Отчёт по лабораторной работе: Анализ факторов стресса студентов с использованием линейной регрессии

Смольяковой Анны Андреевны (группа: U3476)

Целью работы является построение и интерпретация модели линейной регрессии для анализа факторов, влияющих на уровень стресса студентов. Необходимо воспроизвести функционал Excel LINEST средствами Python (библиотеки scikit-learn, numpy, statsmodels), рассчитать показатели регрессии, исследовать мультиколлинеарность, отобрать значимые признаки и проверить модель на выполнение условий теоремы Гаусса–Маркова.

Информация о выбранном датасете

Название: Student Stress Monitoring Dataset или «Набор данных по мониторингу студенческого стресса»

Источник: Kaggle

Датасет содержит результаты анкетирования 1100 студентов в возрасте от 18 до 21 года. Данные собирались через онлайн-опрос (Google Forms) и включают 21 признак, сгруппированный по пяти категориям: психологические, физиологические, академические, социальные и средовые факторы.

Пропуски и дубликаты: отсутствуют

Тип данных: числовые шкалы от 0 до 5 (порядковая шкала, интерпретированная как количественная)

Зависимой переменной (целевой) выступает уровень стресса -stress_level (от 0 до 2, где 0 – стресс отсутствует, 1 – эустресс - «положительный» стресс, который мобилизует организм для решения задач, при этом не причиняя вреда, 2 – дистресс - негативная форма стресса, которая возникает при длительном воздействии неблагоприятных факторов и приводит к истощению психики и организма).

Первоначальные *факторы* (20 признаков):

Психологические: anxiety_level (уровень тревоги), self_esteem (самооценка) , mental_health_history (история психического здоровья), depression (депрессия)

Физиологические: headache (головная боль), blood_pressure (артериальное давление), sleep_quality (качество сна), breathing_problem (проблемы с дыханием)

Социальные: social_support (поддержка социума), peer_pressure (давление со стороны сверстников), extracurricular_activities (внеклассные занятия), bullying (буллинг)

Академические: academic_performance (академическое представление), study_load (учебная нагрузка), teacher_student_relationship (взаимоотношения ученика и учителя), future_career_concerns (обеспокоенность будущей карьерой)

Средовые: noise_level (уровень шума), living_conditions (условия жизни) , safety (безопасность), basic_needs (базовые потребности).

Все признаки числовые (порядковые шкалы), поэтому дополнительного кодирования не потребовалось. Пропусков не обнаружено. Вектор признаков X сформирован из 20 факторов, целевая переменная y — stress_level.

На первом шаге была построена *полная модель* линейной регрессии, включающая все 20 факторов. Модель показала достаточно высокое качество: коэффициент детерминации R2 составил около 0.796, что означает, что почти 80% вариации уровня стресса объясняется включёнными в модель признаками. Среднеквадратическая ошибка оказалась низкой (примерно 0.14), а показатель системного эффекта факторов (доля вариации зависимой переменной, объяснённая всей системой факторов, с учётом их совместного влияния) – около 79% изменчивости уровня стресса среди студентов объясняется системой факторов, включённых в модель (психологические, социальные, академические, физиологические и средовые). Мера мультиколлинеарности (VIF) в пределах 2–5 (обычно допустимо). Выделяется social_support - при последующей проверке на матрице корреляций и построении модели 2 убирается. Несмотря на хорошие общие показатели, модель была перегружена факторами, часть из которых дублировала влияние друг друга. Это потребовало выполнения корреляционного анализа и отбора наиболее информативных признаков.

Для устранения мультиколлинеарности была построена *матрица парных корреляций* между всеми факторами и зависимой переменной.

