**Национальный исследовательский университет**

**“Высшая школа экономики”**

**Московский институт электроники и математики им. А.Н. Тихонова, Москва**

Направление 09.04.01. Информатика и вычислительная техника

Магистерская программа "Компьютерные системы и сети"

Отчет о самостоятельной работе по МАГОЛЕГО

“Основы анализа данных”

Бригада №9:

Маттеи Юлия Павловна, 1 курс, МКЛ221

Москва, 2022

Оглавление

[1. Общая постановка задачи 3](#_Toc117410676)

[1.1. Описание прикладной области и данных 3](#_Toc117410677)

[1.2. Основные гипотезы, которые планируется проверить в рамках исследования 4](#_Toc117410678)

[2 Предварительный анализ собранных данных 4](#_Toc117410679)

[2.1 Анализ особенностей данных: потенциальные ошибки и пропущенные значения, группы и выбросы 4](#_Toc117410680)

[2.1.1 Анализ количественных переменных 4](#_Toc117410681)

[2.2.2. Анализ качественных переменных. 8](#_Toc117410682)

[2.2 Анализ статистической связи. 9](#_Toc117410683)

[2.2.1 Графический анализ пары «целевая переменная – качественная объясняющая переменная». 9](#_Toc117410684)

[2.2.2 Графический анализ пары «числовая зависимая переменная – числовая независимая переменная». 10](#_Toc117410685)

[2.2.3 Анализ статистической взаимосвязи между независимыми переменными. 12](#_Toc117410686)

[2.2.4 Предварительная проверка гипотез 12](#_Toc117410687)

[3 Проверка гипотез с помощью моделирования 13](#_Toc117410688)

[3.1. Построение базовой модели. 13](#_Toc117410689)

[3.2. Проверка гипотез с помощью моделирования 14](#_Toc117410690)

[3.3. Оптимизация итоговой модели, сравнение качества моделей. 16](#_Toc117410691)

[4. Заключение 17](#_Toc117410692)

[Приложение 18](#_Toc117410693)

# 1. Общая постановка задачи

Задачей данного исследования является анализ статистической взаимосвязи между общим числом арендованных велосипедов за день, этот показатель мы будем называть целевой переменной, и множеством других показателей, которые мы будем называть объясняющими переменными: статус дня недели (выходной (или праздник) или рабочий день), погодные условия, температура, влажность, скорость ветра.

## 1.1. Описание прикладной области и данных

Выбрана следующая тема исследования: **аренда велосипедов американской системы Capital Bikeshare в зависимости от экзогенных и погодных условий в 2011 и 2012 годах**. В таблице 1 представлены учтенные в анализе характеристики, от которых зависит выгода компании по аренды велосипедов. Собраны данные за 2 полных года, поэтому время года (сезон), данные даты учитывать не будем; количество наблюдений - 731.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № | Характеристика  объекта/явления | Название  переменной | Шкала  измерения | Роль:  целевая/объясняющая |
| 1 | Рабочий или выходной день (1: рабочий, 0: выходной или праздник) | workingday | Номинальная | объясняющая |
| 2 | Погодные условия (1: ясно, малооблачно; 2: туман и облачно, туман и малооблачно, туман; 3: слабый снег или слабый дождь, гроза, тучи и слабый дождь; 4: сильный дождь, град, грозовая буря, туман и снег, плотный туман) | weathersit | Порядковая | объясняющая |
| 3 | Температура, в долях от максимальной зафиксированной (t-t\_min)/(t\_max-t\_min), t\_min=-8, t\_max=+39) | temp | Относительная | объясняющая |
| 4 | Влажность (доля) | hum | Относительная | объясняющая |
| 5 | Скорость ветра (км/ч) | windspeed | Относительная | объясняющая |
| 6 | Количество арендованных велосипедов | cnt | Относительная | целевая |

Таблица 1. Описание факторов, учтенных в анализе

Ссылка на данные: <https://code.datasciencedojo.com/datasciencedojo/datasets/blob/master/Bike%20Sharing/day.csv>

## 1.2. Основные гипотезы, которые планируется проверить в рамках исследования

Сформулируем три гипотезы:

Простая - Чем выше скорость ветра, тем меньшее количество велосипедов арендуется.

Сложная - В ясную, хорошую погоду сдается больше велосипедов, чем в плохую, причем в выходные и праздники значение еще выше.

Сложная - Чем ниже температура, тем меньше арендуют велосипеды, особенно при высокой влажности.

# 2 Предварительный анализ собранных данных

## 2.1 Анализ особенностей данных: потенциальные ошибки и пропущенные значения, группы и выбросы

### 2.1.1 Анализ количественных переменных

Проанализируем имеющиеся количественные переменные: температура, влажность, скорость ветра, количество арендованных велосипедов.

1. Температура

Необходимо обратить внимание, что в исходных данных температура представлена в нормализованном виде, таким образом, тип шкалы сменился с интервальной на относительную.

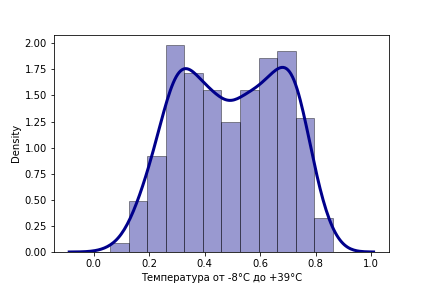


Рисунок 1. Гистограмма: частота температур по дням в нормализованном виде

Распределение бимодальное, наблюдаются два пика около 0.3 и 0.7, то есть около +6°C и +24°C.

|  |  |
| --- | --- |
| Основные статистики | Температура |
| Число наблюдений | 731.0 |
| Среднее | 0.495 |
| Стандартное отклонение | 0.183 |
| Минимум | 0.059 |
| 25% | 0.337 |
| 50% | 0.498 |
| 75% | 0.655 |
| Максимум | 0.862 |

Таблица 2. Основные статистики для показателя температуры

В остальном распределение близко к стандартному: в целом значения незначительно сдвинуты влево (показатель асимметрии равен -0.05), график уплощен относительно нормального (эксцесс равен -1.12 (используется метод библиотеки pandas, по которому для нормального распределения эксцесс равен нулю)), медиана (50%) и среднее крайне близки. Для данных, покрывающих все значения градусника за год, распределение достаточно ожидаемое: мало дней, в которые очень холодно, и мало дней, когда очень жарко, наблюдается два выраженных сезона холодной (около +6°C) и теплой (около +24°C) погоды, причем холодных дней в году немного больше.

