шерстюгина Анастасия

Москва, 2022

ОГЛАВЛЕНИЕ

ОТБОР ДАННЫХ

Для работы был выбран следующий датасет:

[Источник]

В нем представлены наблюдения о московских квартирах. Выберем 500 случайных двухкомнатных квартир для анализа.

```
ds = ds[ds['rooms'] == 2]
ds = ds.sample(n=500, random_state=42)
ds = ds.set_index(np.arange(0, 500))
ds
```

	price	repair	year_built_empty	house_age	closest_subway	dist_to_subway	sq	lat	lon	subway_dist_to_center	...	subway
0	15900000	1	0	0	25	1471	0.0	55.853728	37.620986	10400	...	
1	10400000	2	0	44	5	795	0.0	55.896822	37.574348	16430	...	
2	15000000	4	0	92	191	623	0.0	55.713659	37.611763	3744	...	
3	10600000	2	0	62	80	1038	0.0	55.696875	37.739932	10096	...	
4	10200000	2	0	42	9	1532	0.0	55.868862	37.639984	13519	...	
...	...	...	...	...	...	...	...	...	...	...	...	
495	14900000	3	0	120	125	1114	0.0	55.722274	37.647450	2771	...	
496	14000000	2	0	62	3	515	0.0	55.804557	37.632791	6348	...	
497	14500000	3	0	3	9	896	0.0	55.875152	37.676250	13519	...	
498	8600000	2	0	40	130	1817	2.0	55.645078	37.717586	11933	...	
499	8500000	3	0	59	81	1612	0.0	55.698280	37.785620	10612	...	

500 rows × 22 columns Open in new tab

Из имеющихся в датасете переменных в качестве регрессоров были выбраны следующие переменные:

Описание данных

price – стоимость квартиры (в рублях)

houseAge – возраст дома (в годах)

distToSubway – расстояние от дома до метро (в метрах)

subwayDistToCenter – расстояние от метро до центра (в метрах)

footage – площадь квартиры (в квадратных метрах)

maxFloor – количество этажей в доме

floor – этаж

firstFloor – фиктивная переменная, принимающая значение:

$$\begin{cases} 0, \text{если квартира расположена не на 1 этаже} \\ 1, \text{если квартира расположена на 1 этаже} \end{cases}$$

Имеется $n = 500$ наблюдений.

ПРОБЛЕМА МУЛЬТИКОЛЛИНЕАРНОСТИ

В первую очередь, нужно проверить коррелированность переменных в нашей множественной регрессии. Строгая мультиколлинеарность в модели отсутствует (так нет явного линейного выражения одного параметра через другой), однако нужно проверить частичную мультиколлинеарность. Чтобы такой проблемы в модели не возникло, необходимо:

1) Чтобы парные коэффициенты корреляции между регрессорами по абсолютно величине не были большими (> 0.9);
2) Чтобы коэффициенты VIF были не больше 10.

Матрица корреляции

Построим в Excel матрицу корреляции:

	price	*house_age*	*dist_to_subway*	*subway_dist_to_center*	*footage*	*max_floor*	*floor*	*first_floor*
price	1							
house_age	-0,28426	1						
dist_to_subway	-0,18877	-0,110814	1					
subway_dist_to_center	-0,3684	-0,479763	0,154700968	1				
footage	0,617725	-0,521996	0,001019976	0,181140188	1			
max_floor	0,363811	-0,765371	0,027178897	0,332173567	0,49011	1		
floor	0,182631	-0,488244	0,04000917	0,245564214	0,280554	0,631582	1	
first_floor	-0,12481	0,2681183	-0,128116846	-0,08637673	-0,03232	-0,22008	-0,35256	1

Можно сделать следующие выводы о парных коэффициентах корреляции между регрессорами:

- В большей степени на стоимость квартиры оказывают влияние площадь и количество этажей в доме;
- Чем старше дом, в котором расположена квартира, тем он менее высотный, а также с более маленькими по площади квартирами. Это действительно так, потому что в течение последних лет в основном строятся жилые комплексы с большим числом этажей;
- Дома более ранних построек расположены у станций метро, находящихся преимущественно возле центра. Действительно, центр Москвы в основном заполнен старыми домами, а окраина – новостройками;
- В новых домах квартиры имеют большую площадь;
- Чем больше в доме этажей, тем больше вероятность, что квартира расположена на высоких этажах. Именно поэтому наблюдается положительная корреляция между max_floor и floor.

Поскольку не наблюдается больших (> 0.9) парных коэффициентов корреляции между регрессорами, можно говорить об отсутствии существенной мультиколлинеарности.

Коэффициенты VIF

Дополнительно проверим отсутствие мультиколлинеарности, вычислив VIF коэффициенты.

Коэффициент VIF, соответствующий регрессору houseAge

Проведем соответствующий регрессионный анализ в Excel:

ВЫВОД ИТОГОВ								
Регрессионная статистика								
Множественный R	0,83074065		VIF					
R-квадрат	0,690130027		3,227160063					
Нормированный R-	0,685721308							
Стандартная ошиб	11,9690203							
Наблюдения	500							
Дисперсионный анализ								
	df	*SS*	*MS*	*F*	*Значимость F*			
Регрессия	7	156976,1741	22425,16773	156,5375358	7,8147E-121			
Остаток	492	70482,66388	143,2574469					
Итого	499	227458,838						
	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	104,9347712	3,77205576	27,81898729	4,9606E-103	97,52344605	112,3460964	97,52344605	112,3460964
price	-9,08969E-07	3,29239E-07	-2,760817498	0,005981054	-1,55586E-06	-2,62081E-07	-1,55586E-06	-2,62081E-07
dist_to_subway	-0,002422469	0,001071536	-2,260744816	0,024211152	-0,00452782	-0,000317118	-0,00452782	-0,000317118
subway_dist_to_cen	-0,001368153	0,000161034	-8,496044575	2,34731E-16	-0,001684553	-0,001051754	-0,001684553	-0,001051754
footage	-0,283474384	0,091344964	-3,103338946	0,002023612	-0,462948727	-0,10400004	-0,462948727	-0,10400004
max_floor	-1,857982036	0,128860447	-14,41855956	1,45419E-39	-2,111166702	-1,60479737	-2,111166702	-1,60479737
floor	0,133286876	0,133231896	1,000412667	0,317602606	-0,1284868	0,395060553	-0,1284868	0,395060553
first_floor	7,483203525	1,991613917	3,757356514	0,0001923	3,570085767	11,39632128	3,570085767	11,39632128

