

**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**

Выполнили: студент  
Лемешева София Андреевна,  
4 курс, группа U3475  
Преподаватель:  
Духанов Алексей Валентинович

**Санкт-Петербург  
2025**

## СОДЕРЖАНИЕ

Цель работы.....	3
Информация о выбранном датасете.....	3
Описание обработки данных.....	3
Матрица корреляций.....	5
Матрица корреляций (Pearson):.....	5
Коэффициенты новой линейной регрессии и значений указанных показателей.....	6
Результаты исследования целесообразности исключения факторов.....	7
Результаты исследования модели на предмет соответствия теореме Гаусса-Маркова.....	7
Общий вывод:.....	8

## Цель работы

Целью данной работы является приобретение практических навыков в проведении факторного анализа с использованием модели линейной множественной регрессии. Это включает построение модели зависимости целевой переменной от нескольких факторов, отбор значимых факторов, оценку качества модели и проверку ее соответствия ключевым предположениям теоремы Гаусса-Маркова. В результате ожидается понимание, как факторы влияют на зависимую переменную, и интерпретация статистических показателей для обоснованных выводов.

## Информация о выбранном датасете

Название датасета: `synthetic_coffee_health_10000.xlsx` (в коде используется название `"for_G.xlsx"` для удобства).

Выбранные количественные колонки:

- Зависимая переменная (y) формируется из категориальной колонки 'Stress\_Level' (Low/Medium/High): 'Y' — числовое представление уровня стресса (1 — Low, 2 — Medium, 3 — High; трактуется как количественная для регрессионного анализа).
- Факторы (независимые переменные, x): 'Age' (возраст), 'Coffee\_Intake' (потребление кофе), 'Caffeine\_mg' (количество кофеина в мг), 'Sleep\_Hours' (часы сна), 'BMI' (индекс массы тела), 'Heart\_Rate' (частота сердечных сокращений), 'Physical\_Activity\_Hours' (часы физической активности).
- Датасет содержит 10 000 наблюдений без явных временных рядов, но подходит для регрессионного моделирования.

## Описание обработки данных

Данные были загружены из файла Excel с использованием второй строки в качестве заголовков колонок (`header=1`). Для обработки пропущенных значений (`missing values`) применена функция `data.dropna()`, которая удаляет все строки, содержащие NaN. В предоставленном датасете пропущенных значений не обнаружено, поэтому этот шаг не изменил объем данных (осталось 10 000 наблюдений). Это обеспечивает отсутствие

пустых значений в выбранных колонках, как требуется в задании, и предотвращает ошибки в регрессионном анализе.

Также в анализе не принимали участие такие колонки как Gender, Country, Coffee\_Intake, Caffeine\_mg, Sleep\_Quality, Health\_Issues, Occupation, Alcohol\_Consumption - поскольку данные колонки имеют категориальный тип данных (потенциально, они могли бы приниматься за зависимую переменную, однако для этих целей была принята колонка Stress\_Level, поскольку она принимает только 3 значения (описано выше))

### **Коэффициенты первой линейной регрессии и значения указанных показателей**

Первая модель линейной регрессии построена на всех выбранных факторах с использованием метода наименьших квадратов (OLS) из библиотеки statsmodels.

Коэффициенты регрессии (из summary модели):

- const (свободный член): 4.2086
- Age: -3.444e-05
- Coffee\_Intake: -0.0613
- Caffeine\_mg: 0.0006
- Sleep\_Hours: -0.4268
- BMI (индекс массы тела): 0.0013
- Heart\_Rate: -8.85e-05
- Physical\_Activity\_Hours: -0.0003

Значения показателей:

- Коэффициент детерминации ( $R^2$ ): 0.6297 (модель объясняет около 63% вариации зависимой переменной Y; это указывает на умеренную объяснительную силу, но есть пространство для улучшения).
- Средняя квадратическая ошибка (MSE): 0.1599 (среднее квадратичное отклонение предсказаний от реальных значений; низкое значение говорит о хорошей точности, но сравнение с второй моделью покажет изменения).