1. Влажность

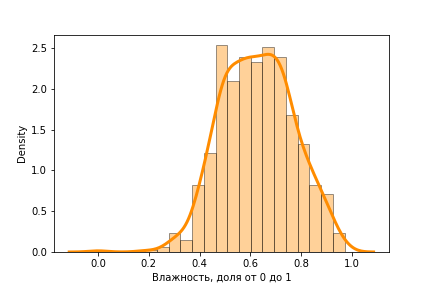


Рисунок 2. Гистограмма: частота значений влажности по дням, в долях

Распределение одномодальное. Мода сильно отстоит от среднего и медианы, что влияет на показатели симметрии (-0.07) и эксцесса (-0.065), значения которых отрицательны, несмотря на очевидное общее смещение значений вправо относительного среднего (0.5).

|  |  |
| --- | --- |
| Основные статистики | Влажность |
| Число наблюдений | 731.0 |
| Среднее | 0.627 |
| Стандартное отклонение | 0.142 |
| Минимум | 0.000 |
| 25% | 0.520 |
| 50% | 0.626 |
| 75% | 0.730 |
| Максимум | 0.972 |

Таблица 3. Основные статистики для показателя влажности

Можно сделать вывод, что для климата описываемого региона стандартна средняя или слегка повышенная влажность.

1. Скорость ветра

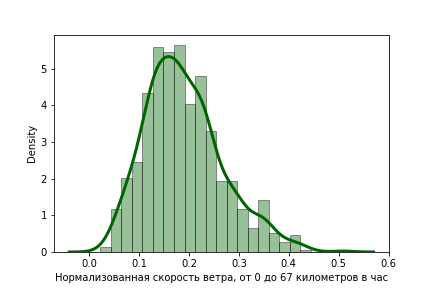


Рисунок 3. Гистограмма: частота скоростей ветра по дням в нормализованном виде

Распределение близко к нормальному. Значения моды, медианы и среднего очень близки. Коэффициент асимметрии равен 0.677, эксцесса — 0.41.

|  |  |
| --- | --- |
| Основные статистики | Скорость ветра |
| Число наблюдений | 731.0 |
| Среднее | 0.190 |
| Стандартное отклонение | 0.078 |
| Минимум | 0.022 |
| 25% | 0.135 |
| 50% | 0.181 |
| 75% | 0.233 |
| Максимум | 0.507 |

Таблица 4. Основные статистики для показателя скорости ветра

По шкале Бофорта, 20 км/ч — это умеренный ветер. Для изучаемого региона характерен умеренный ветер.

1. Количество арендованных велосипедов

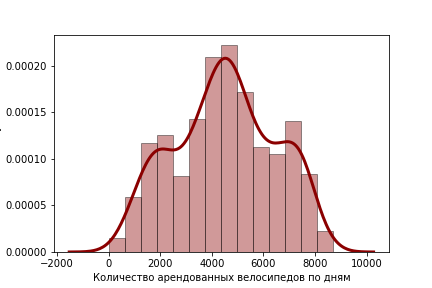


Рисунок 4. Гистограмма: число арендованных велосипедов по дням

|  |  |
| --- | --- |
| Основные статистики | Арендованные велосипеды |
| Число наблюдений | 731.0 |
| Среднее | 4504.35 |
| Стандартное отклонение | 1937.21 |
| Минимум | 22.0 |
| 25% | 3152.0 |
| 50% | 4548.0 |
| 75% | 5956.0 |
| Максимум | 8714.0 |

Таблица 5. Основные статистики для показателя аренды велосипедов

Асимметрии -0.047, эксцесса -0.81.

Для целевой переменной применим правило трех сигм. Среднеквадратичное отклонение равно 1937.21, подсчитаем долю значений параметра, лежащих на интервале от -1307,28 (т. е. 4504.35-1937.21\*3) до 10 315,98 (4504.35+1937.21\*3). На этом интервале лежат все значения переменной, таким образом, никакие значения не определяем как выбросные.

Для каждой из четырех рассмотренных количественных характеристик наблюдаются значения симметрии и эксцесса в пределах от -1 до 1, что говорит о незначительных отклонениях от нормального распределения и позволяет использовать данные для дальнейшего анализа. Пропущенных значений нет.

### 2.2.2. Анализ качественных переменных.

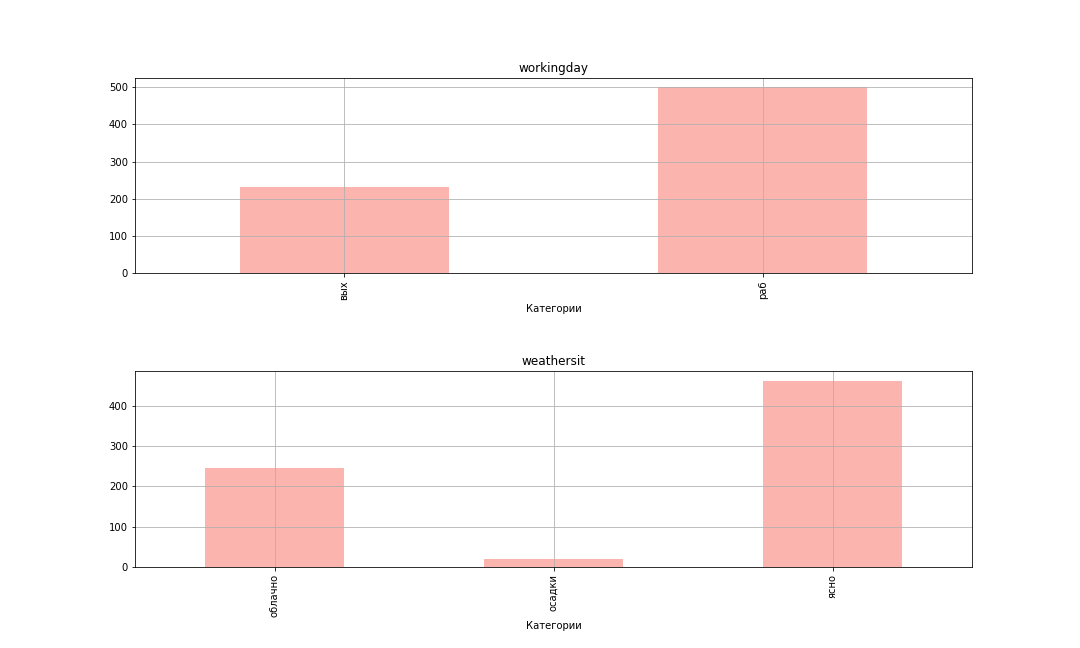
.

