Задание 1. Реализация линейной множественной регрессии

Карцев Михаил Дмитриевич U3475 АДПУР 7.2

Цель работы: сформировать и оценить линейную множественную регрессионную модель для предсказания субъективного качества сна на основе отобранных количественных факторов, выполнить отбор значимых признаков, сравнить полную и сокращенную модели по коэффициенту детерминации, MSE и системному эффекту, проверить целесообразность исключения факторов по критерию Фишера и оценить выполнение условий Гаусса–Маркова для корректности оценок.

Постановка задачи: с помощью средств MS Excel построение линейной многофакторной модели с отбором значимых факторов и выполнить оценку соответствия модели условия Гаусса-Маркова.

Исходный датасет:

Был взят датасет «Screen Time vs Mental Wellness Survey - 2025» с сайта Kaggle. Этот набор данных содержит информацию, полученную от 400 участников опроса, о том, как их ежедневное использование экранов влияет на психическое благополучие. В связи с растущей распространенностью цифровых устройств в нашей жизни понимание связи между временем, проведенным за экраном, качеством сна, стрессом и продуктивностью становится важнейшей областью исследований в области науки о данных, психологии и общественного здравоохранения. Пустые значения отсутствуют.

В качестве факторов (x1, x2, x3, x4) были взяты:

– X1 - возраст

– X2 - экранное время, потраченное на работу/учебу в день в среднем

– X3 - экранное время, потраченное на развлечение (видео, игры, социальные сети и т.д.) в день в среднем

– X4 - среднее время сна за ночь

В качестве зависимого параметра (y) была взята субъективная оценка качества сна респондентов (где 1 – очень плохо, 5 – очень хорошо)

1. Первая линейная регрессия (по всем факторам)

	m4	m3	m2	m1	b
Коэфф	0,451	-0,025	-0,015	-0,004	-1,442
SE	0,032	0,013	0,014	0,003	0,287
R^2; SE_y	0,386	0,514	#Н/Д	#Н/Д	#Н/Д
F; df	62,173	395,000	#Н/Д	#Н/Д	#Н/Д
SSreg; Ssresid	65,602	104,196	#Н/Д	#Н/Д	#Н/Д

Коэффициент детерминации	**0,386352712**
Средняя квадратическая ошибка	**0,263786773**
Показатель системного эффекта факторов	**-86,16091705**
Мера мультиколлинеарности	**-0,052011458**

t-статистика	14,03976879	1,979218899	1,02796526	1,133428186
	>tкр	>tкр	<tкр	<tкр
p-value	3,72451E-96	8,94607E-05	0,0404459	0,023938972
	<α	<α	>α	>α
tкр	1,967			
α	0,05			

Полная модель использует четыре фактора: возраст, экранное время для работы, экранное время для развлечений и часы сна, при этом главный вклад дает продолжительность сна (положительный коэффициент около 0,451), а развлекательное экранное время связано с меньшей оценкой сна (отрицательный коэффициент около −0,025), тогда как возраст и рабочее экранное время имеют намного более слабые эффекты (около −0,004 и −0,015 соответственно). Метрики качества у полной модели умеренные: $R^2 \approx 0,386$ и MSE ≈ 0,264, то есть модель объясняет заметную, но не основную часть разброса субъективной оценки качества сна по шкале 1–5. В отчете также зафиксированы «системный эффект факторов» около −86,16 и «мера мультиколлинеарности» около −0,052, что означает перекрытие парных связей факторов с целевой переменной и отражает то, что часть «вклада» факторов в оценку сна у данных пересекается между собой.

2. Матрица корреляций

	x1	*x2*	***x3***	***x4***	*y*
x1	1				
x2	0,07444263	1			
x3	0,01303851	-0,2864136	1		
x4	0,05350326	-0,1332431	-0,2557467	1	
y	-0,0178161	-0,1008538	-0,2245234	0,61438142	1

Проверка значимости коэффициентов корреляции с y					
Фактор	r(y, x)	r^2	t-статистика	p-value	Значимость
x1	-0,0178161	0,00031741	-0,3554874	0,47751634	Не значим
x2	-0,1008538	0,01017149	-2,0223383	6,294E-05	Слабо значим
x3	-0,2245234	0,05041074	-4,5965837	2,154E-18	Значим
x4	0,61438142	0,37746453	15,5345099	1,863E-108	Высоко значим

Корреляции с целевой переменной показывают простую картину: самая сильная связь у часов сна (y и x4 – 0,61), связь с развлечениями умеренно отрицательная (y и x3 – −0,22), тогда как возраст и рабочее экранное время связаны с оценкой сна слабо по модулю. Между самими факторами связи невысокие, что говорит о том, что факторы в целом не «дублируют» друг друга напрямую, а конкуренция за объяснение оценки сна возникает из- за пересечения их индивидуальных связей с целевой переменной. Такая структура логично