$$VIF = \frac{1}{1 - R^2} = 3.2272$$

Коэффициент VIF, соответствующий регрессору distToSubway

Проведем соответствующий регрессионный анализ в Excel:

ВЫВОД ИТОГОВ								
Регрессионная статистика								
Множественный R	0,309679494		VIF					
R-квадрат	0,095901389		1,106074037					
Нормированный R-квадрат	0,083038198							
Стандартная ошибка	500,9855649							
Наблюдения	500							
Дисперсионный анализ								
	df	*SS*	*MS*	*F*	*Значимость F*			
Регрессия	7	13098592,28	1871227,469	7,455489433	1,54481E-08			
Остаток	492	123485375,8	250986,5362					
Итого	499	136583968,1						
	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	1570,764416	243,1546589	6,459939624	2,52053E-10	1093,014785	2048,514047	1093,014785	2048,514047
price	-7,08324E-05	1,35151E-05	-5,240984174	2,37434E-07	-9,73869E-05	-4,4278E-05	-9,7387E-05	-4,4278E-05
house_age	-4,244157385	1,877327045	-2,260744816	0,024211152	-7,9327246	-0,555590171	-7,9327246	-0,555590171
subway_dist_to_center	-0,011300807	0,007199914	-1,569575312	0,11715682	-0,025447178	0,002845565	-0,02544718	0,002845565
footage	11,12253507	3,827949049	2,9056121	0,003830489	3,601390904	18,64367924	3,601390904	18,64367924
max_floor	-2,882187146	6,4317845	-0,448116249	0,654266564	-15,51934032	9,754966032	-15,5193403	9,754966032
floor	-3,227327499	5,580441004	-0,578328397	0,563307045	-14,19176323	7,73710823	-14,1917632	7,73710823
first_floor	-270,0445957	83,66915812	-3,22752854	0,00133193	-434,4375364	-105,6516551	-434,437536	-105,6516551

$$VIF = \frac{1}{1 - R^2} = 1.1061$$

Коэффициент VIF, соответствующий регрессору subwayDistToCenter

Проведем соответствующий регрессионный анализ в Excel:

ВЫВОД ИТОГОВ								
Регрессионная статистика								
Множественный R	0,766472377		VIF					
R-квадрат	0,587479904		2,424124329					
Нормированный R-квадрат	0,581610716							
Стандартная ошибка	3129,181687							
Наблюдения	500							
Дисперсионный анализ								
	df	*SS*	*MS*	*F*	*Значимость F*			
Регрессия	7	6860796984	980113854,8	100,0955957	1,94322E-90			
Остаток	492	4817554789	9791778,027					
Итого	499	11678351773						
	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	19402,25939	1318,001755	14,72096628	6,60214E-41	16812,65303	21991,86576	16812,65303	21991,86576
price	-0,001285879	6,45226E-05	-19,92912335	3,18286E-65	-0,001412653	-0,001159105	-0,001412653	-0,001159105
dist_to_subway	-0,440880183	0,280891385	-1,569575312	0,11715682	-0,992774832	0,111014467	-0,992774832	0,111014467
house_age	-93,51451977	11,00683017	-8,496044575	2,34731E-16	-115,1407106	-71,88832896	-115,1407106	-71,88832896
footage	208,7860296	22,20081605	9,404430412	1,98621E-19	165,1659249	252,4061342	165,1659249	252,4061342
max_floor	73,13779884	40,0459376	1,82634752	0,068403764	-5,54435346	151,8199511	-5,54435346	151,8199511
floor	-0,234013662	34,86756636	-0,0067115	0,994647759	-68,74171585	68,27368852	-68,74171585	68,27368852
first_floor	-740,2017534	527,0499279	-1,404424352	0,160823492	-1775,748058	295,3445513	-1775,748058	295,3445513

$$VIF = \frac{1}{1 - R^2} = 2.4241$$

Коэффициент VIF, соответствующий регрессору footage

Проведем соответствующий регрессионный анализ в Excel:

ВЫВОД ИТОГОВ								
Регрессионная статистика								
Множественный R	0,780352061		VIF					
R-квадрат	0,608949339		2,557213423					
Нормированный R-квадрат	0,60338561							
Стандартная ошибка	5,850351176							
Наблюдения	500							
Дисперсионный анализ								
	df	*SS*	*MS*	*F*	*Значимость F*			
Регрессия	7	26222,68231	3746,097473	109,4498577	4,14053E-96			
Остаток	492	16839,49157	34,22660889					
Итого	499	43062,17388						
	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	15,96836171	2,868494564	5,566809124	4,27134E-08	10,33235118	21,60437225	10,33235118	21,60437225
price	2,3634E-06	1,22255E-07	19,33169746	2,32403E-62	2,1232E-06	2,60361E-06	2,1232E-06	2,60361E-06
house_age	-0,067726789	0,021823845	-3,103338946	0,002023612	-0,110606222	-0,024847355	-0,110606222	-0,024847355
dist_to_subway	0,001516761	0,000522011	2,9056121	0,003830489	0,000491116	0,002542407	0,000491116	0,002542407
subway_dist_to_center	0,0007298	7,76017E-05	9,404430412	1,98621E-19	0,000577328	0,000882271	0,000577328	0,000882271
max_floor	-0,029877047	0,075111597	-0,397768767	0,690973168	-0,177456115	0,11770202	-0,177456115	0,11770202
floor	0,066390269	0,065120024	1,019506202	0,30846376	-0,061557383	0,19433792	-0,061557383	0,19433792
first_floor	4,866326199	0,962668291	5,055039463	6,07989E-07	2,974878087	6,757774311	2,974878087	6,757774311

$$VIF = \frac{1}{1 - R^2} = 2.5572$$

Коэффициент VIF, соответствующий регрессору maxFloor

Проведем соответствующий регрессионный анализ в Excel:

