АНАЛИЗ ВЫБОРОЧНЫХ СОВОКУПНОСТЕЙ

ВВЕДЕНИЕ

В 2016 году рынок жилой недвижимости в России был на пике перегрева. Цены на все типы жилья в столице, а также индексы стоимости упали относительно 2015 года, продолжая общий нисходящий тренд. Эксперты сравнивали рынок жилой недвижимости с рынком нефти 2013-2014 годов, когда цены показывали отрицательную динамику в сравнении с предыдущим периодом, при этом производители (в случае рынка жилой недвижимости - застройщики) продолжали наращивание предложения [1]. На конец 2022 году аналитики прогнозируют наступление ситуации на рынке жилой недвижимости, схожей с ситуацией в 2016 году, вследствие окончания действия программы низких ипотечных ставок [2]. Вследствие предположения о том, что рынок недвижимости в среднем стабилен, было решено исследовать регрессоры, влияющие на стоимость квартиры. При этом было решено ориентироваться на крупный город России в силу отсутствия возможных данных для всех показателей во всех городах, к примеру, влияние расстояния квартиры от метро может иметь влияние, однако для городов без метрополитена данное влияние будет неактуально. Следовательно, невозможно в рамках текущей осведомленностью базами данных построить репрезентативную модель по выборке для всей страны. Поэтому было решено построить модели, описывающие влияние на цену квартиры в Москве, и подобрать наиболее оптимальную. Также, помимо этого, было решено сформулировать дополнительны гипотезы, которые помогут наиболее лучше понять ситуацию на рынке жилой недвижимости в Москве.

Постановка задач

1. Определить число и тип наблюдений;
2. Определить, есть ли отсутствующие значения в ячейках;
3. Проанализировать максимальные, медианные, средние и минимальные значения для количественных переменных;
4. Определить наличие выбросов по методу Тьюки для зависимой переменной
5. Определить, существует ли разница в ценах квартир в зависимости от географии расположения;
6. Построить диаграмму рассеяния и оценить коэффициенты корреляции между переменными
7. Построить первоначальную модель, оценить валидность модели. Найти наиболее оптимальную и интерпретируемую модель

Выбор и описание данных

После формулирования гипотез следует второй наиболее трудозатратный этап - поиск данных. Под наши гипотезы соответственно необходимыми параметрами являются цены на жилье в Москве, а также возможные регрессоры, которые описывали бы переменную-отклик. Для работы с гипотезами были выбраны данные с официального портала открытых данных правительства Москвы [3] за 2016 год. Данный датасет содержит информацию о следующих параметрах:

1. n – номер квартиры по порядку
2. price – цена квартиры в $1000
3. totsp – общая площадь квартиры, кв.м.
4. livesp жилая площадь квартиры, кв.м.
5. kitsp – площадь кухни, кв.м.
6. dist – расстояние от центра в км.

7. metrdist – расстояние до метро в минутах
8. walk – 1 – пешком от метро, 0 – на транспорте
9. brick 1 – кирпичный, монолит ж/б, 0 – другой
10. floor 1 – этаж кроме первого и последнего, 0 – иначе.
11. code – число от 1 до 8, при помощи которого мы группируем наблюдения по подвыборкам:
    - 11.1. Наблюдения сгруппированы на севере, вокруг Калужско-Рижской линии метрополитена
    - 11.2. Север, вокруг Серпуховско-Тимирязевской линии метрополитена
    - 11.3. Северо-запад, вокруг Замоскворецкой линии метрополитена
    - 11.4. Северо-запад, вокруг Таганско-Краснопресненской линии метрополитена
    - 11.5. Юго-восток, вокруг Люблинской линии метрополитена
    - 11.6. Юго-восток, вокруг Таганско-Краснопресненской линии метрополитена
    - 11.7. Восток, вокруг Калининской линии метрополитена
    - 11.8. Восток, вокруг Арбатско-Покровской линии метрополитена

Этап анализа датасета

```
# Установим директорию
setwd("C:/Users/User/Desktop")

# Присвоим данным переменную "flats".
flats <- read.table('flats_moscow.txt', sep='\t', header=T)

# Удалим 1 столбец в дата фрейме, поскольку он содержит лишь нумерацию данных по порядку.
flats <- flats[,-1]
```

Задание 1

Выборка имеет 2040 наблюдений и 10 переменных, при этом исходя из описания данных, можно заметить, что 4 переменные: walk, brick, floor, code являются факторными. С помощью функции string было выявлено, что исходный тип ранее упомянутых переменных является целочисленное значение (integer). Для дальнейшего исследования корректно трансформировать данные переменные в факторные.

