2 ЛАБОРАТОРНАЯ РАБОТА №2. ЛИНЕЙНАЯ РЕГРЕССИЯ

ЦЕЛЬ РАБОТЫ – Изучение основ регрессионного анализа и реализация модели линейной регрессии на языке Python.

2.1 Теоретические сведения

2.1.1 Регрессионный анализ

В регрессионном анализе рассматривается связь между одной переменной, называемой *зависимой переменной*, и несколькими другими, называемыми *независимыми переменными*. Эта связь представляется с помощью *математической модели*, т.е. уравнения, которое связывает зависимую переменную с независимыми с учетом множества соответствующих предположений. Независимые переменные связаны с зависимой посредством *функции регрессии*, которая в свою очередь имеет набор неизвестных *параметров*. Если функция линейна относительно параметров (но необязательно линейна относительно независимых переменных), то говорят о ***линейной модели регрессии***. В противном случае модель называется *нелинейной*.

При построении модели линейной регрессии мы предполагаем, что между зависимой случайной величиной Y и независимой случайной величиной X существует связь, которую можно записать в общей форме:

$$Y = f(X) + e,$$

где f – фиксированная, но неизвестная функция от X, e – ошибка, которая не зависит от X.

Можно сказать, что функция f выражает ***систематическую*** информацию о Y, содержащуюся в X.

Регрессионный анализ используют по двум причинам. Во-первых, *описание* зависимости между переменными помогает установить наличие возможной причинной связи. Во-вторых, для получения *предиктора (предсказателя)* для зависимой переменной. Эта возможность особенно важна в тех случаях, когда прямые измерения зависимой переменной затруднены или дорого стоят.

Величина линейной зависимости между двумя переменными измеряется посредством коэффициента корреляции. Ниже будет показано, что методы регрессионного и корреляционного анализов тесно связаны между собой.

Приведем пример задачи, где возможно применение линейной регрессии. Предположим, у нас есть данные о игроках баскетбольной команды, которые включают в себя такие параметры, как рост, вес, возраст и средненедельный объем спонсорской поддержки. (табл. 2.1).

Таблица 2.1 – Информация о баскетбольной команде

Номер игрока	Возраст, лет	Рост, см	Вес, фунты	Спонсорская помощь

1	29	192	218	561
2	35	218	251	60
3	22	197	221	1312
4	22	192	219	1359
5	29	198	223	362
6	21	166	188	1536
7	25	195	221	694
8	21	182	199	1678
9	27	189	199	385
10	24	205	232	1416
11	29	206	246	314
12	23	185	207	1497
13	24	172	183	1383
14	24	169	183	1034
15	29	185	197	178
16	30	215	232	434
17	29	158	184	162
18	27	190	207	648
19	28	195	235	481
20	32	192	200	427
21	31	202	220	542
22	32	184	213	12
23	22	190	215	1179
24	21	178	193	1078
25	31	185	200	213
26	19	191	218	1855
27	32	196	235	47
28	22	198	221	1409
29	27	207	247	1065
30	25	201	244	1111

На основании приведенных данных используя метод линейной регрессии можно, например, построить функцию, которая предсказывает величину спон-

сорской помощи в зависимости от возраста игрока. На рис. 2.1 приведена диаграмма рассеяния для переменных возраст (X) и спонсорская помощь (Y). Из диаграммы видно, что с увеличением возраст в среднем спонсорская помощь падает. Метод линейной регрессии позволяет выразить данную взаимосвязь более строгим образом.