- Показатель системного эффекта факторов (F-статистика): 2426.90 (Prob (F-statistic): 0.00, что подтверждает общую значимость модели на уровне  $p < 0.05$ ; факторы в совокупности влияют на Y).
- Мера мультиколлинеарности (VIF для каждого фактора):
  - Age: 1.0006
  - Coffee\_Intake: 2684.94 (очень высокое, указывает на сильную мультиколлинеарность с Caffeine\_mg)
  - Caffeine\_mg: 2685.12 (аналогично, высокая корреляция с Coffee\_Intake)
  - Sleep\_Hours: 1.0386
  - BMI (индекс массы тела): 1.0007
  - Heart\_Rate: 1.0044
  - Physical\_Activity\_Hours: 1.0003

Высокие VIF для Coffee\_Intake и Caffeine\_mg (свыше 10) сигнализируют о проблеме мультиколлинеарности, что может искажать оценки коэффициентов и их значимость.

## Матрица корреляций

Матрица корреляций (Pearson):

	Y	Age	Coffee_Intake	Caffeine_mg	Sleep_Hours	BMI	Heart_Rate	Physical_Activity_Hours
Y	1.0000	-0.0044	0.1497	0.1499	-0.7935	0.0012	0.0273	0.0067
Age	-0.0044	1.0000	-0.0122	-0.0118	0.0050	0.0086	-0.0002	0.0059
Coffee_Intake	0.1497	-0.0122	1.0000	0.9998	-0.1903	-0.0083	0.0601	0.0048

Caffeine_mg	0.1499	-0.0118	0.9998	1.0000	-0.1905	-0.0087	0.0600	0.0050
Sleep_Hours	-0.7935	0.0050	-0.1903	-0.1905	1.0000	0.0085	-0.0362	-0.0112
BMI (индекс массы тела)	0.0012	0.0086	-0.0083	-0.0087	0.0085	1.0000	-0.0094	0.0020
Heart_Rate	0.0273	-0.0002	0.0601	0.0600	-0.0362	-0.0094	1.0000	-0.0029
Physical_Activity_Hours	0.0067	0.0059	0.0048	0.0050	-0.0112	0.0020	-0.0029	1.0000

Отобранные факторы: Sleep\_Hours, BMI (индекс массы тела).

Обоснование отбора: Факторы выбраны на основе двух критериев — статистической значимости коэффициентов в первой модели ( $p\text{-value} < 0.05$  по  $t$ -статистике) и корреляции с зависимой переменной  $Y$  ( $|\text{corr}| > 0.1$ ).

- Sleep\_Hours:  $p=0.000$  ( $<0.05$ ),  $\text{corr}=-0.7935$  ( $>0.1$  по модулю); сильная отрицательная связь с  $Y$ .
- BMI (индекс массы тела):  $p=0.192$  (изначально  $>0.05$ , но выбран как второй по значимости для минимума 2 факторов; в коде использована логика топ-2 по наименьшим  $p\text{-values}$ , где Sleep\_Hours — лидер, BMI — следующий).
- Другие факторы (например, Coffee\_Intake:  $p=0.668$ ,  $\text{corr}=0.1497$ ) исключены из-за низкой значимости или слабой корреляции.

Статистика значимости:  $t$ -статистика сравнивается с критическим значением  $t$  (для  $df=9992$ ,  $\alpha=0.05$ , критическое  $t \approx 1.96$ ). Для Sleep\_Hours

$t=-127.943$  ( $>|1.96|$ ), значимо; для BMI  $t=1.305$  ( $<1.96$ , но сохранен для минимума факторов).

### **Коэффициенты новой линейной регрессии и значений указанных показателей**

Новая модель построена только на отобранных факторах.