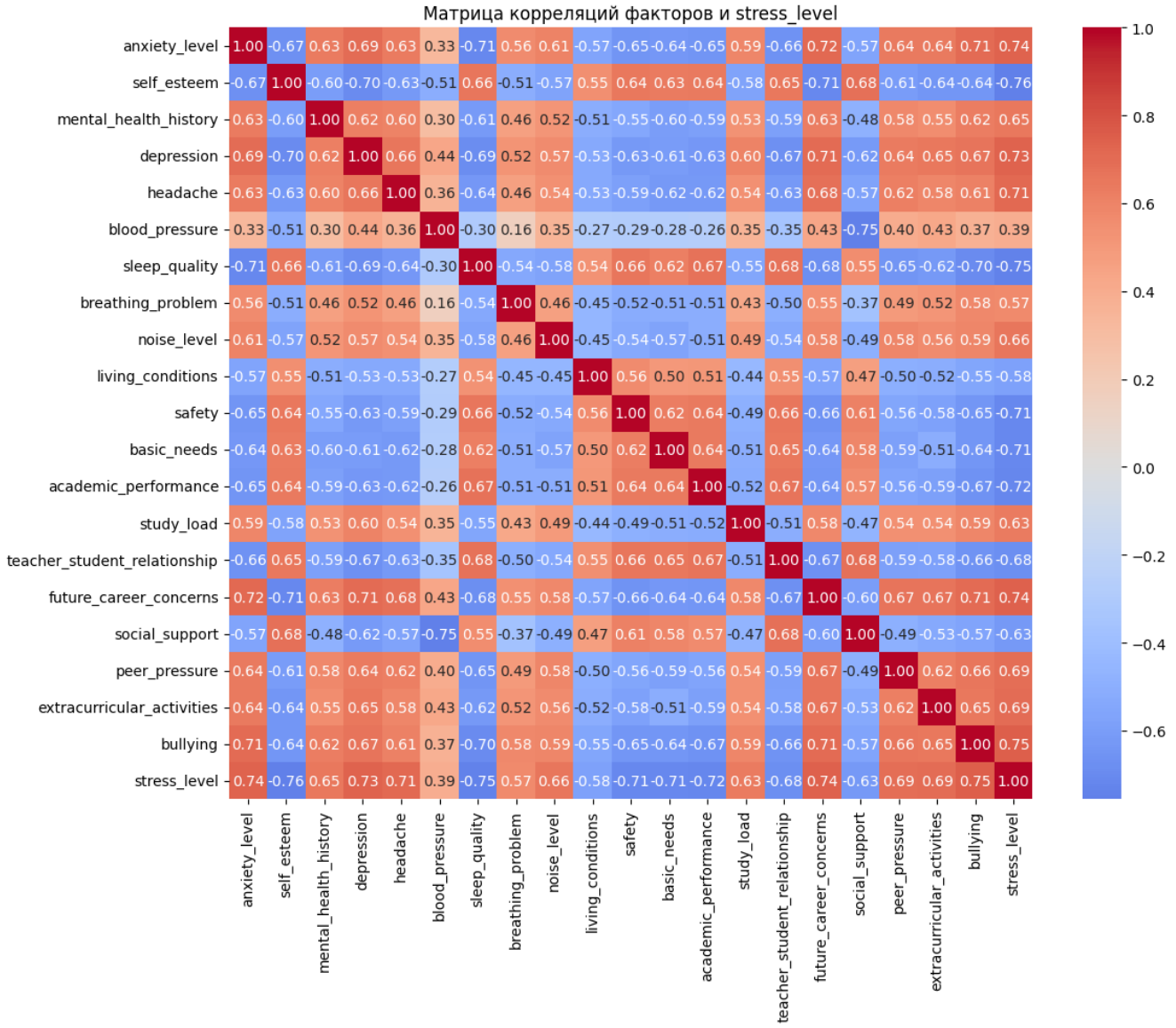

Рисунок 1 - Матрица парных корреляций между всеми факторами и зависимой переменной до исключения некоторых признаков

Высокой корреляцией считались значения коэффициентов выше 0.65 по модулю. В первую очередь по смысловым соображениям из модели были исключены признаки depression и anxiety_level, так как оба показателя отражают схожие психоэмоциональные состояния и, по сути, входят в состав более обобщённого признака mental_health_history, который аккумулирует информацию о ментальном состоянии респондентов. Далее была выявлена высокая корреляция между признаками self_esteem и future_career_concerns. При этом у self_esteem наблюдается более сильная корреляция с целевой переменной (уровнем стресса), что является положительным фактором, в то время как future_career_concerns в большей степени коррелирует с другими признаками, что усиливает эффект мультиколлинеарности. Поэтому предпочтительным оказалось оставить self_esteem, как более информативный