ВЫВОД ИТОГОВ								
Регрессионная статистика								
Множественный R	0,836180068		VIF					
R-квадрат	0,699197106		3,324436098					
Нормированный R-квадрат	0,69491739							
Стандартная ошибка	3,5109332							
Наблюдения	500							
Дисперсионный анализ								
	df	*SS*	*MS*	*F*	*Значимость F*			
Регрессия	7	14097,03725	2013,862464	163,3746515	5,4228E-124			
Остаток	492	6064,712752	12,32665194					
Итого	499	20161,75						
	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	10,78246889	1,706969004	6,316733855	5,9832E-10	7,42862071	14,13631708	7,42862071	14,13631708
price	4,10285E-07	9,55486E-08	4,293997028	2,11484E-05	2,22552E-07	5,98019E-07	2,22552E-07	5,98019E-07
house_age	-0,159870906	0,011087856	-14,41855956	1,45419E-39	-0,181656296	-0,138085517	-0,181656296	-0,138085517
dist_to_subway	-0,000141552	0,000315883	-0,448116249	0,654266564	-0,000762198	0,000479094	-0,000762198	0,000479094
subway_dist_to_center	9,20716E-05	5,04129E-05	1,82634752	0,068403764	-6,97966E-06	0,000191123	-6,97966E-06	0,000191123
footage	-0,010760165	0,027051307	-0,397768767	0,690973168	-0,063910502	0,042390172	-0,063910502	0,042390172
floor	0,407999877	0,034527323	11,81672498	1,54584E-28	0,340160685	0,47583907	0,340160685	0,47583907
first_floor	1,593246234	0,58816291	2,708851929	0,006987001	0,437625303	2,748867165	0,437625303	2,748867165

$$VIF = \frac{1}{1 - R^2} = 3.3244$$

Коэффициент VIF, соответствующий регрессору floor

Проведем соответствующий регрессионный анализ в Excel:

ВЫВОД ИТОГОВ								
Регрессионная статистика								
Множественный R	0,67301462		VIF					
R-квадрат	0,452948679		1,827982058					
Нормированный R-квадра	0,445165428							
Стандартная ошибка	4,046006466							
Наблюдения	500							
Дисперсионный анализ								
	df	*SS*	*MS*	*F*	*Значимость F*			
Регрессия	7	6668,669186	952,6670265	58,19531038	1,45813E-60			
Остаток	492	8054,122814	16,37016832					
Итого	499	14722,792						
	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	0,449684411	2,045225548	0,219870327	0,826063342	-3,568769338	4,46813816	-3,568769338	4,46813816
price	-1,82133E-07	1,11854E-07	-1,628310948	0,104099103	-4,01903E-07	3,76373E-08	-4,01903E-07	3,76373E-08
house_age	0,015230821	0,015224539	1,000412667	0,317602606	-0,014682312	0,045143955	-0,014682312	0,045143955
dist_to_subway	-0,000210497	0,000363975	-0,578328397	0,563307045	-0,000925634	0,00050464	-0,000925634	0,00050464
subway_dist_to_center	-3,91231E-07	5,82926E-05	-0,0067115	0,994647759	-0,000114924	0,000114142	-0,000114924	0,000114142
footage	0,031753653	0,031146111	1,019506202	0,30846376	-0,029442143	0,09294945	-0,029442143	0,09294945
max_floor	0,541836234	0,045853334	11,81672498	1,54584E-28	0,451743725	0,631928743	0,451743725	0,631928743
first_floor	-4,421585961	0,65309129	-6,770241815	3,6728E-11	-5,704778	-3,138393921	-5,704778	-3,138393921

$$VIF = \frac{1}{1 - R^2} = 1.8280$$

Коэффициент VIF, соответствующий регрессору firstFloor

Проведем соответствующий регрессионный анализ в Excel:

ВЫВОД ИТОГОВ								
Регрессионная статистика								
Множественный R	0,453474519		VIF					
R-квадрат	0,20563914		1,258873706					
Нормированный R-квадрат	0,194337257							
Стандартная ошибка	0,267133083							
Наблюдения	500							
Дисперсионный анализ								
	df	*SS*	*MS*	*F*	*Значимость F*			
Регрессия	7	9,088838691	1,298405527	18,19512331	1,47112E-21			
Остаток	492	35,10916131	0,071360084					
Итого	499	44,198						
	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	-0,05478355	0,135017791	-0,405750603	0,685102252	-0,320066149	0,210499049	-0,320066149	0,210499049
price	-3,12828E-08	7,26934E-09	-4,303392787	2,0302E-05	-4,55656E-08	-1,70001E-08	-4,55656E-08	-1,70001E-08
house_age	0,003727569	0,000992072	3,757356514	0,0001923	0,001778348	0,00567679	0,001778348	0,00567679
dist_to_subway	-7,67786E-05	2,37887E-05	-3,22752854	0,00133193	-0,000123519	-3,00387E-05	-0,000123519	-3,00387E-05
subway_dist_to_center	-5,39441E-06	3,84101E-06	-1,404424352	0,160823492	-1,29412E-05	2,1524E-06	-1,29412E-05	2,1524E-06
footage	0,01014595	0,002007096	5,055039463	6,07989E-07	0,006202413	0,014089486	0,006202413	0,014089486
max_floor	0,009223444	0,003404927	2,708851929	0,006987001	0,002533452	0,015913436	0,002533452	0,015913436
floor	-0,019274374	0,002846925	-6,770241815	3,6728E-11	-0,024868006	-0,013680742	-0,024868006	-0,013680742

$$VIF = \frac{1}{1 - R^2} = 1.2589$$

Вывод

Коэффициенты VIF всех регрессоров оказались меньше 10. Следовательно, существенной мультиколлинеарности в модели не наблюдается.

