**Отчёт по итоговому домашнему заданию  
по предмету “**[**Прикладные методы математической статистики**](https://lms.hse.ru/student.php?lessons_ID=143541)**”**

Выполнен студентами группы БПИ204:  
Пеганов Никита  
Куляхтина Алёна  
Завалинский Даниил

**Описание данных и особенности работы с ними**

Выбранные нами данные взяты из статьи <https://medium.com/@max.bobkov/machine-learning-moscow-flats-appraising-25a1e9f171db>. По словам автора, они получены в результате парсинга сайта [www.cian.ru](http://www.cian.ru), специализирующегося на продаже квартир. Также данные очищены «от дублей, объявлений с одинаковым текстом в описании, от строк с нулевой площадью кухни, без материала стен и этажей в доме».

Обработка данных происходила в среде jupyter notebook на языке python. Код решения можно найти в файле processing.ipynb.

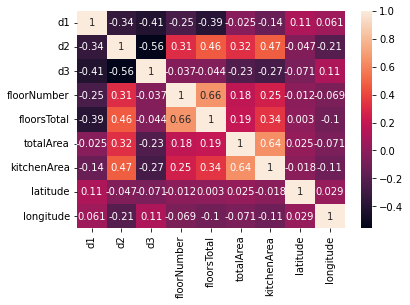
**Описание переменных, используемых в анализе**

Рассмотрим переменные, использованные в датасете:  
**wallsMaterial** — **материал стен  
floorNumber** — **на каком этаже расположена квартира**  
**floorsTotal** — общее число этажей в доме  
**totalArea** — общая площадь квартиры  
**kitchenArea** — площадь кухни  
**latitude** — широта расположения дома на карте  
**longitude** — долгота расположения дома на карте  
**price** — цена квартиры

Для проведения исследования мы разбили переменную wallsMaterial на несколько dummy-переменных:  
**d1** — является ли дом кирпичным  
**d2** — является ли дом монолитным  
**d3** — является ли дом панельным   
**d4** — все остальные материалы

Во избежание мультиколлинеарности переменная d4 не учитывалась, так как является композицией нескольких переменных.

Проверим объясняющие переменные на мультиколлинеарность. Для этого составим матрицу корреляции числовых переменных:



Матрица корреляции показывает, что значительных попарных корреляций между переменными в данных нет. Они не превосходят 0,7. Это означает, что возможных проблем с качеством получаемых оценок не предвидится.

**Описание полученных регрессионных моделей**

С помощью модуля анализа данных Excel посчитаны коэффициенты линейной регрессии. Ее вид:

Где – является ли дом кирпичным, – является ли дом монолитным, – является ли дом панельным, – номер этажа квартиры, – общее число этажей в квартире, – площадь квартиры, – площадь кухни, – широта расположения квартиры, – долгота расположения квартиры.

Это значит, что, если дом кирпичный, цена квартиры в нем падает на То есть кирпичные дома дешевле остальных.

Если дом монолитный, цена возрастает на То есть монолитные дома дороже остальных.

Если дом панельный, цена падает на То есть панельные дома значительно дешевле остальных.

При изменении номера этажа на единицу цена меняется на в ту же сторону, следовательно, чем выше этаж, тем квартира дороже.

При изменении количества этажей на единицу цена изменяется на в обратную изменению кол-ва этажей сторону, квартиры в малоэтажных домах дороже.

При изменении площади квартиры на единицу цена меняется на в ту же сторону.

При изменении площади кухни на 1 квадратный метр цена меняется на в обратную сторону, квадратный метр площади кухни дешевле, чем квадратный метр жилой площади.

При изменении широты координат квартиры на карте на 0,001 цена изменится на в ту же сторону, недвижимость на севере Москвы дороже.

При изменении долготы координат квартиры на карте на 0,001 цена изменится на в ту же сторону, недвижимость на востоке Москвы дороже.

Таблицы с значениями коэффициентов, доверительными интервалами и значимости модели:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Регрессионная статистика* | |  |  |  |  |  |
| Множественный R | 0,760156 |  |  |  |  |  |
| R-квадрат | 0,577838 |  |  |  |  |  |
| Нормированный R-квадрат | 0,577778 |  |  |  |  |  |
| Стандартная ошибка | 19796299 |  |  |  |  |  |
| Наблюдения | 63931 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| Дисперсионный анализ | | |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Значимость F* |  |
| Регрессия | 9 | 3,43E+19 | 3,81E+18 | 9721,375 | 0 |  |
| Остаток | 63921 | 2,51E+19 | 3,92E+14 |  |  |  |
| Итого | 63930 | 5,93E+19 |  |  |  |  |
|  |  |  |  |  |  |  |
|  | *Коэффициенты* | *Стандартная ошибка* | *t-статистика* | *P-Значение* | *Нижние 95%* | *Верхние 95%* |
| Y-пересечение | -7,3E+08 | 48303788 | -15,0158 | 7,07E-51 | -8,2E+08 | -6,3E+08 |
| Переменная X 1 | -478078 | 327620,1 | -1,45925 | 0,144502 | -1120214 | 164057,6 |
| Переменная X 2 | 2650402 | 338107,2 | 7,838941 | 4,61E-15 | 1987712 | 3313093 |
| Переменная X 3 | -1681820 | 304037,2 | -5,53163 | 3,19E-08 | -2277734 | -1085907 |
| Переменная X 4 | 15427,11 | 16333,41 | 0,944512 | 0,344912 | -16586,4 | 47440,62 |
| Переменная X 5 | -274289 | 13773,81 | -19,9138 | 5,72E-88 | -301286 | -247293 |
| Переменная X 6 | 536083,3 | 2350,323 | 228,0892 | 0 | 531476,7 | 540690 |
| Переменная X 7 | -157675 | 19027,32 | -8,28676 | 1,19E-16 | -194968 | -120381 |
| Переменная X 8 | 9847215 | 791885,8 | 12,43515 | 1,85E-35 | 8295118 | 11399312 |
| Переменная X 9 | 4344588 | 545530,5 | 7,96397 | 1,69E-15 | 3275348 | 5413829 |