Рисунок 5. Столбчатые диаграммы: верхняя - категории рабочий/выходной день, нижняя - категории погодных условий облачно/осадки/ясно

|  |  |  |
| --- | --- | --- |
|  | Абсолютное количество | Процент |
| выходной | 500 | 68,4% |
| рабочий | 231 | 31,6% |
| сумма | 731 | 100% |

Таблица 6. Частоты значений по категориям для переменной выходной/рабочий день

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Исходные категории | | Укрупненные категории | |
|  | Абсолютное | Процент | Абсолютное | Процент |
| ясно | 463 | 63,3% | 463 | 63,3% |
| облачно | 247 | 33,8% | 268 | 36,7% |
| осадки | 21 | **2,8%** |
| сумма | 731 | 100% | 731 | 100% |

Таблица 7. Частоты значений по категориям для переменной погодных условий

Во-первых, оказалось, что в 2011-2012 годах не были зафиксированы дни с очень плохой, штормовой погодой (4 из 4). Во-вторых, категория “осадки” (3 из 4) получила долю в только 3%, присоединим ее к единственной соседней, “облачно” (2 из 4). Далее будем переменную погодных условий считать состоящей из двух категорий: “ясно” и “облачно”.

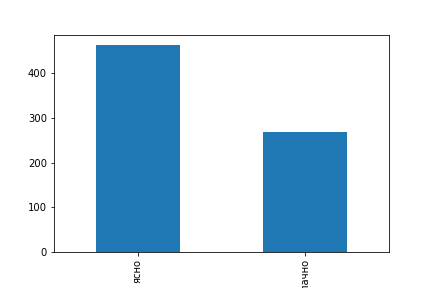


Рисунок 6. Столбчатая диаграмма для двух категорий погодных условий: ясно / облачно

## 2.2 Анализ статистической связи.

### 2.2.1 Графический анализ пары «целевая переменная – качественная объясняющая переменная».

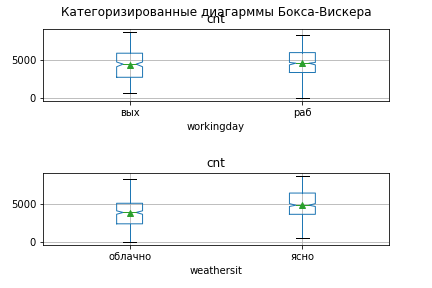


Рисунок 7. Зависимость аренды велосипедов от дня недели (выходной или рабочий) и погодных условий

По рисунку 7 видно:

1) в рабочие дни велосипеды арендуют немного больше, чем в выходные, статистически отличие средних между выборками не значительно: по непараметрическому дисперсионному критерию Крускала-Уоллиса p-value = 0.12. Нельзя принять гипотезу H1 о том, что в рабочие дни велосипед более популярный вид транспорта, чем в выходные.

2) в ясные дни арендуется больше велосипедов, чем в облачные и пасмурные дни. Принимаем гипотезу: по критерию Крускала-Уоллиса p-value <0.001 (= 7\*10-12).

### 2.2.2 Графический анализ пары «числовая зависимая переменная – числовая независимая переменная».

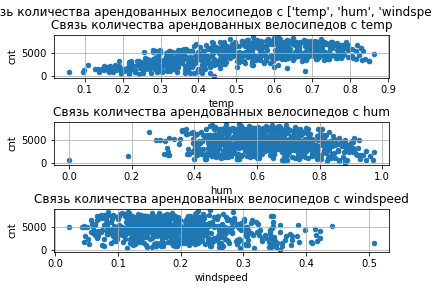


Рисунок 8. Связь между количественными переменными (температура, влажность, скорость ветра) с количеством арендованных велосипедов

По рисунку 8 можно предположить:

1. Велосипеды сдаются в теплую погоду (15-25 градусов) лучше, чем в холодную (0-15 градусов).
2. В экстремально влажные дни велосипедами меньше пользуются, чем в дни с обычной, умеренной влажностью.
3. При высокой скорости ветра арендуется меньше велосипедов, чем при умеренном, обычном ветре.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **temp** | **hum** | **windspeed** | **cnt** |
| **temp** | 1 | 0,126962939 | -0,15794412 | 0,627494009 |
| **hum** | 0,126962939 | 1 | -0,248489099 | -0,100658562 |
| **windspeed** | -0,15794412 | -0,248489099 | 1 | -0,234544997 |
| **cnt** | 0,627494009 | -0,100658562 | -0,234544997 | 1 |
|  |  |  |  |  |
|  | **temp** | **hum** | **windspeed** | **cnt** |
| **temp** | 0 | 0,000580089 | 1,78686E-05 | 2,81062E-81 |
| **hum** | 0,000580089 | 0 | 9,48796E-12 | 0,006454143 |
| **windspeed** | 1,78686E-05 | 9,48796E-12 | 0 | 1,35996E-10 |
| **cnt** | 2,81062E-81 | 0,006454143 | 1,35996E-10 | 0 |

Таблица 8. Коэффициент Пирсона и оценка его значимости для всех количественных переменных

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **temp** | **hum** | **windspeed** | **cnt** |
| **temp** | 1 | 0,129903165 | -0,147152904 | 0,622034473 |
| **hum** | 0,129903165 | 1 | -0,239006478 | -0,098125824 |
| **windspeed** | -0,147152904 | -0,239006478 | 1 | -0,21719701 |
| **cnt** | **0,622034473** | -0,098125824 | -0,21719701 | 1 |
|  |  |  |  |  |
|  | **temp** | **hum** | **windspeed** | **cnt** |
| **temp** | 0 | 0,00042974 | 6,5098E-05 | 1,67455E-79 |
| **hum** | 0,00042974 | 0 | 5,9073E-11 | 0,00793323 |
| **windspeed** | 6,5098E-05 | 5,9073E-11 | 0 | 2,96943E-09 |
| **cnt** | **1,67455E-79** | 0,00793323 | 2,96943E-09 | 0 |

Таблица 9. Коэффициент Спирмена и оценка его значимости для всех количественных переменных

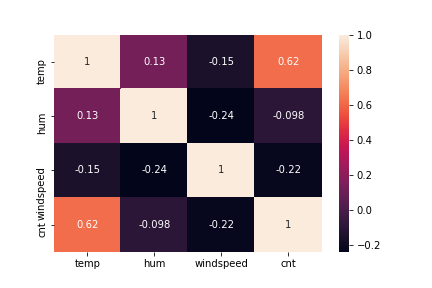
Мы посчитали коэффициенты Спирмена и Пирсона, чтобы сравнить, отличаются ли значения этих коэффициентов для какой-либо пары переменных, серьезных отличий не наблюдается. Коэффициент Спирмена актуален, в отличие от Пирсона, для отношения между температурой, для которой распределение не нормальное, и количеством арендованных велосипедов. 