Задание 2

```
##   price   totsp   livesp   kitsp    dist metrdist    walk
##       0       0       0       0       0       0       0
##   brick   floor    code
##       0       0       0
```

Пропущенных значений в ячейках не наблюдается.

Задание 3

```
##      price          totsp           livesp          kitsp
##  Min.   : 50   Min.   : 44.0   Min.   : 28.0   Min.   : 5.0
##  1st Qu.: 95   1st Qu.: 62.0   1st Qu.: 42.0   1st Qu.: 7.0
##  Median :115   Median : 73.5   Median : 45.0   Median : 9.0
##  Mean   :127   Mean   : 73.1   Mean   : 46.3   Mean   : 8.9
##  3rd Qu.:142   3rd Qu.: 79.0   3rd Qu.: 50.0   3rd Qu.:10.0
##  Max.   :730   Max.   :192.0   Max.   :102.0   Max.   :25.0
##       dist          metrdist
##  Min.   : 3.0   Min.   : 1.00
##  1st Qu.: 9.0   1st Qu.: 5.00
##  Median :12.0   Median : 7.00
##  Mean   :11.0   Mean   : 8.12
##  3rd Qu.:13.5   3rd Qu.:10.00
##  Max.   :17.0   Max.   :20.00
```

Можно заметить, что практически для всех количественных переменных справедливо можно отметить равенство средних и медианных значений, что наталкивает на мысль о нормальности распределения этих параметров, за исключением переменной price. Для

переменной price также можно заметить, что 75% квартир находятся примерно в одном ценовом сегменты, находясь между значениями в 50 тыс. и 142 тыс; при этом 25% квартир имеют цену от 142 тыс. до 730 тыс., что предварительно говорит либо о высокоценовом районе расположения квартир с наибольшей стоимостью, либо о влияния определенных параметров на общую цены квартиры в Москве, либо об этих предпосылках в совокупности. Для остальных переменных сильных различий не замечено.

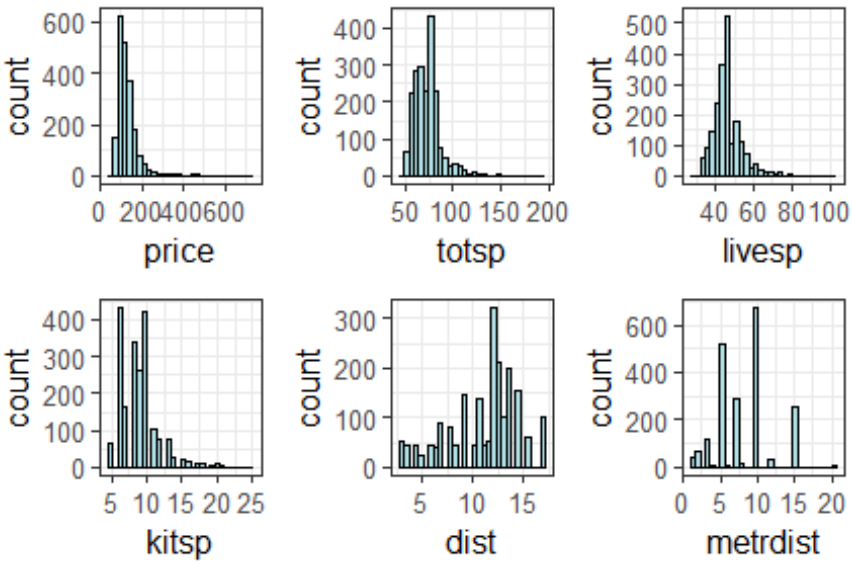

Рис. 1. Плотности распределения переменных

По графикам плотности распределения переменных действительно можно наблюдать некоторое сходство с нормальных распределением, однако с наличием тяжелых хвостов, которые явно свидетельствуют о особо влиятельных наблюдениях.

Задание 4

По методу Тьюки было выявлено 98 значений-выбросов для переменной цены, что неудивительно, поскольку при изначальном анализе гистограммы цены на квартиры в Москве был замечен тяжелый правый хвост, свидетельствующий о необычно дорогих квартирах. Следовательно, в новую переменную tidy_flats были отправлены очищенные от выбросов данные.

