ЛЕКЦИЯ 9. Интеллектуальный анализ данных: регрессия

Классификация и регрессия являются одними из важнейших задач анализа данных. Их объединение не случайно, поскольку в самой постановке задач классификации и регрессии много общего. Действительно, как классификационная, так и регрессионная модель находят закономерности между входными и выходными переменными. Но если входные и выходные переменные модели непрерывные перед нами задача регрессии. Если выходная переменная одна, и она является дискретной (метка класса), то речь идет о задаче классификации.

Линейная и логистическая регрессия.

Задача линейной регрессии заключается в нахождении коэффициентов уравнения линейной регрессии, которое имеет вид:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n \qquad (4.1)$$

где

y – выходная (зависимая) переменная модели;

$x_1, x_2, \ldots x_n$ – входные (независимые) переменные;

b_i – коэффициенты линейной регрессии, называемые также параметрами модели (b_0 – свободный член).

Задача линейной регрессии заключается в подборе коэффициентов b_i уравнения (4.1) таким образом, чтобы на заданный входной вектор $X = (x_1, x_2, \ldots x_n)$ регрессионная модель формировала желаемое выходное значение y.

Одним из наиболее востребованных приложений линейной регрессии является прогнозирование. В этом случае входными переменными модели x_i являются наблюдения из прошлого (предикторы), а y – прогнозируемое значение. Несмотря на свою универсальность, линейная регрессионная модель не всегда пригодна для качественного предсказания зависимой переменной. Когда для решения задачи строят модель линейной регрессии, на значения зависимой переменной обычно не налагают никаких ограничений. Но на практике такие ограничения могут быть весьма существенными. Например, выходная переменная может быть категориальной или бинарной. В таких случаях приходится использовать различные специальные модификации регрессии, одной из которых является логистическая регрессия, предназначенная для пред сказания зависимой переменной, принимающей значения в интервале от 0 до 1. Такая ситуация характерна для задач оценки вероятности некоторого события на основе значений независимых переменных.

Кроме того, логистическая регрессия используется для решения задач бинарной классификации, в которых выходная переменная может принимать только два значения 0 или 1, «Да» или «Нет» и т. д.

Таким образом, логистическая регрессия служит не для предсказания значений зависимой переменной, а скорее для оценки вероятности того, что зависимая переменная примет заданное значение.

Предположим, что выходная переменная y может принимать два возможных значения 0 и 1. Основываясь на доступных данных, можно вычислить, вероятности их появления: $P(y = 0) = 1 - p$; $P(y = 1) = p$. Иными словами, вероятность появления одноuо значения равна 1 минус вероятность появления другого, поскольку одно из них появится обязательно и их общая вероятность равна 1. Для определения этих вероятностей используется логистическая регрессия:

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (4.2)$$

Правая часть формулы (4.2) эквивалентна обычному уравнению линейной регрессии (4.1). Однако вместо непрерывной выходной переменной y в левой части отношения вероятностей двух взаимоисключающих событий (в нашем примере вероятность появления 0 и вероятность появления 1).

Функция вида log (p/(1-p)) называется логит-преобразованием и обозначается logit(p). Использование логит-преобразование позволяет ограничить диапазон изменения выходной переменной в пределах [0; 1].

Пример

Предположим, что регрессионная модель задана уравнением:

$$logit(p) = 1{,}5 - 0{,}6x_1 + 0{,}4x_2 - 0{,}3x_3.$$

Имеется наблюдение $(x_1; x_2; x_3) = (1,0,1)$. Используя логистическую модель, можно оценить вероятность появления выходного значения 1, то есть $P(Y = 1)$.

Для этоuо сначала вычислим соответствующее логит-преобразование:

$$logit(p) = 1{,}5 - 0{,}6 \times 1 + 0{,}4 \times 0 - 0{,}3 \times 1 = 0{,}6$$

То есть $log\left(\frac{p_i}{1-p_i}\right) = 0{,}6$

Затем можно вычислить вероятность появления значения 1:

$$P(Y = 1) = \frac{e^{0{,}6}}{1 + e^{0{,}6}} = 0{,}65$$

Таким образом, можно заключить, что выходное значение $Y = 1$ более вероятно, чем $Y = 0$.

Даже этот несложный пример показывает, что, несмотря на относительную простоту, логистическая регрессия может быть очень полезной при решении задачи предсказания, если наилучшим прогнозом считается наиболее вероятное значение.