Рисунок 9. Таблица сопряженности (Спирман) для количественных переменных

Наиболее наглядно показаны зависимости между переменными на рисунке 9. Среди них единственная связь, для которой есть хотя бы средняя сила, - между переменными температура и количество арендованных велосипедов, согласно таблице 9 эта связь статистически значима, подтверждается направление и сила связи по выдвинутой гипотезе 1. Для гипотез 2 и 3: почти нет связи между влажностью и арендой велосипедов, есть слабая обратная связь между силой ветра и арендой велосипедов (чем сильнее ветер, тем меньше арендуют велосипеды).

### 2.2.3 Анализ статистической взаимосвязи между независимыми переменными.

- для количественных переменных

Будем обращаться к данным таблиц 8 и 9. Возможна слабая отрицательная связь между переменными влажности и скорости ветра, при построении модели следует учесть возможность мультиколлинеарности. В остальном можно утверждать об отсутствии связи между объясняющими (независимыми) переменными, данные количественных переменных подходят для множественной линейной регрессионной модели.

- для качественных переменных

Одна переменная номинальная, вторая порядковая, используем критерий хи-квадрат.

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Погодные условия (weathersit) | |
| Облачно | Ясно |
| Рабочий ли день (workingday) | Выходной | 75 | 156 |
| Рабочий | 193 | 307 |

Таблица 10. Таблица сопряженности: статус дня недели и погодные условия

p-value = 0.12, V-Крамера = 0.056, ассоциация между переменными крайне слаба: не отвергаем нулевую гипотезу об отсутствии связи между двумя качественными объясняющими переменными.

### 2.2.4 Предварительная проверка гипотез

Выдвинутая до предварительного анализа данных гипотеза о связи силы ветра и количестве арендуемых велосипедов не подтвердилась. Между переменными наблюдается только связь слабой силы.

Гипотеза о связи хорошей (ясной) погоды и выходного дня с большим числом сданных велосипедов будет проверена в п. 3, пока что известно о статистически значимой связи между хорошей погодой и популярностью велосипедов и об отсутствии сильной корреляции между выходным или будним днем с арендой велосипедов.

Гипотеза о связи низкой температуры и высокой влажности (снега или холодного дождя) с низкой популярностью велосипедов также проверяется в п. 3, пока что известно о средней силе положительной связи между температурой и арендой велосипедов и об отсутствии связи между влажностью и арендой велосипедов.

# 3 Проверка гипотез с помощью моделирования

## 3.1. Построение базовой модели.

Построим множественную линейную регрессию с фиктивными переменными. Целевая переменная – количество арендованных велосипедов (cnt), фиктивные - workingday\_раб и weathersit\_облачно, базовые – температура (temp), влажность (hum) и скорость ветра (windspeed). Формула: cnt = a0 +a1\* workingday\_раб + a2\* weathersit\_облачно + a3\* temp + a4\*hum + a5\*windspeed + v.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | \*\*\*\*\*\* | Оценка | базовой | модели | \*\*\*\*\*\* |  |
|  | OLS | Regression | Results |  |  |  |
| ============================================================================== | | | | | | |
| Dep. | Variable: | cnt | R-squared: | 0.456 |  |  |
| Model: | OLS |  | Adj.R-squared: | 0.451 |  |  |
| Method: | Least | Squares | F-statistic: | 97.07 |  |  |
| Date: | Sat, 22 Oct 2022 |  | Prob (F-statistic): | 3.27E-74 |  |  |
| Time: | 23:41:19 |  | Log-Likelihood: | -5075.7 |  |  |
| No. | Observations: | 585 | AIC: | 1.02E+04 |  |  |
| Df | Residuals: | 579 | BIC: | 1.02E+04 |  |  |
| Df | Model: | 5 |  |  |  |  |
| Covariance | Type: | nonrobust | |  |  |  |
| ====================================================================================== | | | | | | |
|  | coef | Std err | t | P>|t| | [0.025 | 0.975] |  |
| -------------------------------------------------------------------------------------- | | | | | | |  |
| Const | 3868.843 | 395.356 | 9.786 | 0 | 3092.336 | 4645.349 |
| workingday\_раб | 203.6706 | 129.518 | 1.573 | 0.116 | -50.711 | 458.052 |
| weathersit\_облачно | 33.5615 | 145.28 | 0.231 | 0.817 | -251.778 | 318.901 |
| Temp | 6625.804 | 336.231 | 19.706 | 0 | 5965.422 | 7286.185 |
| Hum | -3006.26 | 493.122 | -6.096 | 0 | -3974.79 | -2037.74 |
| windspeed | -4798.5 | 771.234 | -6.222 | 0 | -6313.25 | -3283.74 |
| ============================================================================== | | | | | | |
| Omnibus: | 5.053 | Durbin-Watson: |  | 2.021 |  |  |
| Prob(Omnibus): | 0.08 | Jarque-Bera | (JB): | 4.624 |  |  |
| Skew: | 0.16 | Prob(JB): |  | 0.0991 |  |  |
| Kurtosis: | 2.704 | Cond. | No. | 21.9 |  |  |
| ============================================================================== | | | | | | |

Таблица 11. Оценка базовой модели

По таблице 11, незначимыми переменными являются weathersit\_облачно и workingday\_раб, однако удалить их для оптимизации невозможно, так как они нужны для проверки гипотезы (2).

По таблице 12 все количественные переменные имеют высокий уровень метрики VIF, наблюдаются высокие уровни мультиколлинеарности, однако все три переменные нужны для проверки гипотез, не можем оптимизировать модель, удаляя какую-либо из них.

|  |  |  |
| --- | --- | --- |
|  | **vars** | **VIF** |
| **0** | temp | 7.202347 |
| **1** | hum | 8.801869 |
| **2** | windspeed | 4.499157 |

Таблица 12. Оценка мультиколлинеарности количественных переменных базовой модели

По таблице 13, значения критерия Уайта стремятся к нулю, присутствует гетероскедастичность.

|  |  |
| --- | --- |
|  | **0** |
| **LM** | 53.57571 |
| **LM\_P** | 2.14E-05 |
| **F** | 3.170081 |
| **F\_P** | 1.22E-05 |

Таблица 13. Оценка гетероскедастичности базовой модели

Критерии оценки качества базовой модели: Adj R2 = 0.451, AIC (Акаике) = 1.02E+04.

## 3.2. Проверка гипотез с помощью моделирования

Гипотеза (1) об отрицательной связи скорости ветра и количества арендованных велосипедов подтверждается по таблице 11: переменная windspeed значима и ее коэффициент отрицателен. Моделирование подтверждает, что с понижением ветрености повышается число арендуемых велосипедов.