Затем в программе Excel взяты натуральные логарифмы от значений всех объясняющих переменных, а также от цены квартир. На них также построена линейная регрессия с помощью модуля анализа данных. Полученная логарифмическая регрессия имеет вид:

При увеличении номера этажа на 1% цена увеличивается на 0,019357%. При увеличении количества этажей на 1% цена увеличивается на 0,005159%. При увеличении площади квартиры на 1% цена увеличивается на 1,102143%. При увеличении площади кухни на 1% цена увеличивается на 0,122064%. При увеличении широты квартиры на 1% цена увеличивается на 46,40707%. При увеличении долготы на 1% цена увеличивается на 5,806979%.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Регрессионная статистика* | |  |  |  |  |  |
| Множественный R | 0,876736 |  |  |  |  |  |
| R-квадрат | 0,768666 |  |  |  |  |  |
| Нормированный R-квадрат | 0,768634 |  |  |  |  |  |
| Стандартная ошибка | 0,323763 |  |  |  |  |  |
| Наблюдения | 63931 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| Дисперсионный анализ | | |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Значимость F* |  |
| Регрессия | 9 | 22263,63 | 2473,736 | 23599,36 | 0 |  |
| Остаток | 63921 | 6700,339 | 0,104822 |  |  |  |
| Итого | 63930 | 28963,97 |  |  |  |  |
|  |  |  |  |  |  |  |
|  | *Коэффициенты* | *Стандартная ошибка* | *t-статистика* | *P-Значение* | *Нижние 95%* | *Верхние 95%* |
| Y-пересечение | -163,088 | 3,129813 | -52,1081 | 0 | -169,223 | -156,954 |
| Переменная X 1 | 0,130361 | 0,005418 | 24,06119 | 2,4E-127 | 0,119741 | 0,14098 |
| Переменная X 2 | 0,105977 | 0,005551 | 19,09324 | 4,84E-81 | 0,095098 | 0,116856 |
| Переменная X 3 | -0,14728 | 0,005011 | -29,3892 | 1,4E-188 | -0,1571 | -0,13746 |
| Переменная X 4 | 0,019452 | 0,001823 | 10,67094 | 1,47E-26 | 0,015879 | 0,023025 |
| Переменная X 5 | 0,005184 | 0,003324 | 1,559442 | 0,118897 | -0,00133 | 0,0117 |
| Переменная X 6 | 1,101585 | 0,003816 | 288,713 | 0 | 1,094106 | 1,109063 |
| Переменная X 7 | 0,122598 | 0,004745 | 25,83581 | 2E-146 | 0,113298 | 0,131899 |
| Переменная X 8 | 38,31237 | 0,722594 | 53,02061 | 0 | 36,89608 | 39,72865 |
| Переменная X 9 | 5,672806 | 0,335414 | 16,91286 | 4,99E-64 | 5,015395 | 6,330218 |

Таблицы с значениями коэффициентов, доверительными интервалами и значимости модели:

**Описание проведенных тестов - Тест Чоу**

Для проведения теста Чоу выборка была разделена на две подвыборки – дома, в которых не более 12ти этажей, и те, в которых 13 и больше – подвыборки A и B соответственно.

n = 63931 – общее число наблюдений в полной выборке

nA = 29550 – число наблюдений в выборке А

nB = 34381 – число наблюдений в выборке B

k = 7 – число регрессоров без константы модели

SL = 5% - уровень значимости

1. Оцениваем регрессию для полной выборки и находим сумму квадратов остатков

RSSall = 25 206 816 903 486 800 000

1. Оцениваем регрессии по подвыборкам A и B и находим сумму квадратов остатков для каждой

RSSA = 17 432 539 921 279 500 000

RSSB = 5 507 061 025 055 580 000

1. Рассчитываем значение статистики:
2. Критическая точка:
3. Получим, что Тнабл ∈ [0; Ткр].

**Вывод: гипотеза не отвергается, и мы оцениваем единую модель по всей выборке.**