Задание 5

```
##
## The Median Test for tidy_flats$price ~ tidy_flats$code
```

```
##
## Chi Square = 112   DF = 7   P.Value 3.31e-21
## Median = 114
##
##   Median   r Min Max   Q25 Q75
## 1    125 258  72 210 100.0 144
## 2    100 212  70 210  93.0 117
## 3    125 303  70 210  98.0 150
## 4    130 191  72 212 105.0 156
## 5    113 331  71 210  99.0 125
## 6    102 255  56 208  83.5 125
## 7    108 222  50 200  93.2 126
## 8    127 170  77 200 103.5 150
##
## Post Hoc Analysis
##
## Groups according to probability of treatment differences and alpha level.
##
## Treatments with the same letter are not significantly different.
##
##   tidy_flats$price groups
## 4              130      a
## 8              127      a
## 1              125      a
## 3              125      a
## 5              113      b
## 7              108     bc
## 6              102     cd
## 2              100      d
```

С помощью функции median.test, которая проверяет гипотезу H0: нет значимой разницы в медианных значениях квартир в разных географических областях, было выяснено, что

существует значимая разница хотя бы для одного медианного значения, поскольку p.value оказалось меньше 0.05.

Также, построив график “ящик с усами” для цены квартир в зависимости от места расположения, можно рассмотреть среднестатистическое положений цен в каждой из рассматриваемых областей.

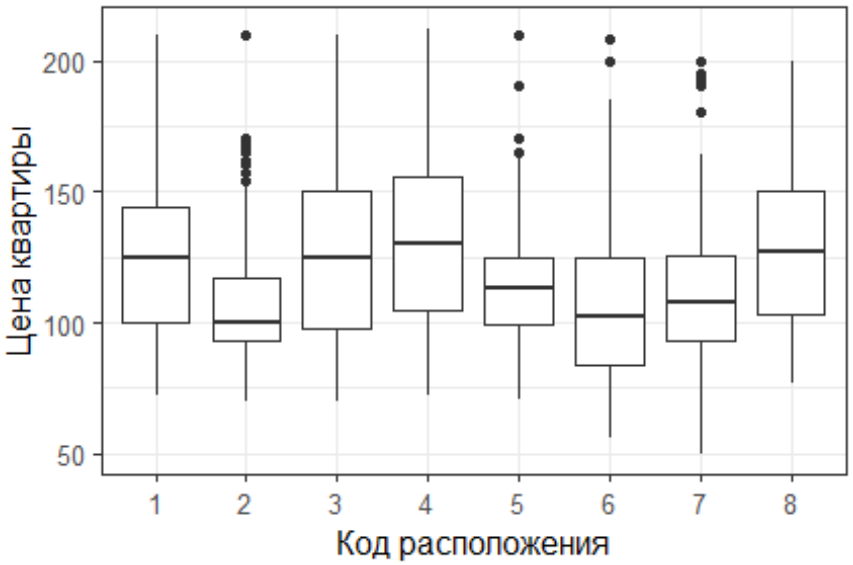

Рис. 2. График “ящик с усами” для цены квартиры в зависимости от места расположения

Исходя из графика также заметно значительное отличие медианных значений цены в Москве по местоположению, в особенности выделяются Северо-запад, вокруг Замоскворецкой линии метрополитена, и Северо-запад, вокруг Таганско-Краснопресненской линии метрополитена

Задание 6

Построим диаграмму рассеивания для предварительного анализа переменных.