Для гипотезы (2) модифицируем модель: cnt = a0 +a1\*workingday\_раб + (a10 + a11\* workingday\_раб)\* weathersit\_облачно + a3\*temp + a4\*hum + a5\*windspeed + v = a0 + a1\*workingday\_раб + a10\*weathersit\_облачно + a11\*workingday\_раб\*weathersit\_облачно + a3\* temp + a4\*hum + a5\*windspeed + v. Если гипотеза справедлива, то a10 < 0, a11 > 0 и значим.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | \*\*\*\*\*\* | Оценка | базовой | модели | \*\*\*\*\*\* |  |
|  | OLS | Regression | Results |  |  |  |
| ============================================================================== | | | | | | |
| Dep. | Variable: | cnt | R-squared: | 0.462 |  |  |
| Model: | OLS |  | Adj. R-squared: | 0.457 |  |  |
| Method: | Least | Squares | F-statistic: | 82.87 |  |  |
| Date: | Sun, 23 Oct 2022 | | Prob (F-statistic): | 1.12E-74 |  |  |
| Time: | 7:51:36 |  | Log-Likelihood: | -5072.3 |  |  |
| No. | Observations: | 585 | AIC: | 1.02E+04 |  |  |
| Df | Residuals: | 578 | BIC: | 1.02E+04 |  |  |
| Df | Model: | 6 |  |  |  |  |
| Covariance | Type: | nonrobust | |  |  |  |
| ====================================================================================== | | | | | | |
|  | Coef | std err | t | P>|t| | [0.025 | 0.975] |
| -------------------------------------------------------------------------------------- | | | | | | |
| const | 3515.72 | 408.642 | 8.603 | 0 | 2713.115 | 4318.325 |
| workingday\_раб | 393.8609 | 157.003 | 2.509 | 0.012 | 85.495 | 702.227 |
| weathersit\_облачно | 105.1604 | 254.878 | 0.413 | 0.68 | -395.44 | 605.761 |
| Temp | 6449.25 | 337.565 | 19.105 | 0 | 5786.247 | 7112.254 |
| hum | -2404.54 | 531.13 | -4.527 | 0 | -3447.72 | -1361.36 |
| windspeed | -4556.35 | 774.984 | -5.879 | 0 | -6078.48 | -3034.23 |
| ww | -525 | 276.268 | -1.9 | 0.058 | -1067.61 | 17.611 |
| ============================================================================== | | | | | | |
| Omnibus: | 6.389 |  | Durbin-Watson: | 2.005 |  |  |
| Prob(Omnibus): | 0.041 | Jarque-Bera | (JB): | 5.572 |  |  |
| Skew: | 0.169 |  | Prob(JB): | 0.0617 |  |  |
| Kurtosis: | 2.662 | Cond. | No. | 23.2 |  |  |
| ============================================================================== | | | | | | |

Таблица 14. Модифицированная модель для гипотезы (2)

Новая переменная ww, содержащая связь между переменными рабочий/выходной день и ясная/облачная погода, значима при P >|t| < 0.1 и принимает отрицательное значение при значениях рабочий день и облачная (пасмурная) погода, а значит, мы принимаем гипотезу (2). Моделирование подтверждает, что в хорошую погоду на выходных и праздниках люди больше арендуют велосипеды, чем в плохую погоду в будни.

В целом, работа модели улучшена: снизились значения показателя значимости (P >|t|) для всех переменных, не значительных по базовой модели, для модифицированной модели переменная workingday\_раб стала значимой, критерии качества работы модели улучшили показатели. Adj R2 = 0.457, что больше на 0.06, чем в базовой модели, критерий Акаике остался на прежнем уровне (1.02E+04).

Модифицируем базовую модель для гипотезы (3): cnt = a0 +a1\* workingday\_раб + a2\* weathersit\_облачно + (a20 + a21\*hum)\*temp + a4\*hum + a5\*windspeed + v = a0 +a1\* workingday\_раб + a2\* weathersit\_облачно + a20\*temp + a21\*hum\*temp + a4\*hum + a5\*windspeed + v. Если гипотеза справедлива, то a20 < 0, a21 > 0 и значим.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | \*\*\*\*\*\* | Оценка | базовой | модели | \*\*\*\*\*\* |  |
|  | OLS | Regression | Results |  |  |  |
| ============================================================================== | | | | | | |
| Dep. | Variable: | cnt | R-squared: | 0.459 |  |  |
| Model: | OLS | Adj. | R-squared: | 0.453 |  |  |
| Method: | Least | Squares | F-statistic: | 81.76 |  |  |
| Date: | Sun, 23 Oct 2022 | | Prob (F-statistic): | 6.68E-74 |  |  |
| Time: | 8:16:41 |  | Log-Likelihood: | -5074.1 |  |  |
| No. | Observations: | 585 | AIC: | 1.02E+04 |  |  |
| Df | Residuals: | 578 | BIC: | 1.02E+04 |  |  |
| Df | Model: | 6 |  |  |  |  |
| Covariance | Type: | nonrobust | |  |  |  |
| ====================================================================================== | | | | | | |
|  | Coef | std err | t | P>|t| | [0.025 | 0.975] |
| -------------------------------------------------------------------------------------- | | | | | | |
| const | 3667.296 | 839.307 | 4.369 | 0 | 2018.832 | 5315.759 |
| workingday\_раб | 224.7468 | 129.747 | 1.732 | 0.084 | -30.086 | 479.58 |
| weathersit\_облачно | -281.219 | 154.077 | -1.825 | 0.068 | -583.838 | 21.399 |
| temp | 6330.509 | 1585.909 | 3.992 | 0 | 3215.662 | 9445.356 |
| hum | -2491.77 | 1347.148 | -1.85 | 0.065 | -5137.68 | 154.129 |
| windspeed | -4563.82 | 781.236 | -5.842 | 0 | -6098.23 | -3029.41 |
| ht | 252.6386 | 2606.096 | 0.097 | 0.923 | -4865.93 | 5371.21 |
| ============================================================================== | | | | | | |
| Omnibus: | 6.704 |  | Durbin-Watson: | 2.012 |  |  |
| Prob(Omnibus): | 0.035 | Jarque-Bera | (JB): | 5.661 |  |  |
| Skew: | 0.162 |  | Prob(JB): | 0.059 |  |  |
| Kurtosis: | 2.643 | Cond. | No. | 90.2 |  |  |
| ============================================================================== | | | | | | |

Таблица 15. Модифицированная модель для гипотезы (3)

Переменная ht, связывающая влажность с температурой, оказывается незначимой. Гипотезу (3) принять не можем. Интересно, что в таком случае workingday\_раб и weathersit\_облачно становятся значимыми, «шумы» около этих переменных пропадают.