ЛИНЕЙНАЯ МОДЕЛЬ

Построим следующую регрессионную модель зависимости стоимости квартиры от всех остальных параметров:

$$price_i = \beta_0 + \beta_1 \cdot houseAge_i + \beta_2 \cdot distToSubway_i + \beta_3 \cdot subwayDistToCenter_i + \beta_4 \cdot footage_i + \beta_5 \cdot maxFloor_i + \beta_6 \cdot floor_i + \beta_7 \cdot firstFloor_i + \varepsilon_i,$$
$$i = 1 \dots 500;$$

Проверка значимости регрессии в целом

Уровень значимости в рамках данного Отчета $\alpha = 0.05$.

Тестирование гипотезы $\boldsymbol{H_0}: \begin{cases} \boldsymbol{\beta_1 = 0} \\ \dots \\ \boldsymbol{\beta_7 = 0} \end{cases}$ против гипотезы $\boldsymbol{H_1}: \begin{bmatrix} \boldsymbol{\beta_1 \neq 0} \\ \dots \\ \boldsymbol{\beta_7 \neq 0} \end{bmatrix}$

Для тестирования данной гипотезы используется тестовая статистика:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-k-1}{k};$$

Известно, что при верной гипотезе H_0 данная тестовая статистика имеет распределение:

$$F^{H_0} \sim F(k, n-k-1)$$
$$F^{H_0} \sim F(7, 500-7-1)$$
$$F^{H_0} \sim F(7, 492)$$

Проведем регрессионный анализ в Excel:

ВЫВОД ИТОГОВ					
Регрессионная статистика					
Множественный R	0,836318314				
R-квадрат	0,699428323				
Нормированный R-квадрат	0,695151897				
Стандартная ошибка	1626396,749				
Наблюдения	500				
Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	7	3,0284E+15	4,32629E+14	163,5543965	4,4918E-124
Остаток	492	1,30142E+15	2,64517E+12		
Итого	499	4,32982E+15			

Если в формулу тестовой статистики мы подставим все известные данные, то получим наблюдаемое значение тестовой статистики:

$$F_{\text{набл}} = \frac{0.6994}{1-0.6994} \cdot \frac{492}{7} = 163.5544;$$

Область, в которой H_0 не отвергается:

$$[0; F_{\text{кр}}] = [0; finv(1-\alpha, k, n-k-1)] = [0; finv(0.95,7,492)] = [0; 2.0282]$$

Поскольку $F_{\text{набл}} \notin [0; F_{\text{кр}}]$, гипотеза H_0 отвергается в пользу альтернативной. Следовательно, регрессия значима в целом.

Алгоритм для проверки значимости переменных

Уровень значимости $\alpha = 0.05$.

Тестирование гипотезы $\boldsymbol{H_0}: \boldsymbol{\beta_i = b}$ против гипотезы $\boldsymbol{H_1}: \boldsymbol{\beta_i \neq b}$

Для тестирования данной гипотезы используется тестовая статистика:

$$T = \frac{\widehat{\beta}_i - b}{\sqrt{\widehat{D}(\widehat{\beta}_i)}};$$

Известно, что при верной гипотезе H_0 данная тестовая статистика имеет распределение:

$$T^{H_0} \sim t(n-k-1)$$
$$T^{H_0} \sim t(492)$$

Область, в которой H_0 не отвергается:

$$[-T_{\text{кр}}; T_{\text{кр}}] = \left[tinv\left(\frac{\alpha}{2}, n-k-1\right); tinv\left(1-\frac{\alpha}{2}, n-k-1\right)\right]$$
$$= [tinv(0.025,492); tinv(0.975,492)] = [-1.9648;\ 1.9648]$$

Если $T_{\text{набл}} \notin [-T_{\text{кр}}; T_{\text{кр}}]$, то гипотеза H_0 отвергается в пользу альтернативной.

Если $T_{\text{набл}} \in [-T_{\text{кр}}; T_{\text{кр}}]$, то гипотеза H_0 принимается.

$$p - value(T_{\text{набл}}) = 2tcdf(-|T_{\text{набл}}|, n-k-1) = 2tcdf(-|T_{\text{набл}}|, 492)$$

Если $p - value < \alpha$, то гипотеза H_0 отвергается в пользу альтернативной.

Если $p - value > \alpha$, то гипотеза H_0 принимается.

Проверка значимости переменных

С помощью регрессионного анализа, проведенного в Excel, оценим значимость переменных по $t-$статистике и $p - value$.