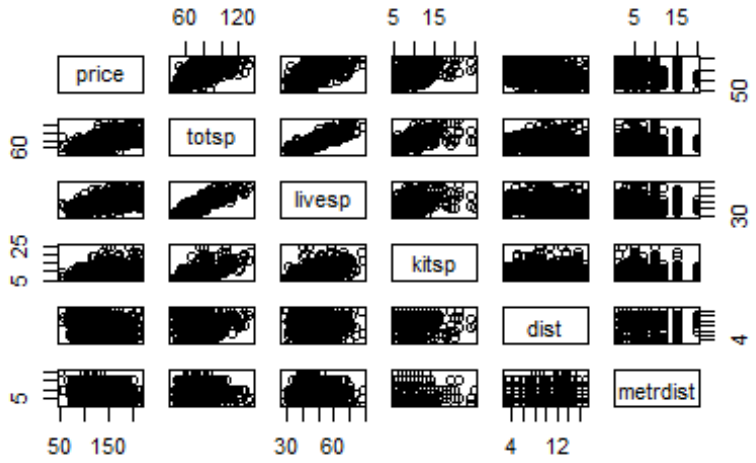

Рис. 3. Диаграмма рассеивания

Исходя из графика виден линейный паттерн зависимости цены квартиры от общей площади квартиры, жилой площади и площади кухни. При этом также заметна связь между общей площадью и жилой площадью, общей площадью и площадью кухни, а также связь между общей площадью и площадью кухни, что предварительно говорит о наличии мультиколлинеарности между регрессорами при включении их в модель. При этом поскольку для переменных расстояние до центра и расстояния до метро зависимости от остальных переменных практически не наблюдается.

Оценим коэффициенты корреляции

```
##          price   totsp  livesp   kitsp    dist metrdist
## price    1.000  0.7560  0.7296  0.5972 -0.3316  -0.1521
## totsp    0.756  1.0000  0.8622  0.7815 -0.1147  -0.0414
## livesp   0.730  0.8622  1.0000  0.5735 -0.1972  -0.0521
## kitsp    0.597  0.7815  0.5735  1.0000 -0.0619  -0.0285
## dist    -0.332 -0.1147 -0.1972 -0.0619  1.0000   0.0992
## metrdist -0.152 -0.0414 -0.0521 -0.0285  0.0992   1.0000
```

Исходя из коэффициентов корреляции заметны аналогичные наблюдения о зависимости между переменными, упомянутыми ранее при рассмотрении диаграммы рассеивания. При этом поскольку влияние переменных расстояния до центра и расстояния до метро

зависимости от остальных переменных не наблюдается, не стоит включать их в возможную модель.

С помощью графика “ящик с усами” оценим влияние дамми переменных на цену квартиры.

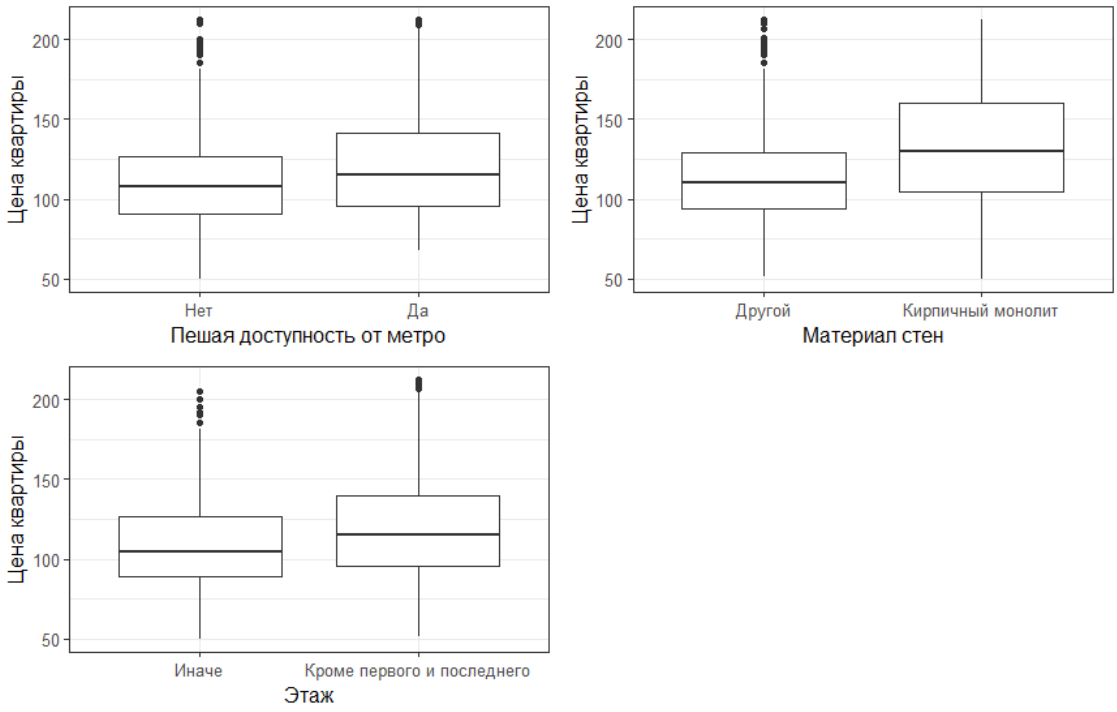

Рис. 4. Графики “ящик с усами” для дамми переменных

Исходя из графиков можно заметить возможную зависимость между ценой и материалом цен, а также ценой и этажом расположения квартиры, что ожидаемо. Однако также заметно некоторое влияние пешей доступности от метро на цену квартиры, тем не менее, оно кажется не столь значимым. Следовательно, в предварительную модель также не будем включать данную переменную. Однако, возможно стоит включить ее при поиски оптимальной модели.