Попробуем создать модель, учитывающую оба эти преобразования. Cnt = a0 + a1\*workingday\_раб + a10\*weathersit\_облачно + a11\*workingday\_раб\*weathersit\_облачно + a20\*temp + a21\*hum\*temp + a4\*hum + a5\*windspeed + v.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | \*\*\*\*\*\* | Оценка | базовой | модели | \*\*\*\*\*\* |  |
|  | OLS | Regression | Results |  |  |  |
| ============================================================================== | | | | | | |
| Dep. | Variable: | cnt | R-squared: | 0.462 |  |  |
| Model: | OLS | Adj. | R-squared: | 0.456 |  |  |
| Method: | Least | Squares | F-statistic: | 70.91 |  |  |
| Date: | Sun, 23 Oct 2022 | | (F-statistic): | 1.06E-73 |  |  |
| Time: | 8:38:02 |  | Log-Likelihood: | -5072.3 |  |  |
| No. | Observations: | 585 | AIC: | 1.02E+04 |  |  |
| Df | Residuals: | 577 | BIC: | 1.02E+04 |  |  |
| Df | Model: | 7 |  |  |  |  |
| Covariance | Type: | nonrobust | |  |  |  |
| ====================================================================================== | | | | | | |
|  | Coef | std err | t | P>|t| | [0.025 | 0.975] |
| -------------------------------------------------------------------------------------- | | | | | | |
| Const | 3591.79 | 838.364 | 4.284 | 0 | 1945.172 | 5238.408 |
| workingday\_раб | 394.8399 | 157.42 | 2.508 | 0.012 | 85.655 | 704.025 |
| weathersit\_облачно | 105.4001 | 255.107 | 0.413 | 0.68 | -395.652 | 606.452 |
| Temp | 6288.55 | 1582.5 | 3.974 | 0 | 3180.388 | 9396.712 |
| Hum | -2532.88 | 1344.295 | -1.884 | 0.06 | -5173.19 | 107.429 |
| Windspeed | -4564.38 | 779.481 | -5.856 | 0 | -6095.35 | -3033.41 |
| Ww | -525.103 | 276.506 | -1.899 | 0.058 | -1068.19 | 17.979 |
| Ht | 270.2829 | 2600.256 | 0.104 | 0.917 | -4836.84 | 5377.404 |
| ============================================================================== | | | | | | |
| Omnibus: | 6.391 |  | Durbin-Watson: | 2.006 |  |  |
| Prob(Omnibus): | 0.041 | Jarque-Bera | (JB): | 5.585 |  |  |
| Skew: | 0.17 |  | Prob(JB): | 0.0613 |  |  |
| Kurtosis: | 2.663 | Cond. | No. | 92.3 |  |  |
| ============================================================================== | | | | | | |

Таблица 16. Модифицированная модель для гипотез 2 и 3 одновременно

Переменная ht в такой комбинации также незначима. Эта модель учитывает все сформулированные гипотезы, назовем ее итоговой.

## 3.3. Оптимизация итоговой модели, сравнение качества моделей.

Итоговую модель можно оптимизировать, удалив незначимую переменную ht. Так получится уже рассмотренная модель, модифицированная для второй гипотезы.

Итого, примем в качестве наилучшей модель №2 (таблица 14).

|  |  |  |
| --- | --- | --- |
|  | Adj R2 | Akaike |
| Базовая | 0.451 | 1.02E+04 |
| **Для гипотезы (2)** | **0.457** | **1.02E+04** |
| Для гипотезы (3) | 0.453 | 1.02E+04 |
| Для гипотез (2) и (3) | 0.456 | 1.02E+04 |

Таблица 17. Сравнение качества включенных в отчет моделей

# 4. Заключение

Гипотеза об обратной связи скорости ветра и популярности аренды велосипедов в день подтвердилась по данным всех моделей, хотя это не было ясно из графического анализа и проверок корреляции.

Гипотеза о связи ясной погоды по выходным с популярностью аренды велосипедов подтверждена для моделей №2 и №4.

Гипотеза об обратной связи температуры и влажности с популярностью аренды велосипедов не подтвердилась ни для модели №3, ни для модели №4.

# Приложение

Скрипт

# -\*- coding: utf-8 -\*-

"""

@author: YM

"""

import os

os.chdir("c:/мкс222\_09\_01/")

import pandas as pd

import numpy as np

import scipy as sp

import matplotlib as mpl

import matplotlib.pyplot as plt

import seaborn as sns

pth\_a = './data/bike\_sharing\_dataset\_day.csv'

BKSH = pd.read\_csv(pth\_a)

#обрезать датасет до нужных колонок

BS = BKSH[['workingday', 'weathersit', 'temp', 'hum','windspeed','cnt']].copy()

#гистограмма и график плотности распределения для температуры, страница 5, рисунок 1

sns.distplot(BS[['temp']], hist=True, kde=True,

bins='fd', color = 'darkblue',

hist\_kws={'edgecolor':'black'},

kde\_kws={'linewidth': 3},

axlabel = 'Температура от -8°C до +39°C')

plt.savefig('./graphics/bikes\_temphist.png', format='png')

plt.show()

BS[['temp']].describe()

BS[['temp']].skew()

BS[['temp']].kurtosis()

#гистограмма для влажности, страница 4, рисунок 2

sns.distplot(BS[['hum']], hist=True, kde=True,

bins='fd', color = 'darkorange',

hist\_kws={'edgecolor':'black'},

kde\_kws={'linewidth': 3},

axlabel = 'Влажность, доля от 0 до 1')

plt.savefig('./graphics/bikes\_humhist.png', format='png')

plt.show()

BS[['hum']].describe()

BS[['hum']].skew()

BS[['hum']].kurtosis()

#гистограмма для скорости ветра, страница 6, рисунок 3

sns.distplot(BS[['windspeed']], hist=True, kde=True,

bins='fd', color = 'darkgreen',

hist\_kws={'edgecolor':'black'},

kde\_kws={'linewidth': 3},

axlabel = 'Нормализованная скорость ветра, от 0 до 67 километров в час')

plt.savefig('./graphics/bikes\_windspeedhist.png', format='png')

plt.show()

BS[['windspeed']].describe()

BS[['windspeed']].skew()

BS[['windspeed']].kurtosis()

#гистограмма для аренды велосипедов, страница 7, рисунок 4

sns.distplot(BS[['cnt']], hist=True, kde=True,

bins='fd', color = 'darkred',

hist\_kws={'edgecolor':'black'},

kde\_kws={'linewidth': 3},

axlabel = 'Количество арендованных велосипедов по дням')

plt.savefig('./graphics/bikes\_cnthist.png', format='png')

plt.show()

BS[['cnt']].describe()

BS[['cnt']].skew()

BS[['cnt']].kurtosis()

#качественные данные: переименование категорий

BKS = BS.copy()