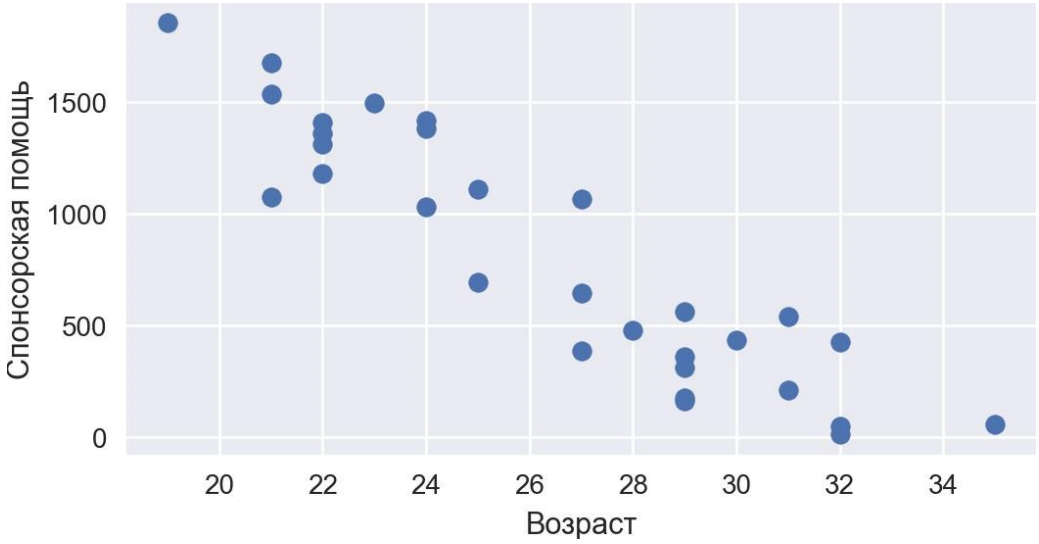

Рис. 2.1. – Диаграммы рассеяния

2.1.2 Модель линейной регрессии

Если предполагается линейная зависимость между случайными величинами Y и X, то теоретическая модель задается уравнением

$$y_i = w_0 + w_1 x_i + e_i, \;\; i = 1, \dots, N, \tag{2.1}$$

и называется моделью ***простой линейной регрессии*** Y по X. Величины w_0 и w_1 являются неизвестными параметрами, а $e_1, \dots, e_N$ суть некоррелированные ошибки, которые можно считать выборкой случайной величины e со средним 0 и неизвестной дисперсией:

$$E\{e\} = 0, \quad \mathrm{Var}\{e\} = \sigma^2. \tag{2.2}$$

Т.е. для каждого значения $X = x_i$ имеется распределение Y (не обязательно нормальное) со средним значением $w_0 + w_1 x_i$ и дисперсией σ^2.

Таким образом, у линейной регрессии есть три параметра: w_0 – ***пересечение*** (свободный член), w_1 – ***угловой коэффициент*** (коэффициент регрессии), σ^2 – ***дисперсия остатков.***

Таким образом, при построении модели линейной регрессии ставится задача нахождения *оценок* параметров $\widehat{w}_0, \widehat{w}_1, \widehat{\sigma}^2$. В результате будет получена подогнанная линейная функция

$$\hat{f}(x) = \hat{w}_0 + \hat{w}_1 x,$$

которая позволит получить *оценку* зависимой переменой y_i на основании значения независимой переменной x_i

$$\hat{y}_i = \hat{f}(x_i).$$

Зная оценки y_i можно вычислить остатки

$$e_i = y_i - \hat{y}_i = y_i - (\hat{w}_0 + \hat{w}_1 x_i).$$

При получении (обучении) линейной регрессии важную роль играет параметр суммы квадратов отклонений или остатков (*RSS – residual sums of squares*):

$$RSS = \sum_{i=1}^{N} {e_i}^2.$$

Каким образом можно получить оценку неизвестных значений w_0 и w_1 на основе имеющейся выборки данных объема N? Наилучшие оценки $\hat{w}_0$ и $\hat{w}_1$ для w_0 и w_1 получаются минимизацией соответственно по $\hat{w}_0$ и $\hat{w}_1$ суммы квадратов остатков

$$RSS = \sum_{i=1}^{N} (y_i - (\hat{w}_0 + \hat{w}_1 x_i))^2. \qquad (2.3)$$

Для примера покажем, как выглядит график ошибки RSS модели линейной регрессии, построенной для данных, показанных на рис. 2.1. Из (2.3) видно, что RSS зависит от имеющейся выборки данных $(x_i, y_i)_{i=1,..,N}$, а параметрами функции являются коэффициенты регрессии $\hat{w}_0$ и $\hat{w}_1$. График на рис. 2.1 показывает нам, что есть определенные значения параметров $\hat{w}_0$ и $\hat{w}_1$ при которых функция RSS достигает минимального значения.