и устойчивый признак. Дополнительно рассматривалась пара self_esteem и social_support. Хотя у social_support корреляция с другими факторами несколько ниже, его связь с целевой переменной слабее. Так как при построении учебной регрессии приоритет отдаётся именно объясняющей способности модели, решено сохранить self_esteem, а не social_support, несмотря на несколько более высокую корреляцию первого с другими признаками. Аналогичным образом из анализа были исключены teacher_student_relationship и sleep_quality. Оба признака продемонстрировали более низкую корреляцию с зависимой переменной и несколько более высокую корреляцию с другими факторами, что также не в пользу их сохранения. Признак blood_pressure был оставлен, так как он обладает умеренной связью с целевой переменной и низкими корреляциями с остальными признаками. Фактор headache также сохранён, поскольку имеет достаточно сильную корреляцию с уровнем стресса и не демонстрирует значимых связей с другими переменными (коэффициенты корреляции не превышают 0.65, если не учитывать уже исключённые признаки). По аналогии с blood_pressure, признаки breathing_problem, noise_level и living_conditions были признаны слабо коррелирующими между собой и сохранены в модели как самостоятельные факторы, отражающие физиологические и средовые аспекты стресса. Пара safety и bullying показала близкие значения корреляции с целевой переменной, но у bullying наблюдалась более сильная связь с другими признаками. Кроме того, safety является более обобщённым показателем, включающим восприятие личной безопасности и комфортности среды, поэтому в модели сохранён именно этот фактор. Оставшиеся признаки - basic_needs, academic_performance, study_load, peer_pressure и extracurricular_activities - продемонстрировали умеренные корреляции с целевой переменной и не имели сильных взаимосвязей между собой (все коэффициенты ниже порогового значения 0.65). Таким образом, они были оставлены в модели без изменений. В результате из исходных 20 признаков в финальную регрессионную модель вошли 13 факторов: mental_health_history, self_esteem, blood_pressure, headache, breathing_problem, noise_level, living_conditions, safety, basic_needs, academic_performance, study_load, peer_pressure, extracurricular_activities. Исключены были 7 признаков: anxiety_level, depression, future_career_concerns, social_support, teacher_student_relationship, sleep_quality, bullying.