Коэффициенты регрессии:

- const: 4.1965
- Sleep\_Hours: -0.4266
- BMI (индекс массы тела): 0.0013

Значения показателей:

- $R^2$ : 0.6296 (vs первая: 0.6297) — практически не изменился, модель сохраняет объяснительную силу.
- MSE: 0.1599 (vs первая: 0.1599) — идентично, точность не ухудшилась.
- F-статистика: 8497.83 (vs первая: 2426.90) — выше, модель значима (Prob=0.00).
- VIF:
  - Sleep\_Hours: 1.0001
  - BMI (индекс массы тела): 1.0001

Сравнение:  $R^2$  и MSE почти идентичны, F-статистика выросла из-за снижения степеней свободы, VIF значительно улучшился (нет мультиколлинеарности).

Интерпретация: Упрощение модели за счет исключения незначимых факторов не привело к потере качества, но повысило надежность оценок (ниже VIF). Это делает модель более интерпретируемой: Y в основном зависит от часов сна (отрицательно) и слабо от BMI.

### **Результаты исследования целесообразности исключения факторов**

По критерию Фишера (F-test между полной и редуцированной моделями):  $F=0.0826$ ,  $p\text{-value}=0.995$ ,  $df=5$ .

Поскольку  $p > 0.05$ ,  $H_0$  о равенстве моделей не отвергается — исключение факторов целесообразно, так как упрощенная модель не значимо хуже полной. Это подтверждает, что отброшенные факторы не вносили существенный вклад.

## **Результаты исследования модели на предмет соответствия теореме Гаусса-Маркова**

Проверка проведена на второй модели.

- Случайность остатков: Критерий поворотных точек (runs test): Observed runs=5065, Expected $\approx$ 4999.5,  $Z=1.310$ ,  $p=0.190$  ( $>0.05$ ) — остатки случайны ( $H_0$  не отвергается). Асимметрия (skewness): 0.348, Эксцесс (kurtosis): -0.531. Shapiro-Wilk: stat=0.973,  $p<0.001$  — остатки не нормально распределены (отвергается нормальность, но для больших выборок тест чувствителен; skewness и kurtosis близки к 0, распределение приемлемо).
- Независимость остатков: Дарбин-Уотсон: 2.018 (в диапазоне 1.5-2.5) — нет автокорреляции, остатки независимы.
- Равенство суммы остатков нулю: Сумма= $9.81e-11$  ( $\approx 0$ ), Среднее= $9.81e-15$  ( $\approx 0$ ). t-Стьюдента:  $t=2.45e-12$ ,  $p=1.00$  ( $>0.05$ ) — среднее не отличается от 0.

Вывод: модель в целом соответствует условиям Гаусса-Маркова (случайность, независимость, нулевое среднее остатков), несмотря на отклонение от нормальности (что менее критично для больших выборок).

### **Общий вывод:**

В ходе выполнения задания по реализации линейной множественной регрессии на датасете `synthetic_coffee_health_10000.xlsx` (в коде — `for_G.xlsx`) были достигнуты поставленные цели. Получены практические навыки построения и анализа модели регрессии, включая отбор значимых факторов, оценку качества и проверку условий Гаусса-Маркова.

Ключевые результаты:

- Модель объясняет около 63% вариации уровня стресса (Y), что указывает на умеренную предсказательную силу. Основной фактор



влияния — часы сна (Sleep\_Hours, коэффициент -0.4266), показывающий отрицательную зависимость: больше сна снижает стресс.

- Отбор факторов (Sleep\_Hours и BMI) позволил упростить модель без потери качества ( $R^2$  и MSE стабильны, VIF снижен до 1), подтверждено F-тестом ( $p=0.995$ ).
- Модель соответствует большинству условий Гаусса-Маркова: остатки случайны, независимы, с нулевым средним. Отклонение от нормальности не критично для большой выборки ( $n=10000$ ).