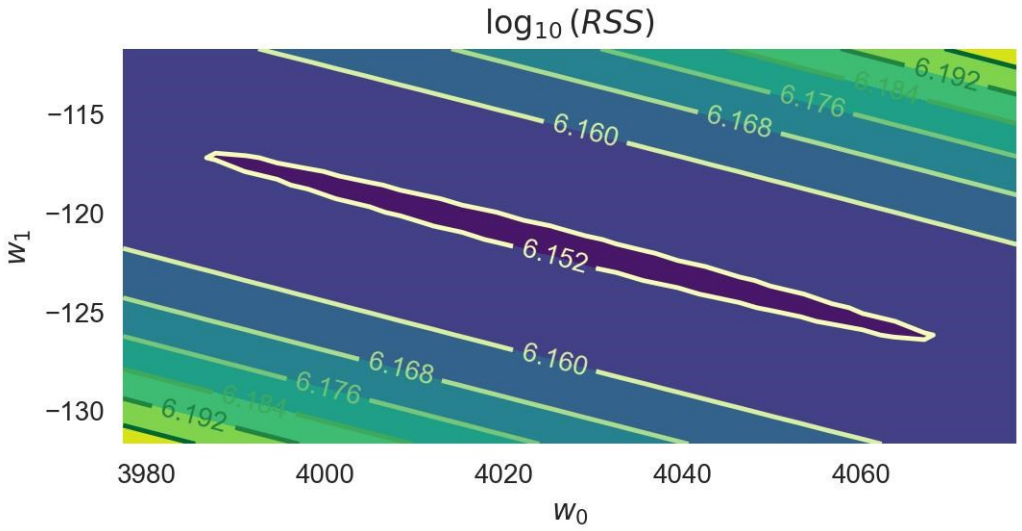

Рис. 2.2. – Функция ошибки RSS в логарифмическом масштабе

Поиск коэффициентов линейной регрессии

Очевидно, что RSS есть мера ошибки, возникающей при аппроксимации данных (x_i, y_i) при помощи регрессионной модели $\hat{y}_i = \hat{w}_0 + \hat{w}_1 x_i$. Естественно, что нас интересует такие значения $\hat{w}_0$ и $\hat{w}_1$ минимизируют ошибку, которые минимизируют RSS.

Для минимизации RSS найдем частные производные по $\hat{w}_0$ и $\hat{w}_1$ и приравняем их к нулю.

$$\frac{\partial RSS}{\partial \hat{w}_0} = \sum_{i=1}^{N} 2(y_i - \hat{w}_0 - \hat{w}_1 x_i)(-1) \triangleq 0$$

$$\Rightarrow \hat{w}_0 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{w}_1 x_i)$$

Т.е.

$$\hat{w}_0 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{w}_1 x_i) = \underbrace{\frac{1}{N}\sum_{i=1}^{N} y_i}_{\bar{y}} - \hat{w}_1 \underbrace{\frac{1}{N}\sum_{i=1}^{N} x_i}_{\bar{x}} = \bar{y} - \hat{w}_1 \bar{x}. \tag{2.4}$$

В выражении (2.4) через $\bar{y}$ и $\bar{x}$ обозначены средние значения соответствующих случайных величин.