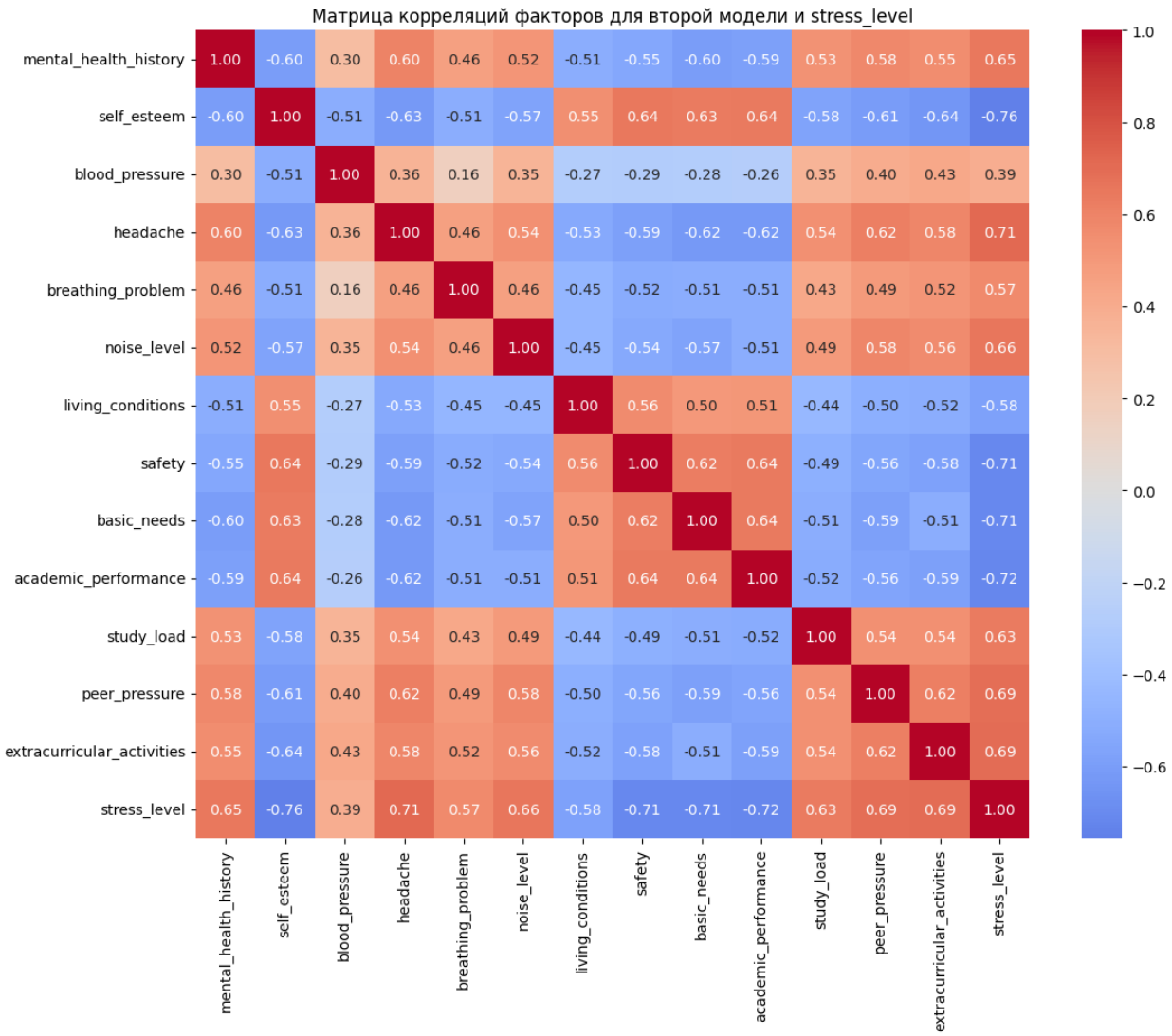

Рисунок 2 - Матрица парных корреляций между факторами и зависимой переменной после исключения некоторых признаков

После исключения коррелирующих переменных *модель была пересчитана заново*. Новое значение коэффициента детерминации составило $R^2 = 0.784$, что лишь незначительно меньше, чем в исходной модели (разница около 1%). Среднеквадратическая ошибка увеличилась несущественно (до 0.145), а скорректированный R^2 остался на уровне 0.782. Главным преимуществом новой модели стало устранение мультиколлинеарности: значения VIF для всех факторов оказались ниже 3, что говорит о независимости признаков. Таким образом, полученная модель стала более стабильной и интерпретируемой при практически неизменном уровне объясняющей способности.

Чтобы убедиться, что удаление признаков не ухудшило модель, было выполнено сравнение двух регрессий с помощью критерия Фишера. Полученное значение статистики F = 9.085. F=9.085 оказалось выше

критического при уровне значимости 0.05, что формально указывает на то, что исключение некоторых факторов ухудшило модель.

Проверка условий теоремы Гаусса–Маркова

Был выполнен критерий поворотных точек (runs test), который подтвердил случайность распределения знаков остатков ($p \approx 0.63$) - p-value> 0.05, значит, нет оснований отвергать гипотезу о случайности остатков. Это означает, что остатки не зависят друг от друга и не имеют систематического характера.

Во-вторых, тест Дарбина–Уотсона показал значение около 2.0, что соответствует отсутствию автокорреляции между остатками.

Дополнительно были рассчитаны коэффициенты асимметрии и эксцесса: skewness ≈ -0.38 и kurtosis ≈ 8.6. Проведен тест на нормальность распределения остатков (тест Жарка–Бера): p-value = 0, значит, отклоняем H_0 о нормальности. Отрицательная асимметрия и большой эксцесс - распределение “слева тяжелое”, с пиками и хвостами. Это не страшно, особенно при большом кол-ве наблюдений: просто тесты могут быть чуть менее точными.

Среднее значение ошибок оказалось близким к нулю, а t-тест показал p-value значительно больше 0.05, что показывает: гипотеза «среднее = 0» не отвергается.

В ходе работы была реализована процедура, аналогичная Excel LINEST, с использованием Python. На основе данных о студенческом стрессе построена регрессионная модель, выявляющая факторы, оказывающие наибольшее влияние на уровень стресса. Первая модель показала высокое качество, но содержала признаки мультиколлинеарности. После отбора факторов удалось получить более устойчивую и практически эквивалентную по точности модель. Проверка условий теоремы Гаусса–Маркова подтвердила корректность построенной модели. В результате получена адекватная и интерпретируемая модель, пригодная для анализа и прогнозирования факторов, определяющих стресс у студентов.