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	7022680,301	758766,2485	9,255393627	6,5595E-19	5531858,382	8513502,219	5531858,382	8513502,219
house_age	-16783,58634	6079,208911	-2,760817498	0,005981054	-28727,99993	-4839,172744	-28727,99993	-4839,172744
dist_to_subway	-746,5085205	142,4367057	-5,240984174	2,37434E-07	-1026,367783	-466,6492578	-1026,367783	-466,6492578
subway_dist_to_center	-347,3693713	17,43023841	-19,92912335	3,18286E-65	-381,6162576	-313,1224849	-381,6162576	-313,1224849
footage	182653,1637	9448,376897	19,33169746	2,32403E-62	164089,0178	201217,3096	164089,0178	201217,3096
max_floor	88042,83059	20503,70087	4,293997028	2,11484E-05	47757,21334	128328,4478	47757,21334	128328,4478
floor	-29429,82287	18073,83467	-1,628310948	0,104099103	-64941,24544	6081,599704	-64941,24544	6081,599704
first_floor	-1159588,163	269459,0572	-4,303392787	2,0302E-05	-1689020,606	-630155,72	-1689020,606	-630155,72

Запишем получившееся уравнение:

$$\widehat{price_i} = 7022680.3 - 16783.586 \cdot houseAge_i - 746.5 \cdot distToSubway_i - 347.37$$
$$\cdot\ subwayDistToCenter_i + 182653.16 \cdot footage_i + 88042.83 \cdot maxFloor_i$$
$$- 29429.82 \cdot floor_i - 1159588.16 \cdot firstFloor_i + \varepsilon_i, \qquad i = 1 \dots 500;$$

Для houseAge

$-2.7608 \notin [-1.9648;\ 1.9648]$, гипотеза H_0 отвергается в пользу альтернативной.

$0.006 < 0.05$, гипотеза H_0 отвергается в пользу альтернативной.

Следовательно, переменная $houseAge$ является *значимой*.

Интерпретация: при увеличении возраста дома на 1 год цена квартиры снижается на 16783.7 рубля.

Для distToSubway

$-5.2410 \notin [-1.9648;\ 1.9648]$, гипотеза H_0 отвергается в пользу альтернативной.

$2.3743\text{E}-07 < 0.05$, гипотеза H_0 отвергается в пользу альтернативной.

Следовательно, переменная $distToSubway$ является *значимой*.

Интерпретация: при увеличении расстояния от дома до ближайшей станции метро на 1 метр цена квартиры снижается на 746.5 рубля.

Для subwayDistToCenter

$-19.9291 \notin [-1.9648;\ 1.9648]$, гипотеза H_0 отвергается в пользу альтернативной.

$3.1829\text{E}-65 < 0.05$, гипотеза H_0 отвергается в пользу альтернативной.

Следовательно, переменная $subwayDistToCenter$ является *значимой*.

Интерпретация: при увеличении расстояния от ближайшей станции метро до центра города на 1 метр цена квартиры снижается на 347.37 рубля.

Для footage

$-19.3317 \notin [-1.9648;\ 1.9648]$, гипотеза H_0 отвергается в пользу альтернативной.

$2.3249\text{E}-62 < 0.05$, гипотеза H_0 отвергается в пользу альтернативной.

Следовательно, переменная $footage$ является *значимой*.

Интерпретация: при увеличении площади квартиры на 1 кв. метр цена квартиры повышается на 182653.16 рубля.

Для maxFloor

$-4.2940 \notin [-1.9648;\ 1.9648]$, гипотеза H_0 отвергается в пользу альтернативной.

$2.1148\text{E}-05 < 0.05$, гипотеза H_0 отвергается в пользу альтернативной.

Следовательно, переменная $maxFloor$ является *значимой*.

Интерпретация: при увеличении количества этажей в доме на 1 цена квартиры повышается на 88042.83 рубля.

Для floor

$-1.6283 \in [-1.9648;\ 1.9648]$, гипотеза H_0 принимается.

$0.1041 > 0.05$, гипотеза H_0 принимается.

Следовательно, переменная $floor$ *не является значимой*.

Интерпретация: при увеличении этажа, на котором расположена квартира, на 1 цена квартиры снижается на 29429.82 рубля.

Для firstFloor

$-4.3034 \notin [-1.9648;\ 1.9648]$, гипотеза H_0 отвергается в пользу альтернативной.

$2.03020\text{E}-05 < 0.05$, гипотеза H_0 отвергается в пользу альтернативной.

Следовательно, переменная $firstFloor$ является *значимой*.

Интерпретация: если квартира расположена на первом этаже, то её стоимость снижается 1159588.163 рубля.

Вывод

Заметим, что в данной линейной модели переменная $floor$ является незначимой. Это означает, что этаж, на котором расположена квартира, слабо влияет на ее стоимость. Но в реальности такая ситуация скорее невозможна. Поэтому выдвинем предположение, что линейная модель несовершенна. Стоит также рассмотреть нелинейную модель.

НЕЛИНЕЙНАЯ МОДЕЛЬ

В качестве нелинейной выбрана *двойная логарифмическая модель*, которая была получена путем преобразования исходной линейной модели, в результате чего есть следующее уравнение:

$$\ln price_i = \beta_0 + \beta_1 \cdot \ln houseAge_i + \beta_2 \cdot \ln distToSubway_i + \beta_3 \cdot \ln subwayDistToCenter_i + \beta_4 \cdot \ln footage_i + \beta_5 \cdot \ln maxFloor_i + \beta_6 \cdot \ln floor_i + \beta_7 \cdot firstFloor_i + \varepsilon_i, \quad i = 1, \dots, 500;$$

Смысл данного преобразования заключается в том, что относительно параметров модель является линейной, то есть к ней можно применять стандартные методы оценивания (единственное существенное отличие заключается в интерпретации полученного результата).