Выполним подстановку (2.4) в (2.3) и получим

$$RSS = \sum_{i=1}^{N}(y_i - (\hat{w}_0 + \hat{w}_1 x_i))^2 = \sum_{i=1}^{N}(y_i - (\bar{y} - \hat{w}_1 \bar{x}) - \hat{w}_1 x_i)^2$$

$$= \sum_{i=1}^{N}(y_i - \bar{y} - \hat{w}_1 (x_i - \bar{x}))^2. \tag{2.5}$$

Используя последнее выражение, найдем производную RSS по $\widehat{w}_1$

$$\frac{\partial RSS}{\partial \widehat{w}_1} = \sum_{i=1}^{N} 2\big(y_i - \bar{y} - \widehat{w}_1(x_i - \bar{x})\big)(-1)(x_i - \bar{x}) \triangleq 0. \tag{2.6}$$

Далее выразим $\widehat{w}_1$:

$$\widehat{w}_1 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{2.7}$$

Нахождение коэффициентов по формулам (2.4) и (2.7) называется подгонкой (fit) модели линейной регрессии по ***методу наименьших квадратов***. Название обусловлено тем, что найденные коэффициенты минимизируют функцию квадратичной ошибки RSS (см. формулу (2.3)).

Анализ модели линейной регрессии

Оценкой уравнения регрессии (или прямой наименьших квадратов) будет

$$\hat{y} = \widehat{w}_0 + \widehat{w}_1 x.$$

Оценкой значения с.в. Y при $X = x_i$ есть $\hat{y}_i = \widehat{w}_0 + \widehat{w}_1 x_i$. Разница между наблюдаемым и оцененным значением Y называется ***отклонением*** или ***остатком***

$$e_i = y_i - \hat{y}_i.$$

Прямая наименьших квадратов доставляет минимум сумме квадратов отклонений $\widehat{RSS} = \sum_{i=1}^{N} e_i^2$.

Используя (2.4) и (2.7) уравнение линейной регрессии можно записать в следующем виде

$$\hat{y} = \underbrace{\bar{y} - \widehat{w}_1 \bar{x}}_{\widehat{w}_0} + \widehat{w}_1 x = \bar{y} + \widehat{w}_1(x - \bar{x}). \tag{2.8}$$

Коэффициент $\widehat{w}_1$ обычно называется ***коэффициентом регрессии***, $\widehat{w}_0$ – ***свободным членом уравнения регрессии***.

Следует заметить, что коэффициент регрессии содержит умноженную на N ковариацию между X и Y и умноженную на N дисперсию X в знаменателе

$$\widehat{w}_1 = \frac{N \cdot \mathrm{Cov}(X, Y)}{N \cdot \mathrm{Var}\{X\}} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}\{X\}}.$$

Отметим ещё несколько полезных вещей. Согласно выражению (2.8) прямая регрессии проходит через точку $(\bar{x}, \bar{y})$. Прибавление константы ко всем значениям x (параллельный перенос) влияет только на свободный член, но не на коэффициент регрессии, поскольку тот определен в терминах отклонения от среднего значения и при параллельном переносе не изменяется. Поэтому мы можем ***центрировать значения*** x ***относительно нуля***, вычтя из каждого x_i величину $\bar{x}$, тогда свободный член будет равен $\bar{y}$. Можно даже вычесть $\bar{y}$ из всех значений y_i, чтобы получить нулевой свободный член, не изменив задачу существенным образом.

Уравнение линейной регрессии можно существенно упросить, если ***нормировать*** X, поделив все значения x_i на дисперсию $\sigma_X^2 = \mathrm{Var}\{X\}$:

$$x_i{}' = \frac{x_i}{\sigma_X^2}. \tag{2.9}$$

В этом случае $\bar{x}$ заменится на $\bar{x}' = \bar{x}/\sigma_X^2$. Для нормированных значений коэффициент регрессии имеет вид

$$\widehat{w}_1 = \frac{1}{N}\sum_{i=1}^{N}(x_i{}' - \bar{x}')(y_i - \bar{y}) = \mathrm{Cov}\{X', Y\}, \tag{2.10}$$

т.е. в качестве коэффициента регрессии можно взять ковариацию между нормированным признаком и целевой переменной. Таким образом, можно считать, что линейная регрессия состоит из двух шагов:

1. Нормировка признака (независимой переменной), см. выражение (2.9).
2. Вычисление ковариации между целевой переменной и нормированным признаком (2.10).

