Федеральное государственное автономное образовательное учреждение высшего образования

«Национальный исследовательский университет ИТМО»

Факультет технологического менеджмента и инноваций

Анализ данных при принятии управленческих решений

Отчет

Реализация линейной множественной регрессии

Выполнил: студент
Шапоренко Ангелина
Евгеньевна, 4 курс, группа
U3477 Преподаватель:
Духанов Алексей Валентинович

Санкт-Петербург
2025

СОДЕРЖАНИЕ

ЦЕЛЬ РАБОТЫ

Освоить методы построения и анализа многофакторной линейной регрессии, научиться отбирать статистически значимые факторы, оценивать качество модели и проверять её соответствие условиям теоремы Гаусса-Маркова для обеспечения корректности и надёжности полученных результатов.

ИНФОРМАЦИЯ О ВЫБРАННОМ ДАТАСЕТЕ

Название датасета: Bike Sharing

Датасет содержит 17 тысяч строк, включает категориальные, бинарные и числовые признаки. Каждая запись описывает количество арендованных велосипедов Capital Bikeshare в 2011-2012 годах по часам и погодные, сезонные условия.

Выбранные колонки:

- Зависимая переменная (y): ‘count’ – количество арендованных за час велосипедов.
- Факторы (x): ‘temp’ – температура в градусах Цельсия, линейно нормированная; ‘atemp’ – ощущаемая температура в градусах Цельсия, линейно нормированная; ‘hum’ – влажность воздуха, линейно нормированная; ‘windspeed’ – скорость ветра, линейно нормированная.

ОПИСАНИЕ ОБРАБОТКИ ДАННЫХ

Из датасета были убраны следующие колонки: ‘dteday’(дата), поскольку сезонные тренды гипотетически будут нелинейными, ‘season’ (время года), ‘yr’ (год), ‘mnth’ (месяц), ‘hr’ (часы), ‘holiday’ (праздник/нет), ‘weekday’ (день недели), ‘workingday’ (рабочий день/нет), ‘weathersit’ (словесная характеристика погоды), так как они являются категориальными.

Датасет не содержит пропущенных значений, что указано в метаданных. Была осуществлена дополнительная проверка для подтверждения данной информации.

КОЭФФИЦИЕНТЫ ПЕРВОЙ ЛИНЕЙНОЙ РЕГРЕССИИ И ЗНАЧЕНИЯ УКАЗАННЫХ ПОКАЗАТЕЛЕЙ

Была построена первая модель по выбранным непрерывным факторам.

Получившиеся коэффициенты (результата fit_results.summary):

- const (свободный член) = 161.8069
- temp = 85.5765
- atemp = 314.3429
- hum = −275.1803
- windspeed = 42.9793

Самые сильные факторы: atemp, hum. Признаки temp и windspeed тоже значимы, но их вклад меньше.

Основные показатели первой модели:

- F-value = 1474.1302776485202
- R^2 = 0.25339016185201135
- MSE = 24563.141088011096

Построенная модель линейной регрессии является статистически значимой (F = 1474.13, Prob (F-statistic) = 0.00) и объясняет около 25 % вариации зависимой переменной (R^2 = 0.253), следовательно, есть другие важные факторы (возможно, день недели, время суток, сезон, осадки и т. д.). При этом среднеквадратическая ошибка составляет 24 563, что свидетельствует о наличии значительной не предсказанной части данных. Для повышения

качества модели целесообразно рассмотреть добавление дополнительных предикторов и/или нелинейных зависимостей.

Мера мультиколлинеарности (VIF):

VIF вычислялся по признакам temp, atemp, hum, windspeed:

- VIF(temp) ≈ 43.55
- VIF(atemp) ≈ 43.65
- VIF(hum) ≈ 1.10
- VIF(windspeed) ≈ 1.16

VIF сильно больше 10 для temp и atemp – признак сильной мультиколлинеарности между этими двумя переменными.

МАТРИЦА КОРРЕЛЯЦИЙ

Матрица корреляций:

	temp	**atem**	**hum**	**windspeed**	**cnt**
temp	1,000	0.9877	–0.0699	–0.0231	0.4048
atem	0.9877	1.000	–0.0519	–0.0623	0.4009
hum	−0.0699	–0.0519	1.000	–0.2901	–0.3229
windspeed	−0.0231	–0.0623	–0.2901	1.000	0.0932
cnt	0.4048	0.4009	–0.3229	0.0932	1.000

По результатам регрессии значимостью обладают все факторы (p-value < 0.05), но:

1. VIF и матрица корреляции указывают на высокую степень взаимосвязи между temp и atemp (≈ 0.988).
2. windspeed наименее коррелирует с зависимой переменной (≈ 0.093)

На основании этого уберем из рассмотрения temp и windspeed.

Отобранные факторы: atem и hum.

Фактор atem был выбран потому, что надо исключить дублирующие/почти линейно зависимые признаки и оставить один из пары temp/atemp, а atemp имеет более высокую t-статистику и более небольшое относительное стандартное отклонение в модели. Кроме того, прогнозирование показывает, что «ощущаемая температура» лучше отражает поведение людей, чем просто температура воздуха.

