2. Строим регрессию (80 баллов)

Юный аналитик решил провести исследование об успеваемости своих одноклассников. Он предположил, что средний балл (переменная *grade*) зависит от пола ученика (переменная *is_male*, принимающая значение 1, если ученик мужского пола, и 0, если женского).

(i). В результате была получена следующая модель линейной регрессии:

$$grade = 4.16026667 - 0.18740000 * is\_male$$

Заполните приведенную ниже таблицу, не забудьте объяснить, почему ваши расчеты верны. Если для заполнения некоторых ячеек вам необходима дополнительная информация, укажите это.

	Средний балл
Мальчики	?
Девочки	?
Все ученики	?

Считайте, что была использована классическая линейная регрессия — то есть минимизирующая сумму квадратов ошибок.

Решение:

Регрессия будет предсказывать 4.16026667 для всех девочек и 3.97286667 для всех мальчиков. Эти значения будут средними для соответствующих групп, так как сумма квадратов ошибок будет минимизирована при предсказании среднего. Для среднего балла всех учеников необходимо знать соотношение мальчиков и девочек

Критерии:

Max – **20 баллов**

Ответ – **8 баллов** если значения для девочек и для мальчиков верны. **0 баллов** если хотя бы один ответ неверен

Объяснение – **8 баллов**

Невозможность предсказания среднего – **4 балла**

(ii). Одноклассницы юного аналитика возмутились по поводу такой системы кодировки пола. Тогда он решил заменить переменную is_male на переменную is_female, принимающую значение 0, если студент мужского пола, и 1, если женского.

Какими будут коэффициенты новой регрессии и почему?

Решение:

Регрессия снова должна предсказывать 4.16026667 для всех девочек и 3.97286667 для всех мальчиков, чтобы минимизировать сумму квадратов. Это будет достигнуто, если $grade = 3.97286667 + 0.18740000 * is\_female$

Критерии:

Max – **20 баллов**

Коэффициент при переменной будет другой по знаку, но такой же по модулю – **4 балла**

Свободный коэффициент – **8 баллов**

Объяснение обоих коэффициентов – **8 баллов**

(iii). Юный аналитик решил добавить в регрессию обе переменные. Будет ли полученная регрессия лучше объяснять оценки учеников и почему?

Решение:

Нет, не будет, так как никакой дополнительной информации новая переменная не даст, потому что одна переменная линейно зависит от другой.

Критерии:

Max – **8 баллов**

Объяснение + ответ – **8 баллов**

Ответ без объяснения – **1 балл**

(iv). Предположим, что аналитик вместо добавления переменной is_female решил удвоить изначальный набор наблюдений, просто дописав к каждому наблюдению ровно такое же. Как изменятся коэффициенты регрессии и ее качество? Обоснуйте свой ответ.

Решение:

Разделим выборку на две изначальные и построим две регрессии, которые будут одинаковыми. Мы знаем, что каждая регрессия будет минимизировать сумму квадратов ошибок в своей группе. Значит, общая сумма квадратов ошибок также будет минимизирована, если взять регрессию с теми же коэффициентами. При этом качество регрессии не изменится, так как доля объясненной дисперсии в итоговой выборке будет такая же, как в общей.

Критерии:

Max – **32 балла**

Объяснение + ответ для коэффициентов – **20 баллов** (**8 баллов**, если объяснение на пальцах - даны утверждения, из которых можно сделать объяснения)

Объяснение + ответ для качества регрессии – **12 баллов**

Только ответ для коэффициентов – **4 балла**

Только ответ для качества – **4 балла**