согласуется со знаками и относительной величиной коэффициентов в полной модели: больше сна – выше оценка, больше развлечений – ниже оценка, а возраст и рабочие экраны почти не меняют картину в линейной постановке. Были отобраны факторы x3 и x4, поскольку их влияние на зависимый параметр гораздо выше, при этом друг с другом все факторы коррелируют слабо, что позволяет отбирать и рассматривать любые группы. Проверка значимости коэффициентов корреляции с y (n = 400, df = 398, tкр = 1.967 при α = 0.05) показала, что фактор x4 имеет высокую значимость, x3 – просто значим, x2 – слабо значим а x1– не является значимым. Аналогичные результаты можно видеть в проверке значимости факторов в первой и второй линейных регрессиях.

3. Вторая линейная регрессия (с факторами x3 и x4)

	m2	m1	b
Коэфф	0,456	-0,021	-1,656
SE	0,031	0,012	0,254
R^2; SE_y	0,382	0,514	#Н/Д
F; df	122,866	397,000	#Н/Д
SSreg; Ssresid	64,918	104,880	#Н/Д

Коэффициент детерминации	**0,382324826**
Средняя квадратическая ошибка	**0,264180606**
Показатель системного эффекта факторов	**-91,732**
Мера мультиколлинеарности	**-0,046**

t-статистика	14,6058899	1,767447887
	>tкр	>tкр
p-value	6,1006E-101	0,00045602
	<α	<α
tкр	1,967	
α	0,05	

Сокращенная модель оставляет два наиболее информативных фактора - часы сна и развлекательное экранное время - при этом коэффициенты по смыслу остаются такими же: положительный при часах сна (около 0,456) и отрицательный при развлечениях (около −0,021). По метрикам качество практически не меняется: $R^2 \approx 0{,}382$ и MSE ≈ 0,264, то есть по точности описания оценок сна сокращенная модель близка к полной. В отчете «системный эффект факторов» для сокращенной модели около −91,732, а «мера мультиколлинеарности» около −0,046, что указывает на сохраняющееся перекрытие парных связей с целевой переменной, но уже без слабых факторов, практически не влияющих на итоговые метрик.

Критерий Фишера:

D4	0,38635271
D2	0,38232483
F1	**1,296**
f1	2

f2	395
tкр	**3,04**
Уровень значимости	0,95

F1 < t, следовательно гипотеза о том, что исключенные факторы не влияют на y не опровергается, а значит исключение факторов x1 и x2 является целесообразным.

Сравнение моделей

Различия метрик между полной и сокращенной спецификациями малы: R^2 снизился с ≈0,386 до ≈0,382, а MSE изменился с ≈0,2638 до ≈0,2642, то есть качество описания практически сохранилось. По расчету критерия Фишера получено, что в интерпретации отчета подтверждает целесообразность исключения x1 и x2.

4. Условия Гаусса-Маркова

Случайность остатков:

K	261
$\frac{2n-1}{3} - 2\sqrt{\frac{16n-29}{90}}$	249,5061067

Независимость остатков

d	**1,59**
dl	1,728
dh	1,81

Критерий Стьюдента:

p	0,95
f	395
t	**7,2695**
tкр	**1,967**

Случайность остатков по критерию поворотных точек подтверждена, что согласуется с предпосылкой случайного характера последовательности ошибок. Критерий Дарбина–Уотсона дал d ≈ 1,59 < dl, что указывает на положительную автокорреляцию остатков.

Коэффициенты асимметрии и эксцесса

Коэффициент асимметрии	**3,571303517**
Коэффициент эксцесса	**13,72955822**
Sa (стандартное отклонение коэфф)	0,014776953
Se (стандартное отклонение коэфф)	0,057795762

Проверка равенства суммы остатков нулю по критерию Стьюдента показала t ≈ 7,27 > tкр, что отвергает гипотезу о нулевой сумме ошибок. Коэффициенты асимметрии ≈3,57 и эксцесса ≈13,73 с учетом стандартных ошибок указывают на отклонение от нормальности

распределения остатков

Почему такие значения

Зависимая переменная — субъективная порядковая оценка качества сна по шкале 1–5, что ограничивает достижимую долю объясненной дисперсии линейной моделью и согласуется с умеренными значениями коэффициента детерминации в обеих спецификациях. Основной вклад объяснения связан с продолжительностью сна и противоположным по знаку влиянием развлекательного экранного времени, тогда как возраст и рабочее экранное время показывают существенно меньшую линейную связь с субъективной оценкой.

Вывод

Сокращенная линейная модель с факторами x3 и x4 сохраняет метрики качества на уровне полной модели и, по расчету критерия для сравнения моделей, подтверждает целесообразность исключения слабых факторов x1 и x2. Диагностика остатков указывает на положительную автокорреляцию и ненормальность, что отражает особенности данных с субъективной шкалой.