При использовании средств анализа данных Excel были получены следующие данные для нелинейной модели:

ВЫВОД ИТОГОВ									
Регрессионная статистика									
Множественный R	0,836749087								
R-квадрат	0,700149034								
Нормированный R-квадрат	0,695882862								
Стандартная ошибка	0,133847968								
Наблюдения	500								
Дисперсионный анализ									
	df	*SS*	*MS*	*F*	*Значимость F*				
Регрессия	7	20,58134297	2,940191852	164,1164466	2,4949E-124				
Остаток	492	8,814317037	0,017915279						
Итого	499	29,39566							
	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*	
Y-пересечение	15,98284346	0,228686607	69,8897223	1,246E-257	15,53352062	16,4321663	15,53352062	16,4321663	
house_age	-0,018325741	0,008055295	-2,274993048	0,023335773	-0,034152764	-0,002498718	-0,034152764	-0,002498718	
dist_to_subway	-0,04748	0,009706024	-4,891807242	1,35578E-06	-0,066550372	-0,028409629	-0,066550372	-0,028409629	
subway_dist_to_center	-0,275884002	0,013520325	-20,40513034	1,64038E-67	-0,302448701	-0,249319302	-0,302448701	-0,249319302	
footage	0,763440685	0,041473602	18,4078702	5,89003E-58	0,681953461	0,844927908	0,681953461	0,844927908	
max_floor	0,105577879	0,01734452	6,087102811	2,31606E-09	0,071499411	0,139656347	0,071499411	0,139656347	
floor	-0,021284451	0,01051312	-2,024560775	0,043452162	-0,041940603	-0,0006283	-0,041940603	-0,0006283	
first_floor	-0,125594783	0,026204183	-4,79292872	2,18025E-06	-0,177080693	-0,074108873	-0,177080693	-0,074108873	

Запишем полученную модель в виде уравнения:

$$\ln \widehat{price}_i = 15.982 - 0.018 \cdot \ln houseAge_i - 0.047 \cdot \ln distToSubway_i - 0.276 \cdot \ln subwayDistToCenter_i + 0.7634 \cdot \ln footage_i + 0.1056 \cdot \ln maxFloor_i - 0.021 \cdot \ln floor_i - 0.1256 \cdot firstFloor_i, i = 1, \dots, 500;$$

Стоит отметить, что знаки коэффициентов совпадают со знаками, полученными в ходе анализа линейной модели в предыдущем разделе, и не противоречат здравому смыслу.

Проверка значимости регрессии в целом

В рамках Отчета используется уровень значимости $\alpha = 0.05$.

Значимость $F-$статистики равна $2{,}4949E-124 < 0.05$, значит, регрессия *значима в целом*.

Также можно проверить с помощью тестовой статистики:

Тестирование гипотезы $\boldsymbol{H_0}$: $\begin{cases} \boldsymbol{\beta_1 = 0} \\ \quad \dots \\ \boldsymbol{\beta_7 = 0} \end{cases}$ против гипотезы $\boldsymbol{H_1}$: $\begin{bmatrix} \boldsymbol{\beta_1 \neq 0} \\ \quad \dots \\ \boldsymbol{\beta_7 \neq 0} \end{bmatrix}$.

Область, в которой H_0 не отвергается:

$$[0; F_{\text{кр}}] = [0; finv(1-\alpha, k, n-k-1)] = [0; finv(0.95,7,492)] = [0; 2.0282];$$

Согласно данным анализа: $F_{\text{набл}} = 164.116;$

$F_{\text{набл}} \notin [0; F_{\text{кр}}]$, гипотеза H_0 отвергается в пользу альтернативной. Следовательно, регрессия *значима в целом*.

Проверка значимости переменных

Для уровня значимости $\alpha = 0.05$ все переменные оказывают значимое влияние, так как соответствующие им $p - value$ меньше заданного уровня значимости (подробнее алгоритм проверки на значимость описан в предыдущем разделе). Проверить это можно и с помощью $t-$статистики, значение которой для каждого коэффициента не принадлежит области, для которой гипотеза о незначимости не отвергается:

$$[-T_{\text{кр}}; T_{\text{кр}}] = \left[tinv\left(\frac{\alpha}{2}, n-k-1\right); tinv\left(1-\frac{\alpha}{2}, n-k-1\right)\right]$$
$$= [tinv(0.025,492); tinv(0.975,492)] = [-1.9648;\ 1.9648];$$

Как видно из анализа в Excel, в данный диапазон не входит ни одна $t-$статистика.

Важно отметить, что переменная $floor$, характеризующая номер этажа, на котором расположена квартира, в нелинейной модели стала также оказывать значимое влияние (в линейной модели такого не наблюдалось).

Полученные результаты можно интерпретировать следующим образом (стоит учитывать, что некоторые параметры принимают только целочисленные значения, поэтому для них интерпретация является условной в рамках данной модели):

- $houseAge$: увеличение возраста дома на 1% снижает цену на квартиру приблизительно на 0.02%;
- $distToSubway$: увеличение расстояния от дома до ближайшей станции метро на 1% снижает цену на квартиру практически на 0.05%;
- $subwayDistToCenter$: увеличение расстояния от ближайшей станции метро до центра города на 1% снижет цену на квартиру на 0.256%;
- $footage$: увеличение площади квартиры на 1% повышает цену на квартиру на 0.763%;
- $maxFloor$: увеличение количества этажей в доме на 1% повышает цену на квартиру на 0.106%;
- $floor$: увеличение этажа, на котором расположена квартира, на 1% снижает цену на квартиру на 0.021%;
- $firstFloor$: при прочих равных условиях, для квартиры на первом этаже цена на квартиру снижается на 0.126%.

Интерпретация для параметров указывается ПРИ ПРОЧИХ РАВНЫХ УСЛОВИЯХ.

Вывод

Переменная $floor$, характеризующая номер этажа, на котором расположена квартира, в нелинейной модели стала оказывать значимое влияние, чего не наблюдалось в линейной. Также в логарифмической модели незначительно, но увеличился коэффициент $R^2 = 0.7$; таким образом, преобразование в нелинейную модель дало небольшие улучшения.