Задание 7

В первоначальную модель включим в качестве регрессоров площадь квартиры, жилой площади, кухни, а также в качестве дамми переменных - материал стен и этаж. А также вычислим среднюю ошибку аппроксимации.

```
##
## Call:
```

```
## lm(formula = price ~ . - dist - metrdist - walk - code, data = tidy_flats)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -82.45 -12.51  -2.92  10.61  84.37
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept) -12.2062     3.3858   -3.61              0.00032 ***
## totsp         1.2277     0.0951   12.91 < 0.0000000000000002 ***
## livesp        0.3984     0.1410    2.83              0.00477 **
## kitsp         1.8053     0.3081    5.86         0.0000000055 ***
## brick1       15.4063     1.0665   14.45 < 0.0000000000000002 ***
## floor1        6.4987     1.1357    5.72         0.0000000122 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.4 on 1936 degrees of freedom
## Multiple R-squared:  0.557,  Adjusted R-squared:  0.556
## F-statistic:  488 on 5 and 1936 DF,  p-value: <0.0000000000000002

## [1] 0.127
```

Модель (model1) оказалось значимой, при этом коэффициенты в модели также оказались значимыми на уровне 5%, за исключением переменной жилой площади. Доля объясненных скорректированных значений оказалась равна 55.6%, а средняя ошибка аппроксимации - 12,7%. В принципе предварительная модель оказалась довольно хорошо предсказывающей. Однако, как было отмечено ранее, возможно стоить включить дамми переменную пешей доступности от метро.

```
##
## Call:
## lm(formula = price ~ . - dist - metrdist - code, data = tidy_flats)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
## -81.77 -12.47  -3.27  10.08  81.95
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept) -18.8484     3.3644   -5.60        0.0000000242 ***
## totsp         1.2759     0.0928   13.75 < 0.0000000000000002 ***
## livesp        0.3324     0.1376    2.42                0.016 *
## kitsp         1.7842     0.3003    5.94        0.0000000034 ***
## walk1         9.8849     0.9745   10.14 < 0.0000000000000002 ***
## brick1       14.1229     1.0472   13.49 < 0.0000000000000002 ***
## floor1        6.6404     1.1070    6.00        0.0000000024 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.9 on 1935 degrees of freedom
## Multiple R-squared:  0.58,   Adjusted R-squared:  0.579
## F-statistic:  445 on 6 and 1935 DF,  p-value: <0.0000000000000002

## [1] 0.124
```

Новая модель (model2) также осталась значимой, как и ее коэффициенты на уровне 5%, однако коэффициент перед переменной жилой площади оказался значимым лишь на уровне 1%. Тем не менее, увеличилась доля скорректированных объясненных значений до 57.9%, а средняя ошибка аппроксимации упала до уровня 12,4%. В целом можно опираться от второй модели в нахождении наиболее валидной, поскольку она оказалась более лучше предсказывающей значения.

Далее оценим наличие мультиколлинеарности между регрессорами во 2 модели с помощью коэффициента раздутости дисперсии.

```
##  totsp livesp  kitsp   walk  brick  floor
##   6.47   3.89   2.72   1.02   1.15   1.02
```

Заметна значительная раздутость дисперсии переменной общей площади квартиры, что говорит в пользу упрощения модели засчет удаления данного регрессора.

Оценим коэффициент раздутости дисперсии третьей модели (model3), удалив из второй модели переменную общей площади квартиры.

```
## livesp  kitsp   walk  brick  floor
##   1.46   1.40   1.02   1.14   1.02
```

Показатели для текущих переменных оказались меньше 2, что говорит о допустимости новой модели вследствие наличия лишь слабой мультиколлинеарности. Возможно, следовало ввести дополнительный штраф для модели, построив РИДЖ или ЛАССО регрессию, однако для этого необходим детальный анализ коэффициента лямбда на основе кросс-валидации.