Важно отметить ещё один факт: сумма отклонений ε_i, полученных методом наименьших квадратов, равна нулю:

$$\sum_{i=1}^{N}(y_i - (\widehat{w}_0 + \widehat{w}_1 x_i)) = N(\bar{y} - \widehat{w}_0 - \widehat{w}_1\,\bar{x}) = 0. \tag{2.11}$$

Это свойство интуитивно может казаться красивым, но следует помнить, что оно делает линейную регрессию чувствительной к ***выбросам***. Под выбросами понимают точки, отстоящие далеко от прямой регрессии, что нередко случается в результате ошибок измерения.

2.1.3 Линейная регрессия (матричная форма записи)

Описание модели

Простая модель линейной регрессии, рассмотренная выше, учитывает только один предиктор (описательный признак). Однако, большинство задач

аналитического прогнозирования носят многомерный характер. Распространим модель линейной регрессии на случай многомерных данных. Такая модель будет называться ***множественной линейной регрессией***.

Для начала опишем одномерную линейную регрессию в матричной форме:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} w_0 + \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} w_1 + \begin{pmatrix} e_1 \\ \vdots \\ e_N \end{pmatrix}. \tag{2.12}$$

Теперь мы можем перейти к однородным координатам и переписать это выражение следующим образом:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ \vdots \\ e_N \end{pmatrix}. \tag{2.13}$$

Теперь мы можем переписать (2.13) в матричном виде

$$\mathrm{y} = \mathbf{X}\mathrm{w} + \mathbf{e}, \tag{2.14}$$

где X – матрица размерности $N \times 2$, а $\mathbf{y}, \mathbf{e}$ – векторы размерности $N \times 1$, $\mathbf{w} = (w_0 \quad w_1)^T$.

В случае, если предикторов становится больше, например d, то $\mathbf{X}$ просто становится матрицей $N \times (d+1)$, первый столбец которой содержит единицы, а остальные столбцы хранят d предикторов; первым элементом $\mathbf{w}$ является свободный член, а остальные d – коэффициенты регрессии.

Легко заметить, что в формулировке (2.14) отдельное предсказание величины $\hat{y}_i$ осуществляется, путем взятия i-го строки матрицы X, которую можно обозначить, как $\mathbf{X}_{i:}$, и вычисления её *скалярного произведения* с вектором коэффициентов регрессии $\mathbf{w}$:

$$\hat{y}_i = \mathbf{X}_{i:}\mathbf{w} = \langle \mathbf{X}_{:i}, \mathbf{w} \rangle \tag{2.15}$$

Как и в случае простой линейной регрессии определим **функцию ошибки** как сумму квадратов остатков:

$$RSS = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} (y_i - \langle \mathbf{X}_{:i}, \mathbf{w} \rangle)^2 \tag{2.16}$$

Нормальные уравнения

Оценка $\widehat{\mathbf{w}}$ по методу наименьших квадратов (МНК) минимизирует:

$$RSS(\widehat{\boldsymbol{w}}) = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \|\mathbf{y} - \mathbf{X}\widehat{w}\|^2 \tag{2.17}$$

МНК оценка вектора $\widehat{\mathbf{w}}$ вычисляется по формуле:

$$\widehat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{2.18}$$

Давайте это докажем. Начнем с того, что перепишем (2.17) в виде

$$RSS(\widehat{\mathbf{w}}) = (\mathbf{y} - \mathbf{X}\widehat{\mathbf{w}})^T(\mathbf{y} - \mathbf{X}\widehat{\mathbf{w}}).$$

Это квадратичная функция с $d + 1$ параметрами. Продифференцируем её по $\widehat{\mathbf{w}}$:

$$\frac{\partial RSS}{\partial \widehat{\mathbf{w}}} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\widehat{\mathbf{w}}), \quad \frac{\partial^2 RSS}{\partial \widehat{\mathbf{w}} \partial \widehat{\mathbf{w}}^T} = 2\mathbf{X}^T\mathbf{X}$$

Приравняем первую производную к нулю и получим:

$$-2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\widehat{\mathbf{w}}) = 0 \quad \Rightarrow \quad \widehat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathrm{X}^T\mathbf{y}.$$

Выражение (2.18) описывают систему нормальных уравнений.