ТЕСТ ЧОУ

Так как в рассматриваемых моделях присутствует фиктивная переменная $firstFloor$, отвечающая за качественный признак – расположена ли квартира на первом этаже или нет – целесообразность её использования нужно проверить с помощью теста Чоу.

Выборку можно разбить на две подвыборки по данному признаку:

A – квартиры, расположенные не на первом этаже;

B – квартиры, расположенные на первом этаже.

Соответственно для первой подвыборки $firstFloor = 0$, для второй $fisrtFloor = 1$.

Для тестирования будет использоваться первый способ, в ходе которого анализируется модель без фиктивной переменной (так как изначально она включена) на основе трех выборок.

$$price_i = \beta_0 + \beta_1 \cdot houseAge_i + \beta_2 \cdot distToSubway_i + \beta_3 \cdot subwayDistToCenter_i + \beta_4 \cdot footage_i + \beta_5 \cdot maxFloor_i + \beta_6 \cdot floor_i + \varepsilon_i, \qquad i = 1 \dots 500;$$

Для линейной модели

1) Оцениваем регрессию для полной выборки и получаем результат:

ВЫВОД ИТОГОВ					
Регрессионная статистика					
Множественный R	0,829526742				
R-квадрат	0,688114616				
Нормированный R-квадрат	0,684318851				
Стандартная ошибка	1655042,206				
Наблюдения	500				
Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	6	2,97941E+15	4,96569E+14	181,2848154	2,7199E-121
Остаток	493	1,35041E+15	2,73916E+12		
Итого	499	4,32982E+15			

Таким образом, $RSS_{all} = 1350408199555290$;

2) Оцениваем регрессии по соответствующим подвыборкам:

Для подвыборки A

ВЫВОД ИТОГОВ					
Регрессионная статистика					
Множественный R	0,827388593				
R-квадрат	0,684571884				
Нормированный R-квадрат	0,680309342				
Стандартная ошибка	1660698,396				
Наблюдения	451				
Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	6	2,65756E+15	4,42927E+14	160,6017881	6,6995E-108
Остаток	444	1,22452E+15	2,75792E+12		
Итого	450	3,88208E+15			

Для подвыборки B

ВЫВОД ИТОГОВ					
Регрессионная статистика					
Множественный R	0,918111317				
R-квадрат	0,842928391				
Нормированный R-квадрат	0,82048959				
Стандартная ошибка	1192579,239				
Наблюдения	49				
Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	6	3,20565E+14	5,34276E+13	37,56566054	2,39943E-15
Остаток	42	5,97343E+13	1,42225E+12		
Итого	48	3,803E+14			

Таким образом, $RSS_A + RSS_B = 1224516108871380 + 59734300140023.6 = 1284250409011403.6$;

3) С помощью теста Чоу проверяем следующую гипотезу (против альтернативной):

$$\mathrm{H}_0: \begin{cases} \beta_0^A = \beta_0^B \\ \dots \\ \beta_6^A = \beta_6^B \end{cases}; \quad H_1: \begin{bmatrix} \beta_0^A \neq \beta_0^B \\ \dots \\ \beta_6^A \neq \beta_6^B \end{bmatrix}.$$

Значение тестовой статистики: $F_{\text{набл}} = \frac{(RSS_{all} - RSS_A - RSS_B)}{RSS_A + RSS_B} \cdot \frac{n-2k-2}{k+1} \sim F(k+1, n-2k-2)$;

$$n = n_A + n_B = 451 + 49 = 500, \qquad k = 6;$$

$$F_{\text{набл}} = \frac{1350408199555290 - 1284250409011403.6}{1284250409011403.6} \cdot \frac{500 - 12 - 2}{7} = 3.57659;$$

$$F_{\text{набл}} \sim F(7,486);$$

Область, для которой основная гипотеза не отвергается:
$[0; F_{\text{кр}}] = [0; finv(0.95, 7,486)] = [0; 2.0284]$; так как $F_{\text{набл}} \notin [0; 2.0284]$, то основная гипотеза отвергается, а это значит, что при данном уровне значимости $\alpha = 0.05$ цена квартиры зависит от того, расположена ли она на первом этаже или нет. Таким образом, включение фиктивной переменной $firstFloor$ оправдано.

Для нелинейной модели

1) Оцениваем регрессию для полной выборки и получаем результат:

ВЫВОД ИТОГОВ					
Регрессионная статистика					
Множественный R	0,828340857				
R-квадрат	0,686148575				
Нормированный R-квадрат	0,682328882				
Стандартная ошибка	0,136798143				
Наблюдения	500				
Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	6	20,16979022	3,361631703	179,6344918	1,2729E-120
Остаток	493	9,225869784	0,018713732		
Итого	499	29,39566			

Таким образом, $RSS_{all} = 9.22586978405432;$

2) Оцениваем регрессии по соответствующим подвыборкам:

Для подвыборки A:

ВЫВОД ИТОГОВ					
Регрессионная статистика					
Множественный R	0,825584504				
R-квадрат	0,681589773				
Нормированный R-квадрат	0,677286932				
Стандартная ошибка	0,136487528				
Наблюдения	451				
Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	6	17,70536815	2,950894692	158,4045956	5,3642E-107
Остаток	444	8,271207273	0,018628845		
Итого	450	25,97657543			

Для подвыборки B:

ВЫВОД ИТОГОВ					
Регрессионная статистика					
Множественный R	0,912586825				
R-квадрат	0,832814712				
Нормированный R-квадрат	0,8089311				
Стандартная ошибка	0,107146209				
Наблюдения	49				
Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	6	2,401890698	0,400315116	34,86971296	8,69558E-15
Остаток	42	0,482173022	0,01148031		
Итого	48	2,88406372			

Таким образом, $RSS_A + RSS_B = 8.27120727250147 + 0.482173022312562 = 8.753380294814032;$