Простая линейная регрессия

в Data Mining существует большой класс задач, где требуется установить зависимость между признаками (атрибутами, показателями), которые описывают исследуемый процесс или объект предметной области. Для этого строятся различные модели, в которых данные признаки выступают в качестве переменных, Если модель будет корректно отражать зависимость между входными и выходными переменными, то с помощью такой модели можно будет предсказывать значения выходной переменной по заданным значениям входных.

Начнем рассмотрение с простого примера. Владелец овощной лавки планирует оптимизировать закупки и для этого собирает статистику, отражающую зависимость объемов продаж от цены, устанавливаемой на продукты. Предполагается, что розничная цена меняется ежемесячно в зависимости от цены, устанавливаемой поставщиком, на которую, в свою очередь, влияют сезонность, качество товара, ситуация на рынке и т. д. Для розничного продавца оценка того, сколько товара он сможет продать за определенную цену, представляет большой интерес. Если такая оценка будет получена, то станет ясно, сколько и каких продуктов потребуется закупить, например, на месяц при определенной ценовой ситуации на рынке.

Результатом собранных наблюдений явилась зависимость ежемесячных продаж картофеля от установленной цены (табл. 4.1),

Таблица 4.1. Зависимость объема продаж от цены

№ месяца	Цена за 1 кг, x	Количество проданного картофеля, y, кг	Количество проданного картофеля, оцененное с помощью регрессии, $\hat{y}$, кг
1	13	1000	1323,4
2	20	600	305,6
3	17	500	741,8
4	15	1200	1032,6
5	16	1000	887,2
6	12	1500	1468,8
7	16	500	887,2
8	14	1200	1178,0

9	10	1700	1759,6
10	11	2000	1614,2

Цель анализа оценка ожидаемых объемов продаж картофеля в зависимости от установленной цены. Построим модель продаж, где в качестве входной переменной будет использоваться цена, а в качестве выходной объем продаж. На рис. 4.1 представлена диаграмма рассеяния исходных данных.