Оценим новую третью модель.

```
## livesp  kitsp   walk  brick  floor
##   1.46   1.40   1.02   1.14   1.02

##
## Call:
## lm(formula = price ~ . - dist - metrdist - code - totsp, data = tidy_flats)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -86.87 -13.53  -3.38  11.42  90.69
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept) -19.9205     3.5230   -5.65        0.0000000180 ***
## livesp        1.8255     0.0885   20.63 < 0.0000000000000002 ***
## kitsp         4.6650     0.2254   20.70 < 0.0000000000000002 ***
## walk1         9.1999     1.0194    9.03 < 0.0000000000000002 ***
## brick1       13.0946     1.0940   11.97 < 0.0000000000000002 ***
## floor1        6.7931     1.1595    5.86        0.0000000055 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.8 on 1936 degrees of freedom
## Multiple R-squared:  0.539,  Adjusted R-squared:  0.538
## F-statistic:  452 on 5 and 1936 DF,  p-value: <0.0000000000000002

## [1] 0.133
```

Модель оказалась все еще значимой, при этом все коэффициенты также значимы на уровне 5%. Доля скорректированных объясненных значений снизилась незначительно до 53.8%, а средняя ошибка аппроксимации осталась на допустимом уровне в 13%.

Теперь проанализируем остатки для оценки применимости модели.

Сперва с помощью графика расстояния Кука проверим модель на наличие влиятельных наблюдений.

Рис.5. График расстояния Кука

Можно отметить, что влиятельных наблюдений не выявлено, поскольку не оказалось расстояний Кука, превышающих стандартный уровень - 1. Однако относительно уровня, учитывающего число наблюдений в модели, который оказался равен 0.002, присутствуют влиятельные наблюдения. Тем не менее, их общее влияние на модель можно считать незначимым.

Далее проверим разброс дисперсии остатков.

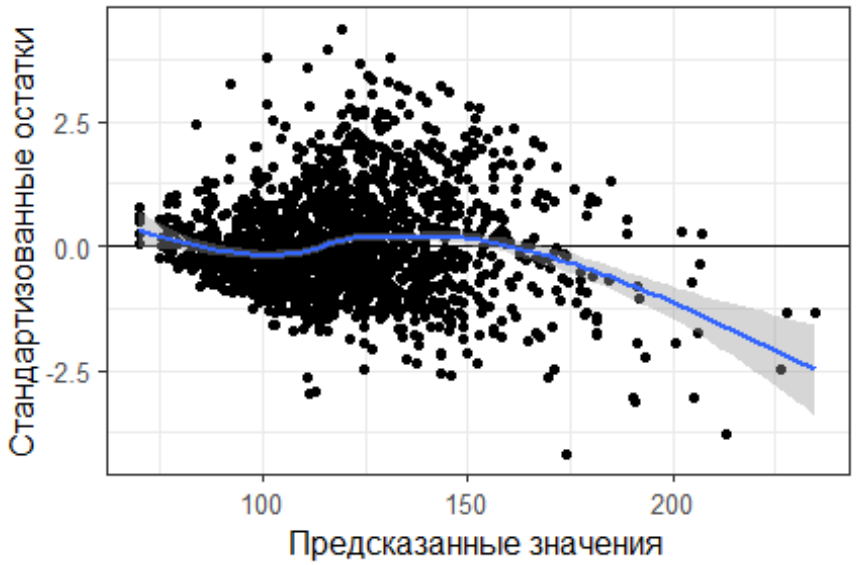

Рис. 6. График остатков от предсказанных значений

Исходя из графика разброса остатков от предсказанных значений можно наблюдать наличие гетероскедастичности, поскольку при увеличении предсказанных значений от 75 до 150 заметен рост разброса дисперсии остатков - стандартный паттерн воронки.