Фактор hum был выбран, так как он не связан мультиколлинеарностью с другими факторами и p-value для hum меньше 0.001.

Статистика по исследованию факторов на статистическую значимость по выбранному критерию: t-статистика сравнивается с критическим значением $t \approx 1.96$. Для atem $t = 6.876$ (>1.96) и $p < 0.05$, для hum $t = |-42.560|$ (>1.96) и $p < 0.05$. Оба коэффициента статистически значимы.

КОЭФФИЦИЕНТЫ НОВОЙ ЛИНЕЙНОЙ РЕГРЕССИИ И ЗНАЧЕНИЙ УКАЗАННЫХ ПОКАЗАТЕЛЕЙ

Была построена вторая модель после исключения факторов temp и windspeed

Коэффициенты регрессии:

- const = 174.6504
- atemp = 406.5809
- hum = −284.7899

Значения показателей:

- F-value = 2930.8970086914005
- R^2 = 0.25225260250598414
- MSE = 24600.566298991125

- VIF (atemp, hum) ≈ 1.0027, 1.0027

После исключения переменных temp и windspeed модель сохранила статистическую значимость ($F = 2930.9$, $p < 0.001$) и практически тот же уровень объясняющей способности ($R^2 = 0.252$). Среднеквадратическая ошибка (MSE = 24 600.6) изменилась несущественно. Также модель была избавлена от мультиколлинеарности (VIF близки к 1). Следовательно упрощённая модель с переменными atemp и hum является более надёжной и более интерпретируемой.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ ЦЕЛЕСООБРАЗНОСТИ ИСКЛЮЧЕНИЯ ФАКТОРОВ

Для того, чтобы проверить нулевую гипотезу, что добавленные в полную модель параметры не дают значимого улучшения по сравнению с упрощённой моделью, использовался F-тест.

Результат: $F = 13.235799388139272$, $p \approx 1.8036e^{-06}$, $df = 2.0$

p-value сильно меньше 0.05, следовательно старая модель эффективнее новой.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ МОДЕЛИ НА ПРЕДМЕТ СООТВЕТСТВИЯ ТЕОРЕМЕ ГАУССА-МАРКОВА

Проверка проводился на второй (упрощенной) модели

Были проверены 4 аспекта: случайность остатков, нормальность, независимость (автокорреляция) и равенство суммы остатков нулю.

1. Случайность остатков:

Результаты Runs Test:

- runs = 2725

- exp_runs = 8690.0
- z = −90.5009
- p-value ≈ 0.0

Фактическое число серий (runs) намного меньше ожидаемого (exp_runs), то есть остатки очень сгруппированы по знакам.

Z-статистика очень большая по модулю при p < 0.05, следовательно, отклонение от случайного порядка экстремально. Это значит, что остатки не случайны и имеет место автокорреляция.

2. Проверка нормальности остатков

Проверка нормальности остатков была проведена с помощью тестов Skew, Kurtosis, Шапиро-Уилка.

- Skew (асимметрия) = 1.2319
- Kurtosis (эксцесс) = 1.9527
- Shapiro–Wilk: W = 0.9194, p-value ≈ 3.1255e-69

skew > 1 – сильная правосторонняя асимметрия распределеиня остатков

kurtosis < 3 – плосковершинное распределение

shapiro_stat < 0.95 при p << 0.05 – распределение далеко от нормального

3. Независимость остатков:

Независимость остатков была проверена с помощью теста Дарбина-Уотсона.

- Durbin–Watson = 0.4200859765008472

DW ≈ 0.42 (значительно < 2) – сильная положительная автокорреляция остатков. Это подтверждает вывод Runs Test.

4. Равенство суммы остатков нулю:

Был проведен t-тест о равенстве среднего значения остатков нулю.

- t ≈ 7.51e-13, p-value ≈ 1.0

  t_stat близок к нулю при p > 0.05 – нет оснований отвергнуть H_0: среднее остатков = 0. Модель в среднем не смещает предсказания (в среднем ошибается около нуля). Это одно из требований Гаусса–Маркова выполнено.

Итог исследования модели на предмет соответствия теореме Гаусса-Маркова: четко выполняются только два критерия: линейность модели (линейна по коэффициентам) и нулевая средняя остатков. Остальные критерии: отсутствие автокорреляции остатков, нормальность распределения остатков нарушены. Следовательно наши коэффициенты не максимально надежны и точны для линейной модели. Модель даёт несмещённые, но неэффективные оценки

ВЫВОД

В ходе выполнения работы были построены и проанализированы две модели для предсказания количества аренды велосипедов по погодным признакам.

Модель объясняет только ~25% вариации целевой переменной – есть значимая доля неопределённости, вероятно связанная с временными факторами и/или неучтёнными переменными.

Полная модель показала мультиколлинеарность между temp и atemp; после исключения temp и windspeed получена стабильная, интерпретируемая модель с atemp и hum, но F-тест показал, что старая модель эффективнее новой.

Проверка на соответствие теореме Гаусса-Маркова показала, что основная проблема модели – сильная автокорреляция и неслучайность остатков, а также отклонение от нормальности.