2.1.4 Градиентный спуск

Для минимизации RSS в выражении (2.16) можно использовать метод градиентного спуска. Напомним, что градиентный спуск – это итеративный метод поиска минимума (максимума) функции многих переменных.

При инициализации в качестве w задается случайная стартовая точка в весовом пространстве.

На каждой итерации происходит обновление весов в соответствии со следующей схемой

$$w_p \leftarrow w_p - \eta \frac{\partial RSS}{\partial w_p}, \tag{2.19}$$

где η – скорость обучения, которая определяет как быстро алгоритм сходится; $p = 1,2, \dots d$.

Чтобы воспользоваться (2.19) требуется вычислить частную производную $\frac{\partial RSS}{\partial w_p}$. Сделаем это.

$$\frac{\partial RSS}{\partial w_p} = \sum_{i=1}^{N} 2 \cdot (y_i - \langle \mathbf{X}_{:i}, \mathbf{w} \rangle) \cdot \frac{\partial (y_i - \langle \mathbf{X}_{:i}, \mathbf{w} \rangle)}{\partial w_p}$$
$$= \sum_{i=1}^{N} 2 \cdot (y_i - \langle \mathbf{X}_{:i}, \mathbf{w} \rangle) \cdot \left(-\mathrm{X}_{p,i}\right). \tag{2.20}$$

Критерием остановки градиентного спуска может служить значение разницы минимизируемой функции на предыдущем и на текущим шаге. Если она меньше наперед заданного значения ϵ, которое обычно выбирается в диапазоне от 10^{-3} до 10^{-6}, то процесс минимизации можно завершать.

Типичный диапазон для скорости обучения $[10^{-6},\ 0{,}1]$.

2.1.5 Оценка качества линейной регрессии

Значение RSS

Значение RSS сообщает вам, какая величина отклонения (дисперсии) остается после подгонки линейной модели, которая измеряется квадратами различий между прогнозируемыми и фактическими целевыми значениями. RSS – это общая дисперсия результата.

Коэффициент детерминации R^2

Мы можем использовать коэффициент детерминации R^2 для оценки эффективности модели линейной регрессии. В основе расчета коэффициента R^2 лежит сравнение оцениваемой модели с другой, «базовой» моделью. В качестве базовой модели используется воображаемая модель, которая всегда возвращает среднее значение зависимой переменной y, которое рассчитано на тестовом наборе (либо на всем наборе данных). Таким образом, для любого x_i базовая модель всегда будет возвращать $\bar{y}$. Итак, коэффициент детерминации вычисляется как

$$R^2 = 1 - \frac{\text{сумма квадратов остатков}}{\text{общая сумма квадратов}} = 1 - \frac{RSS}{TSS}, \tag{2.21}$$

где ***общая сумма квадратов*** (total sum of squares) вычисляется как:

$$TSS = \sum_{i=1}^{N} (y_i - \bar{y})^2, \tag{2.22}$$

Значение R^2 находится в диапазоне $[0, 1)$. Значения, близкие к единице указывают на лучшую эффективность модели. Коэффициент детерминации R^2 также интерпретируют, как величину изменения зависимой переменной y, которая обясняется независимыми (описательными) признаками x.

Заметим также, что коэффициент детерминации R^2 для случая простой линейной регрессии Y по X в точности совпадает квадратом коэффициента корреляции r_{XY}^2.

R^2 показывает, какая часть дисперсии целевой переменной Y может быть объяснена линейной моделью. Значение R^2 колеблется между 0 для моделей, в которых модель вообще не объясняет данные, и 1 для моделей, которые объясняют все отклонения в данных.

Коэффициент R^2 позволяет оценивать эффективность модели независимо от предметной области.

Давайте рассмотрим ещё один пример, где с R^2 не все так гладко. Для примера возьмем данные о 30 игроках профессиональной баскетбольной команды (табл. 2.1).

