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
Терешина Мария Андреевна,
4 курс, группа U3475
Преподаватель:
Духанов Алексей Валентинович

Санкт-Петербург

2025

СОДЕРЖАНИЕ

Цель работы

Целью лабораторной работы является освоение методов построения и анализа моделей линейной множественной регрессии. В ходе выполнения работы проводится исследование значимости факторов, выявление мультиколлинеарности, отбор оптимального подмножества переменных, сравнение полной и сокращенной моделей, а также проверка выполнения условий теоремы Гаусса–Маркова.

Информация о датасете

Для исследования выбран открытый датасет Fish Market, размещенный на платформе Kaggle. Датасет содержит 159 наблюдений, каждое из которых соответствует отдельному экземпляру рыбы. В таблице представлены следующие характеристики: Weight (масса рыбы, г) — зависимая переменная, Length1, Length2, Length3 (три различных измерения длины, мм), Height (высота тела, мм), Width (толщина тела, мм). Категориальная переменная Species была исключена из анализа, так как имеет строковый тип данных и не может быть использована в регрессионной модели.

Предобработка данных

На этапе подготовки данных были оставлены только числовые переменные. Проверка на пропуски показала их отсутствие. Все признаки имели удобный числовой формат, что позволило сразу перейти к построению модели.

Построение полной модели

В первую регрессионную модель были включены все пять факторов: Length1, Length2, Length3, Height, Width. Коэффициент детерминации R^2=0.885, скорректированный R^2=0.882. Среднеквадратичная ошибка (MSE) =

14607.9, показатель системного эффекта факторов η = -3.20. Единственным статистически значимым фактором оказался Height (p=0.0015). Значения VIF для Length1, Length2 и Length3 превышали 400–2000, что свидетельствует о сильной мультиколлинеарности. Height (14.6) и Width (12.3) имели более умеренные значения VIF. Так, полная модель объясняет 88.5% изменчивости массы, но страдает от мультиколлинеарности и незначимости коэффициентов.

Корреляции и отбор факторов

Была построена матрица корреляций между факторами, так также проведена оценка коэффициентов регрессии на статистическую значимость. Переменная Height, несмотря на более умеренную с зависимым признаком (около 0.73), единственная показала статистическую значимость в регрессионной модели (p-value < 0.05).

Таблица 1 – Коэффициенты корреляции

	Length1	Length2	Length3	Height	Width
Length1	1.000	0.999	0.992	0.625	0.867
Length2	0.999	1.000	0.994	0.640	0.874
Length3	0.992	0.994	1.000	0.703	0.879
Height	0.625	0.640	0.703	1.000	0.793
Width	0.867	0.874	0.879	0.793	1.000

Анализ значений матрицы корреляций показывает, что между длинами (Length1, Length2 и Length3) наблюдаются крайне высокие коэффициенты корреляции (близкие к 1), что указывает на сильную мультиколлинеарность. Использование всех трех переменных одновременно приводит к дублированию информации и снижает устойчивость модели.

Среди факторов наибольшую корреляцию с зависимой переменной Weight демонстрирует показатель Length3 (около 0.93), что подтверждает его высокую информативность и значимость при прогнозировании массы рыбы.

Таким образом, для сокращенной модели были отобраны переменные Length3 и Height.

Построение сокращенной модели

Сокращенная модель построена с использованием Height и Length3. R^2=0.863, скорр. R^2=0.861, MSE=17425.1. Значения VIF около 2, что указывает на отсутствие мультиколлинеарности. Оба коэффициента значимы. То есть сокращенная модель обладает чуть меньшей точностью, но отличается устойчивостью и интерпретируемостью.

Сравнение моделей

Сравнение моделей по F-тесту: F=9.836, p=0.0000. Формально полная модель статистически лучше, но практическая ценность ее ниже из-за мультиколлинеарности и незначимости факторов.

Проверка условий теоремы Гаусса–Маркова

Для сокращенной модели сумма остатков равна -0.0. t-тест показал, что среднее значение остатков статистически не отличается от нуля. Критерий поворотных точек (K=93, p=0.0273) выявил нарушение случайности. Критерий Дарбина–Уотсона равен 0.425, что указывает на сильную положительную автокорреляцию. Асимметрия = 0.35, эксцесс = 0.15, что близко к нормальному распределению. Вывод: выполнены условия несмещенности и нормальности, но нарушена независимость остатков. Возможной причиной является исключение фактора Species из построения модели.

Заключение

В ходе работы была реализована множественная линейная регрессия на данных о характеристиках рыб (Fish Market, Kaggle). Полная модель (5 факторов) объяснила 88.5% изменчивости массы, но страдала от мультиколлинеарности. Сокращенная модель (Height и Length3) имела R^2=0.863 и показала более устойчивые результаты. F-тест подтвердил преимущество полной модели, но практическая ценность выше у сокращенной. Проверка условий Гаусса–Маркова подтвердила несмещенность и нормальность, однако выявила автокорреляцию остатков. Можно сделать вывод, что для стабильного практического применения оптимальна сокращенная модель.

Для более подробного исследования можно построить другие варианты линейной регрессии с разными наборами элементов. Кроме того, можно закодировать категориальный признак Species и также использовать его при построении регрессии, а не исключать его.