BKS['workingday'] = BKS['workingday'].apply(lambda x: 'раб' if x==1 else 'вых' )

def recode(x):

if x == 1:

return 'ясно'

elif x == 2:

return 'облачно'

elif x == 3:

return 'осадки'

elif x ==4:

return 'шторм'

else:

pass

BKS['weathersit'] = BKS['weathersit'].apply(recode)

BKS = BKS.astype({'workingday':'category', 'weathersit':'category', 'temp':np.float64, 'hum':np.float64, 'windspeed':np.float64, 'cnt':np.float64})

BKS.dtypes

#качественные данные: столбчатые диаграммы, страница 8, рисунок 5

dfn = BKS.select\_dtypes(include=['O', "category"])

plt.figure(figsize=(15, 9))

plt.subplots\_adjust(wspace=0.5, hspace=0.5)

nplt = 0

nrow = dfn.shape[1]

for s in dfn.columns:

nplt += 1

ax = plt.subplot(nrow, 1, nplt)

ftb = pd.crosstab(dfn[s], s)

ftb.index.name = 'Категории'

ftb.plot.bar(ax=ax, grid=True, legend=False, title=s, colormap='Pastel1')

plt.savefig('./graphics/workingday\_categories.png', format='png')

#страница 8, таблица 6

BKS['workingday'].value\_counts()

BKS['weathersit'].value\_counts()

#по итогам разведочного анализа перекодируем некоторые данные

BKS['weathersit'] = BKS['weathersit'].replace('осадки', 'облачно')

BKS['weathersit'].value\_counts()

#для нового разделения на категории посмотрим столбчатую диаграмму, страница 9, рисунок 6

BKS['weathersit'].value\_counts().plot(kind='bar')

plt.savefig('./graphics/weathersit\_TWOcategories.png', format='png')

#анализ связи

#между качественными объясняющими и количественной целевой, страница 9, рисунок 7

dfn = BKS.copy()

cols = dfn.select\_dtypes(include='category').columns

nrow = len(cols)

fig, ax\_lst = plt.subplots(nrow, 1)

fig.figsize=(15, 12)

nplt = -1

for s in cols:

nplt += 1

# Доверительные интервалы строятся методом бутстрепа

dfn.boxplot(column='cnt', by=s, ax=ax\_lst[nplt], grid=True, notch=True,

bootstrap=50, showmeans=True, color=None)

fig.subplots\_adjust(wspace=0.5, hspace=1.0)

# Общая подпись к графикам

fig.suptitle('Категоризированные диагарммы Бокса-Вискера')

plt.savefig('./graphics/bikes\_whisker.png', format='png')

#критерий Крускала-Уоллиса для строгой проверки предположений, страница 10

from scipy.stats import kruskal

# Качественная переменная - 'workingday'

# Создаем подвыборки

sel\_yes = BKS['workingday']=='вых'

x\_1 = BKS.loc[sel\_yes, 'cnt']

sel\_no = BKS['workingday']=='раб'

x\_2 = BKS.loc[sel\_no, 'cnt']

# Используем криетрий Крускала-Уоллиса

bike\_sig = kruskal(x\_1, x\_2)

# Сохраняем текстовый отчет

with open('./output/bikes\_workingday\_STAT.txt', 'w') as fln:

print('Критерий Крускала-Уоллиса для переменных \'workingday\' и \'cnt\'',

file=fln)

print(bike\_sig, file=fln)

sel\_yes = BKS['weathersit']=='ясно'

x\_1 = BKS.loc[sel\_yes, 'cnt']

sel\_no = BKS['weathersit']=='облачно'

x\_2 = BKS.loc[sel\_no, 'cnt']

# Используем криетрий Крускала-Уоллиса

bike\_sig = kruskal(x\_1, x\_2)

# Сохраняем текстовый отчет

with open('./output/bikes\_weathersit\_STAT.txt', 'w') as fln:

print('Критерий Крускала-Уоллиса для переменных \'weathersit\' и \'cnt\'',

file=fln)

print(bike\_sig, file=fln)

#между количественными объясняющими и количественной целевой, страница 10, рисунок 8

dfn = BKS.select\_dtypes(include='float64')

nrow = dfn.shape[1] - 1 # Учитываем, что одна переменная целевая - ось 'Y'

fig, ax\_lst = plt.subplots(nrow, 1)

fig.figsize=(15, 9)

nplt = -1

for s in dfn.columns[:-1]: # Последняя переменная - целевая ('Y')

nplt += 1

dfn.plot.scatter(s, 'cnt', ax=ax\_lst[nplt])

ax\_lst[nplt].grid(visible=True)

ax\_lst[nplt].set\_title(f'Связь количества арендованных велосипедов с {s}')

fig.subplots\_adjust(wspace=0.5, hspace=1.0)

"""

Общая подпись к графикам

Используем форматированные 'f'-строки

{} - позиция для подстановки значения

"""

fig.suptitle(f'Связь количества арендованных велосипедов с {list(dfn.columns[:-1])}')

plt.savefig('./graphics/bikes\_scat.png', format='png')

# Анализ корреляции между количественными переменными, страницы 10-11, таблицы 8-9

# Используем библиотеку scipy

from scipy.stats import pearsonr

from scipy.stats import spearmanr

BK = BKS.select\_dtypes(include='float')

# Здесь будут значения оценок коэффициента корреляции Пирсона

C\_P = pd.DataFrame([], index=BK.columns, columns=BK.columns)

# Здесь будут значения значимости оценок коэффициента корреляции Пирсона

P\_P = pd.DataFrame([], index=BK.columns, columns=BK.columns)

# Здесь будут значения оценок коэффициента корреляции Спирмена

C\_S = pd.DataFrame([], index=BK.columns, columns=BK.columns)

# Здесь будут значения значимости оценок коэффициента корреляции Спирмена

P\_S = pd.DataFrame([], index=BK.columns, columns=BK.columns)

for x in BK.columns:

for y in BK.columns:

C\_P.loc[x,y], P\_P.loc[x,y] = pearsonr(BK[x], BK[y])

C\_S.loc[x,y], P\_S.loc[x,y] = spearmanr(BK[x], BK[y])

# Сохраняем текстовый отчет на разные листы Excel файла

with pd.ExcelWriter('./output/bikes\_STAT.xlsx', engine="openpyxl") as wrt:

# Корреляция Пирсона

C\_P.to\_excel(wrt, sheet\_name='Pirson')

dr = C\_P.shape[0] + 2

P\_P.to\_excel(wrt, startrow=dr, sheet\_name='Pirson') # Значимость

# Корреляция Спирмена

C\_S.to\_excel(wrt, sheet\_name='Spirmen')

dr = C\_S.shape[0] + 2

P\_S.to\_excel(wrt, startrow=dr, sheet\_name='Spirmen') # Значимость

#страница 11, рисунок 9

corr = BK.corr(method = 'spearman')

sns.heatmap(corr, annot = True)

plt.savefig('./graphics/bikes\_heatmap.png', format='png')