3) С помощью теста Чоу проверяем следующую гипотезу:

$$\mathrm{H}_0: \begin{cases} \beta_0^A = \beta_0^B \\ \dots \\ \beta_k^A = \beta_k^B \end{cases}; \quad H_1: \begin{bmatrix} \beta_0^A \neq \beta_0^B \\ \dots \\ \beta_k^A \neq \beta_k^B \end{bmatrix}.$$

Значение тестовой статистики: $F_{\text{набл}} = \frac{(RSS_{all} - RSS_A - RSS_B)}{RSS_A + RSS_B} \cdot \frac{n-2k-2}{k+1} \sim F(k+1, n-2k-2);$

$$n = n_A + n_B = 451 + 49 = 500, \qquad k = 6;$$

$$F_{\text{набл}} = \frac{9.22586978405432 - 8.753380294814032}{8.753380294814032} \cdot \frac{500 - 12 - 2}{7} = 3.74761;$$

$$F_{\text{набл}} \sim F(7,486);$$

Область, для которой основная гипотеза не отвергается:
$[0; F_{\text{кр}}] = [0; finv(0.95, 7,486)] = [0; 2.0284]$; так как $F_{\text{набл}} \notin [0; 2.0284]$, то основная гипотеза отвергается, а это значит, что при данном уровне значимости $\alpha = 0.05$ цена

квартиры зависит от того, расположена ли она на первом этаже или нет. Таким образом, включение фиктивной переменной $firstFloor$ оправдано.

Вывод

Тесты Чоу показали, что использование в модели (как линейной, так и логарифмической) фиктивной переменной $firstFloor$ оказалось оправдано. Рассмотренные в ходе тестирования подвыборки нельзя оценивать единым образом, то есть признак того, расположена квартира на первом этаже или нет, имеет значение.

Тест Рамсея

Данный тест позволяет протестировать необходимость включения в модель степеней независимых переменных. Проверим для переменных вида x_i^2 для линейной модели.

1) Оценим исходную модель (раздел «Линейная модель»). Это короткая модель (R), имеем $RSS_R = 1301421862054060$;
2) Оценим длинную модель вида:
$$price_i = \beta_0 + \beta_1 \cdot houseAge_i + \beta_2 \cdot distToSubway_i + \beta_3 \cdot subwayDistToCenter_i + \beta_4 \cdot footage_i + \beta_5 \cdot maxFloor_i + \beta_6 \cdot floor_i + \beta_7 \cdot firstFloor_i + \delta \cdot \widehat{price_i^2} + u_i, \quad i = 1 \dots 500;$$
В таком случае $RSS_{UR} = 1298336698203590$;
3) Основная гипотеза: $H_0: \delta = 0,\ H_1: \delta \neq 0$.
4) Рассчитаем значение статистики: $T_{\text{набл}} = \frac{RSS_R - RSS_{UR}}{RSS_{UR}} \cdot (n - k - 2);\ T^{H_0} \sim F(1, n - k - 2)$;
5) Проведем вычисления: $T_{\text{набл}} = \frac{1301421862054060 - 1298336698203590}{1298336698203590} \cdot (500 - 7 - 2) = 1.16673545$;
6) Область, для которой основная гипотеза не отвергается: $[0; T_{\text{кр}}] = [0; finv(0.95, 1{,}491)] = 3.8605$;
7) Так как $T_{\text{набл}} \in [0; 3.8605]$, то основная гипотеза не отвергается, то есть в модели нет пропущенных нелинейных переменных (это супер).

Тест Бокса-Кокса

Данный тест показывает, какую модель лучше выбрать – линейную или логарифмически-линейную.

1) Вычисляем измененное значение зависимой переменной $price$ по следующей формуле:
$$price_i^* = \frac{price_i}{\sqrt[n]{price_1 \cdot \dots \cdot price_n}};\ n = 500.$$
2) Оцениваем модель
$$price_i^* = \beta_0 + \beta_1 \cdot houseAge_i + \beta_2 \cdot distToSubway_i + \beta_3 \cdot subwayDistToCenter_i + \beta_4 \cdot footage_i + \beta_5 \cdot maxFloor_i + \beta_6 \cdot floor_i + \beta_7 \cdot firstFloor_i + \varepsilon_i,\ i = 1 \dots 500;$$
$$RSS_1^* = 9.76015466696509;$$
3) Оцениваем модель
$$\ln price_i^* = \beta_0 + \beta_1 \cdot houseAge_i + \beta_2 \cdot distToSubway_i + \beta_3 \cdot subwayDistToCenter_i + \beta_4 \cdot footage_i + \beta_5 \cdot maxFloor_i + \beta_6 \cdot floor_i + \beta_7 \cdot firstFloor_i + \varepsilon_i, \quad i = 1 \dots 500;$$
$$RSS_2^* = 8.96030982631136;$$
4) Основная гипотеза H_0: модели имеют одинаковое качество;
Альтернативная гипотеза H_1: модели имеют разное качество.
5) Тестовая статистика: $T_{\text{набл}} = \frac{n}{2} \cdot \left|\ln \frac{RSS_2^*}{RSS_1^*}\right| = \frac{500}{2} \cdot \left|\ln \frac{8.96030982631136}{9.76015466696509}\right| = 21.37586$;
$$T^{H_0} \sim \chi^2(1);$$

6) Область, для которой основная гипотеза не отвергается: $\left[0; T_{\text{кр}}\right] = [0; chi2inv(0.95, 1)] = 3.8415$;
7) Так как $T_{\text{набл}} \notin [0; 3.8415]$, то основная гипотеза отвергается, то есть модели имеют разное качество;
8) Так как $RSS_2^* < RSS_1^*$, то модель (2) лучше, чем модель (1). Иначе говоря, в нашем случае логарифмически-линейная модель лучше, чем линейная.