Допустим, что мы хотим построить линейную регрессию, чтобы предсказать размер спонсорской поддержки (y) по описательным признакам (Возраст, Рост, Вес). Рассчитаем несколько моделей каждый раз добавляя по одному новому признаку и посмотрим, как будет изменятся R^2. Результаты представлены в следующей таблице.

Таблица 2.2 – Коэффициент R^2

№	Предиктор	R^2
1	Возраст	0.8427
2	Рост	0.8629
3	Вес	0.8639

Мы видим, что, постепенно добавляя новые признаки в модель коэффициент R^2 увеличивается.

Скорректированный коэффициент детерминации (adjusted R^2 value)

Есть один подвох в использовании коэффициента R^2. Дело в том, что R^2 всегда увеличивается (по крайней мере не уменьшается) с увеличением количества признаков в модели, даже если они вообще не содержат никакой информации о целевой переменной. Поэтому лучше использовать скорректированный R^2, который учитывает количество предикторов, используемых в модели.

$$R^2_{\text{adj}} = 1 - \left(\frac{RSS}{TSS} \times \frac{N-1}{N-K-1}\right), \tag{2.23}$$

где N – число наблюдений, K – число предикторов.

Корректировка заключается в том, что R^2_{adj} учитывает число степеней свободы. Преимущество R^2_{adj} в том, что он будет возрастать при добавлении новых предикторов только в том случае, если оно больше, чем мы ожидаем от добавления случайных значений. Большой недостаток R^2_{adj} в, что мы теперь не можем интерпретировать его, как это было в случае с R^2, который говорит нам о доле «объясненной» дисперсии выхода

Если мы теперь посчитаем R^2_{adj} для данных о спонсорской поддержки, то получим следующие значения.

Таблица 2.3 – Коэффициент R^2_{adj}

№	Предиктор	R^2_{adj}
1	Возраст	0.8427
2	Рост	0.8580
3	Вес	0.8539

Мы видим, что добавление признака Рост позволяет нам несколько улучшить качество предсказания, однако такой фактор, как Вес уже не улучшает работы нашей модели.

2.1.6 Значимость коэффициентов регрессии

Опишем процедуру проверки гипотезы о статистической значимости коэффициентов регрессии.

Вначале мы выдвигаем нулевую гипотезу, что данный коэффициент w_j не важен для модели (*т.е. равен нулю*). Далее принятие/отклонение этой гипотезы выполняется в три этапа:

1. Вычисляется **тестовая статистика**.

2. Вычисляется вероятность того, что значение с.в. с распределением тестовой статистики будет больше и равно значению тестовой статистики. Эта вероятность называется p-значением.

3. P-значение сравнивается с предопределенным порогом значимости, и если p-значение меньше порога, то нулевая гипотеза отклоняется. Стандартные статистические пороги равны 5 и 1%.

Итак, что же мы будем считать?

Дело в том, что найденные нами коэффициенты регрессии w_j являются с.в. и, следовательно, для них можно оценить дисперсию и мат. ожидание. Известно, что w_j распределено по нормальному закону. Дисперсия коэффициента w_j линейной регрессии рассчитывается как

$$\sigma_{w_j}^2 = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x}_i)}, \tag{2.24}$$

где σ^2 – определено в (2.2)). К сожалению, σ^2 мы, как правило, не знаем, поэтому выборочная дисперсия s^2 служит её оценкой:

$$s = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - \left(w_0 + \sum_{j=1}^{d} x_j w_j\right)\right)^2}{N-2}}. \tag{2.25}$$

В результате получаем

$$s_{w_j}^2 = \frac{s^2}{\sum_{i=1}^{N}(x_i - \bar{x}_i)^2}. \tag{2.26}$$

Величина $\frac{(N-2)s_{w_j}^2}{\sigma^2}$ распределена как χ^2 с $N-2$ степенями свободы, поэтому (по определению) статистика данного теста

$$t = \frac{w_j/\sigma}{\sqrt{\frac{(N-2)s_{w_j}^2}{\sigma^2(N-2)}}} = \frac{w_j}{\sqrt{s_{w_j}^2}} \tag{2.27}$$

подчиняется t-распределению с $N-2$ степенями свободы.