Однако также проведем тесты Голдфелда-Квандта и Бройша-Пагана для оценки наличия гетероскедастичности

```
##
##  Goldfeld-Quandt test
##
## data:  model3
## GQ = 1, df1 = 965, df2 = 965, p-value = 0.8
## alternative hypothesis: variance changes from segment 1 to 2

##
##  Breusch-Pagan test
##
## data:  model3
## BP = 279, df = 5, p-value <0.0000000000000002
```

Исходя из результатов, выявлено, что тест Голдфелда-Квандта выдал значение p.value > 0.05, то есть нет оснований отвергать H0 о наличии гомоскедастичности. Однако тест Бройша-Пагана выдал значение p.value < 0.05, вследствие чего H0 о наличии гомоскедастичности должна быть отвергнута. Основываясь на выводах из графика и тестах, можно сделать замечание, что остатки обладают гетероскедастичностью. Вследствие чего следует оценить коэффициенты с помощью робастных ошибок, устойчивых к гетероскедастичности.

```
coeftest(model3, vcov.=vcovHC(model3))

##
## t test of coefficients:
##
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  -19.920      3.898   -5.11         0.0000003524 ***
## livesp         1.826      0.104   17.59 < 0.0000000000000002 ***
## kitsp          4.665      0.276   16.92 < 0.0000000000000002 ***
## walk1          9.200      0.956    9.63 < 0.0000000000000002 ***
## brick1        13.095      1.269   10.32 < 0.0000000000000002 ***
## floor1         6.793      1.126    6.03         0.0000000019 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Тест коэффициентов с учетом ошибок, устойчивых к гетероскедастичности, показал, что коэффициенты остались по-прежнему значимыми на уровне 5%.

Далее также можно построить доверительные интервалы для коэффициентов модели, чьи ошибки устойчивы к гетероскедастичности и сравнить с исходными для модели.

```
##             estimate se_HC left_ci right_ci
## (Intercept)   -19.92 3.898  -27.56   -12.28
## livesp          1.83 0.104    1.62     2.03
## kitsp           4.67 0.276    4.12     5.21
## walk1           9.20 0.956    7.33    11.07
## brick1         13.09 1.269   10.61    15.58
## floor1          6.79 1.126    4.59     9.00
```

```
##                2.5 % 97.5 %
## (Intercept) -26.83 -13.01
## livesp         1.65   2.00
## kitsp          4.22   5.11
## walk1          7.20  11.20
## brick1        10.95  15.24
## floor1         4.52   9.07
```

Можно заметить, что доверительные интервалы для коэффициентов модели, чьи ошибки устойчивы к гетероскедастичности, расширились.

Далее проверим остатки на нормальность.

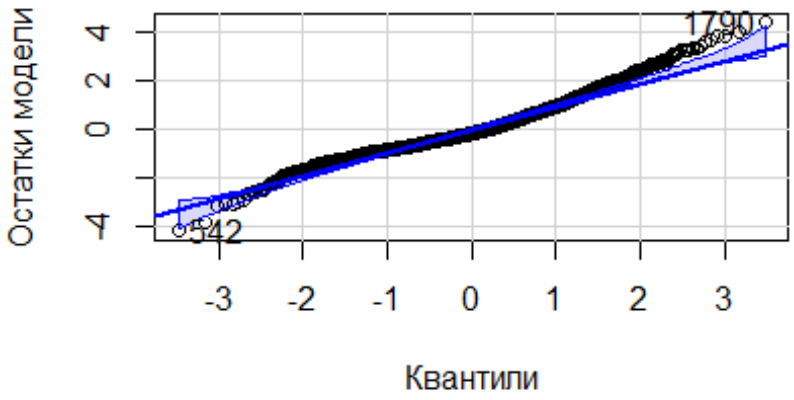

Рис. 7. График Q-Q plot

Исходя из графика, возможно отметить, что остатки не распределены нормально, поскольку лежат вне области квантилей нормального распределения.

Проведем тест Шапиро-Уилка для проверки гипотезы H0 о нормальном распределение остатков.

```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  model3$residuals
## W = 1, p-value <0.0000000000000002
```

По результатам теста заметно, что p.value оказалось меньше 5%, что действительно говорит о ненормальном распределении остатков. Нарушение данной предпосылки довольно существенно и предположительно вызвано либо нелинейностью зависимости переменной отклика от регрессора, либо наличием гетероскедастичности остатков.

Далее следует проверить предпосылку о равенстве математического ожидания остатков 0, однако в силу ненормальности их распределения это будет не корректно. Тем не менее, с помощью t.test проверим гипотезу H0 о равенстве математического ожидания остатков 0.

```
##
##  One Sample t-test
##
## data:  model3$residuals
## t = 0.000000000000002, df = 1941, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.925  0.925
## sample estimates:
##          mean of x
## 0.00000000000000109
```

Исходя из результатов теста, можно увидеть, что p.value оказалось более 5%, вследствие чего нет оснований отвергать гипотезу H0.

Последним шагом проверим предпосылку о наличии автокорреляции между остатками с помощью теста Дарбина-Уотсона.