#таблица корреляции и хи-квадрат для качественных переменных, страница 12, таблица 10

ct\_table = pd.crosstab(BKS['workingday'], BKS['weathersit'])

dt = np.array([[75, 156], [193, 307]])

x = sp.stats.chi2\_contingency(dt)[0]

#calculate Cramer's V

n = np.sum(dt)

minDim = min(dt.shape)-1

np.sqrt((x / n) / minDim)

#моделирование

import statsmodels.api as sm

B = BKS.copy()

# Разбиение данных на тренировочное и тестовое множество

# frac- доля данных в тренировочном множестве

# random\_state - для повторного отбора тех же элементов

B\_train = B.sample(frac=0.8, random\_state=42)

# Символ ~ обозначает отрицание (not)

B\_test = B.loc[~B.index.isin(B\_train.index)]

# Будем накапливать данные о качестве постреонных моделей

# Используем adjR^2 и AIC

mq = pd.DataFrame([], columns=['adjR^2', 'AIC']) # Данные о качестве

"""

Постреоние базовой модели - таблица 11

Базовая модель - линейная регрессия, которая включает в себя

все количественные переменные и фиктивные переменные дял качественных

переменных с учетом коллинеарности. Для каждого качетсвенного показателя

включаются все уровни за исключением одного - базового.

"""

# Формируем целевую переменную

Y = B\_train['cnt']

# Формируем фиктивные (dummy) переменные для всех качественных переменных

DUM = pd.get\_dummies(B\_train[['workingday', 'weathersit']])

# Выбираем переменные для уровней, которые войдут в модель

# Будет исключен один - базовый. ВЛияние включенных уровней на зависимую

# переменную отсчитывается от него

DUM = DUM[['workingday\_раб', 'weathersit\_облачно']]

# Формируем pandas.DataFramee содержащий матрицу X объясняющих переменных

# Добавляем слева фиктивные переменные

X = pd.concat([DUM, B\_train[['temp', 'hum', 'windspeed']]], axis=1)

# Добавляем переменную равную единице для учета константы

X = sm.add\_constant(X)

X = X.astype({'const':'uint8'}) # Сокращаем место джля хранения константы

# Формируем объект, содержащий все исходные данные и методы для оценивания

linreg00 = sm.OLS(Y,X)

# Оцениваем модель

fitmod00 = linreg00.fit()

# Сохраняем результаты оценки в файл

with open('./output/bikes\_STAT\_woWeather.txt', 'a') as fln:

print('\n \*\*\*\*\*\* Оценка базовой модели \*\*\*\*\*\*',

file=fln)

print(fitmod00.summary(), file=fln)

# Проверяем степень мультиколлинеарности только базовой модели, таблица 12

from statsmodels.stats.outliers\_influence import variance\_inflation\_factor

vif = pd.DataFrame() # Для хранения

X\_q = X.select\_dtypes(include='float64')# Только количественные регрессоры

vif["vars"] = X\_q.columns

vif["VIF"] = [variance\_inflation\_factor(X\_q.values, i)

for i in range(X\_q.shape[1])]

# Сохраняем полученные результаты

with pd.ExcelWriter('./output/bikes\_STAT.xlsx', engine="openpyxl",

if\_sheet\_exists='overlay', mode='a') as wrt:

vif.to\_excel(wrt, sheet\_name='vif')

# Проверяем гетероскедастичность базовой модели, таблица 13

# помощью коритерия White(а) и F критерия

from statsmodels.stats.diagnostic import het\_white

e = fitmod00.resid

WHT = pd.DataFrame(het\_white(e, X), index= ['LM', 'LM\_P', 'F', 'F\_P'])

# Сохраняем полученные результаты

with pd.ExcelWriter('./output/bikes\_STAT.xlsx', engine="openpyxl",

if\_sheet\_exists='overlay', mode='a') as wrt:

WHT.to\_excel(wrt, sheet\_name='het')

# Сохраняем данные о качестве модели

q = pd.DataFrame([fitmod00.rsquared\_adj, fitmod00.aic],

index=['adjR^2', 'AIC'], columns=['base\_00']).T

mq = pd.concat([mq, q])

#модификация для гипотезы 2, статус дня к погодным условиям, таблица 14

# Вводим переменную вщаимодействия

X\_1 = X.copy()

X\_1['ww'] = X\_1['workingday\_раб']\*X\_1['weathersit\_облачно']

linreg02 = sm.OLS(Y,X\_1)

fitmod02 = linreg02.fit()

# Сохраняем результаты оценки в файл

with open('./output/bikes\_stat\_ww.txt', 'a') as fln:

print('\n \*\*\*\*\*\* Оценка базовой модели \*\*\*\*\*\*',

file=fln)

print(fitmod02.summary(), file=fln)

# Сохраняем данные о качестве модели

q = pd.DataFrame([fitmod02.rsquared\_adj, fitmod02.aic],

index=['adjR^2', 'AIC'], columns=['hyp\_01']).T

mq = pd.concat([mq, q])

#модификация для гипотезы 3, влажность к температуре, таблица 15

X\_2 = X.copy()

X\_2['ht'] = X\_1['temp']\*X\_1['hum']

linreg03 = sm.OLS(Y,X\_2)

fitmod03 = linreg03.fit()

# Сохраняем результаты оценки в файл

with open('./output/bikes\_stat\_ht.txt', 'a') as fln:

print('\n \*\*\*\*\*\* Оценка базовой модели \*\*\*\*\*\*',

file=fln)

print(fitmod03.summary(), file=fln)

# Сохраняем данные о качестве модели

q = pd.DataFrame([fitmod03.rsquared\_adj, fitmod03.aic],

index=['adjR^2', 'AIC'], columns=['hyp\_01']).T

mq = pd.concat([mq, q])

#модицификация для гипотез 2 и 3, таблица 16

X\_12 = X.copy()

X\_12['ww'] = X\_12['workingday\_раб']\*X\_12['weathersit\_облачно']

X\_12['ht'] = X\_12['temp']\*X\_12['hum']

linreg04 = sm.OLS(Y,X\_12)

fitmod04 = linreg04.fit()

# Сохраняем результаты оценки в файл

with open('./output/bikes\_stat\_pair.txt', 'a') as fln:

print('\n \*\*\*\*\*\* Оценка базовой модели \*\*\*\*\*\*',

file=fln)

print(fitmod04.summary(), file=fln)

# Сохраняем данные о качестве модели

q = pd.DataFrame([fitmod04.rsquared\_adj, fitmod04.aic],

index=['adjR^2', 'AIC'], columns=['hyp\_01']).T

mq = pd.concat([mq, q])