В критерии (2.27) мы делим σ числитель, чтобы в числителе получилось распределение $\mathcal{N}(0,1)$. Напомним, что согласно нашей нулевой гипотезе $E\{w_j\} = 0$.

Далее, используя таблицу t-критерия мы можем определить p-значение и, если оно меньше требуемого уровня значимости, обычно 0,05 или 0,01, то мы отклоняем нулевую гипотезу и говорим, что признак x_j оказывает значительное влияние на модель.

Рассчитаем значимость коэффициентов регрессии для данных о баскетбольной команде.

Таблица 2.4 – Значимость коэффициентов модели линейной регрессии для предсказания спонсорской поддержки баскетбольного игрока

№	**Предиктор**	**Коэффициент w_j**	**Значение t**	**p-значение**
1	Возраст	-128.76	-13.935	4.0515e-14
2	Рост	8.93	3.127	0.004
3	Вес	-2.08	-1.059	0.298

Полученные данные говорят о том, что исключительное значение имеет возраст. Также высокое значение имеет Рост. В свою очередь Вес не имеет статистической значимости (p-значение больше 0,05).

2.2 Порядок выполнения работы

1) Загрузите набор данных согласно варианту (табл. 2.5).

Таблица 2.5 – Наборы данных по вариантам

Номер варианта (n_0)	Набор данных	Зависимая переменная (Y)
1	mtcars	Пробег (миль на галлон/ mpg)
2	mtcars	Полная мощность / hp
3	wine	Цена
4	Fish (категория Bream)	Вес
5	Fish (категория Roach)	Цена
6	Fish (категория Perch)	Цена
7	Real estate (записи с 1 по 50)	Цена
8	Real estate (записи с 100 по 150)	Цена

2) Используя набор данных согласно варианту самостоятельно выберите независимый параметр X_1, далее:

а) Постройте диаграмму рассеяния для выбранных величин X_1 и Y.

б) Рассчитайте коэффициент корреляции $r_{X_1,Y}$.

в) Постройте линейную регрессию Y по X_1 и выведите на экран полученную линию регрессии.

г) Выведите в консоль значение минимизированной суммы квадратов остатков (RSS).

д) Посчитайте коэффициент детерминации R^2 для полученной модели простой линейной регрессии.

3) Выберите дополнительный независимый параметр X_2 и постройте множественную линейную регрессию Y по X_1 и X_2

а) Выведите в консоль найденные параметры множественной линейной регрессии.

б) Постройте (3D) диаграмму рассеяния и плоскость, соответствующую регрессионной функции

в) Выведите в консоль значение минимизированной суммы квадратов остатков (RSS). Стало ли значение RSS меньше, чем при простой линейной регрессии?

г) Посчитайте коэффициент детерминации R^2 для модели множественной линейной регрессии.

д) Поочередно (один за другим) добавьте в модель множественной регрессии все имеющиеся описательные признаки. После каждого очередного добавления пересчитайте коэффициент детерминации R^2. Результаты сведите в таблицу. Что можно сказать по этой таблице?

е) Повторите предыдущий пункт, только на этот раз рассчитывайте скорректированный коэффициент детерминации R^2.

4) Оформить отчет в соответствии с СТП 01-2017.

2.3 Дополнительные задания

1) Изменить одно значение в наборе данных и показать чувствительность линейной регрессии к выбросам.

2) Построить поверхность ошибок $RSS(w_0, w_1)$ используя сетку значений w_0 и w_1. Плоскость w_0–w_1 называют весовым пространством.

3) Выполните нормализацию данных и повторите процесс обучения множественной регрессии. При фиксированной скорости обучения и числе итераций градиентного спуска в каком случае получается лучше результат с применением нормализации или без неё?

4) Построить график изменения функции ошибки RSS в процессе градиентного спуска.

5) Рассчитайте и выведите на экран p-значения для коэффициентов множественной линейной регрессии, полученной в задании №3.