```
##
##  Durbin-Watson test
##
## data:  model3
## DW = 2, p-value = 0.7
## alternative hypothesis: true autocorrelation is greater than 0
```

По результатам теста p.value оказалось более 5%, следовательно, нет оснований отвергнуть H0 об отсутствии автокорреляции между остатками.

ВЫВОД

В итоге мы получили линейную модель следующего вида, построенную на основе МНК:

$$price_i = -19.92 + 1.83livesp_i + 4.67kitsp_i + 9.2walk1_i + 13.09brick1_i + 6.79floor1_i$$

Данная модель, однако, не является совершенно оптимальной, поскольку для нее нарушены предпосылки об адекватности остатков, при этом оценки валидности также не являются наилучшими. Тем не менее, данная модель является одновременно наиболее лаконичной, качественной с точки зрения эффективности оценок коэффициентов и легко интерпретируемой.

Можно выделить наиболее влиятельные коэффициенты для визуализации упрощенной линейной модели. Для это стандартизуем количественные переменные в модели.

```
##
## Call:
## lm(formula = price ~ scale(livesp) + scale(kitsp) + walk + brick +
##     floor, data = tidy_flats)
##
## Coefficients:
##   (Intercept)  scale(livesp)   scale(kitsp)         walk1
##        103.47         11.79         11.56          9.20
##        brick1        floor1
##         13.09          6.79
```

Заметим, что наиболее значимыми оказались коэффициенты перед переменной жилой площади и дамми переменной материала стен. Построим линейную упрощенную модель по этим регрессорам.

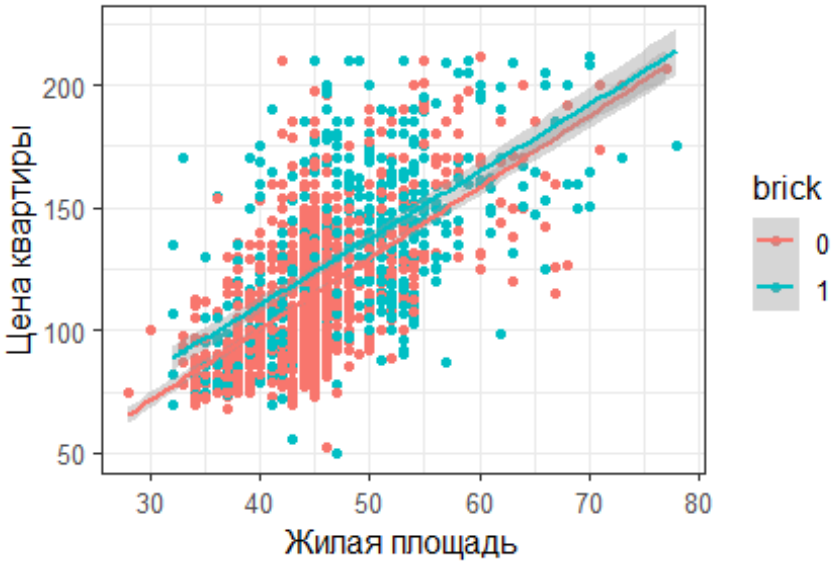

Рис. 8. Упрощенная линейная модель

Из упрощенной модели видна положительная зависимость между ценой квартиры и жилой площадью, при этом в случае строительства стен из кирпичного монолита цена также увеличивается.

СССЫЛКИ НА ИСПОЛЬЗОВАННЫЕ ИСТОЧНИКИ

[1] Индикаторы рынка недвижимости // Обзор рынка недвижимости по итогам 2016 года [Электронный ресурс]. URL: https://www.irn.ru/news/112490.html(дата обращения:19.12.2021), режим доступа - свободный.

[2] Журнал стратегия // Клуб экспертов: прогнозы по рынку недвижимости на 2021-2022 годы [Электронный ресурс]. URL: https://strategyjournal.ru/ekonomika-i-biznes/klub-ekspertov-prognozy-po-rynku-nedvizhimosti-na-2021-2022-gody/(дата обращения:19.12.2021), режим доступа - свободный.

[3] Портал открытых данных правительства Москвы [Электронный ресурс]. URL: https://data.mos.ru/(дата обращения: 10.06.2021), режим доступа - свободный.