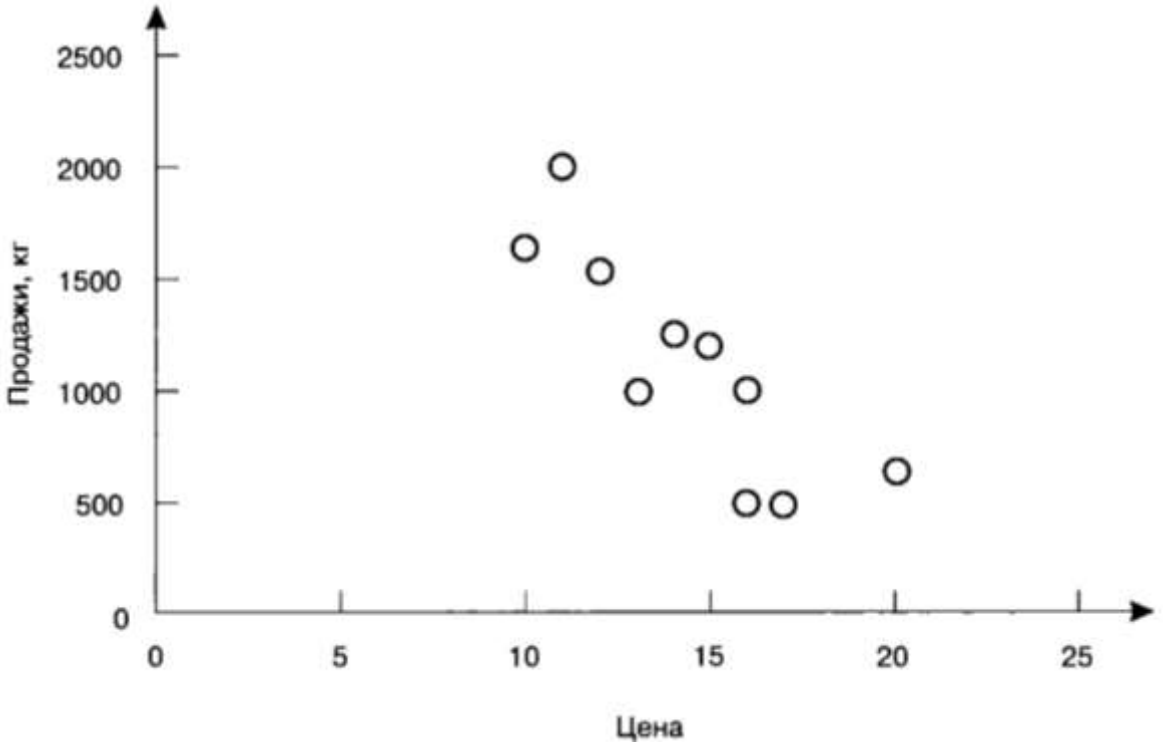

Рис. 4.1. Наблюдаемая зависимость объемов продаж от установленной цены

Даже простой визуальный анализ показывает наличие обратной зависимости между ценой и объемами продаж: с увеличением цены продажи падают. Само по себе это не является неожиданным. Однако с практический точки зрения наибольший интерес представляет количественное описание этой зависимости, а именно какого падения спроса следует ожидать при увеличении цены за единицу.

Если предположить, что зависимость между переменными линейная, то для построения модели достаточно провести прямую линию, проходящую через «облако» точек, соответствующих наблюдениям. Тогда наклон линии покажет, насколько уменьшатся продажи при увеличении цены. Но таких линий можно построить бесконечно много, и только одна из них обеспечит оптимальную оценку объемов продаж. Естественным было бы провести линию таким образом, чтобы рассеяние вдоль нее точек, соответствующих реальным наблюдениям, было минимальным. На практике линию строят так, чтобы сумма квадратов

отклонений наблюдаемых значений от оцененных с помощью данной линейной зависимости была минимальной, то есть:

$$\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2 \to min$$

где n – число наблюдений;

$\widehat{y_i}$ – оценка выходного значения для -го наблюдения, полученная с помощью модели;

y_i – реально наблюдаемое значение объема продаж.

Данный метод приближения линейной зависимости между входными и выходными переменными известен как метод наименьших квадратов (МНК), а линия, построенная с его помощью, называется линией регрессии.

Линия регрессия – это прямая наилучшего приближения для множества пар значений входной и выходной переменной (x, y), выбираемая таким образом, чтобы сумма квадратов расстояний от точек (x_i, y_i) до этой прямой, измеренных вертикально (то есть вдоль оси y), была минимальна, Уравнение, описывающее линию регрессии, называется уравнением регрессии:

$$\hat{y} = b_0 + b_1 x \qquad (4.3)$$

где $\hat{y}$– оценка значения выходной переменной;

b_0 - коэффициент, определяющий точку пересечения линии с осью y, называемый также свободным членом Коэффициент b_1 определяет наклон линии относительно оси x (иногда ero называют *угловым коэффициентом*). Проще говоря,

b_1 – это величина, на которую изменяется значение выходной переменной y при изменении входной переменной x на единицу, Коэффициенты линейного уравнения b_0 и b_1 называются *коэффициентами регрессии*.

В зарубежной литературе регрессию часто называют предсказанием (prediction). При этом входную переменную называют предсказывающей переменной, или предиктором (predictor variable), а выходную переменную – предсказываемой (predicted variable).

Таким образом, задача построения модели линейной регрессии сводится к нахождению таких коэффициентов b_0 и b_1, для которых сумма квадратов ошибок, то есть разностей между реально наблюдаемыми значениями выходной переменной y_i и их оценками $\hat{y}$, была бы минимальна. Уравнение регрессии с учетом ошибки между наблюдаемым и оцененным значениями будет

$$\hat{y} = b_0 + b_1 x + \varepsilon$$

где ε – ошибка.

Тогда сумму квадратов ошибок по всем наблюдениям можно вычислить следующим образом:

$$E = \sum_{i=1}^{n} \varepsilon^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_i x_i)^2 \quad (4.4)$$

Можно найти значения b_0 и b_1, которые минимизируют $\sum_{i=1}^{n} \varepsilon^2$, путем дифференцирования уравнения (4.3) по b_0 и b_1. Частные производные для уравнения (4.4) по b_0 и b_1 соответственно будут:

$$\frac{\partial E}{\partial b_0} = -2 \sum_{i=1}^{n} (\hat{y}_i - b_0 - b_i x_i)\,; \frac{\partial E}{\partial b_1} = -2 \sum_{i=1}^{n} (\hat{y}_i - b_0 - b_i x_i). \quad (4.5)$$

В точке, где функция минимальна, ее производная обращается в ноль. Поэтому нас интересуют значения b_0 и b_1, которые обращают (4.5) в ноль, то есть

$$\sum_{i=1}^{n} (y_i - b_0 - b_i x_i)^2 = 0; \sum_{i=1}^{n} x_i (y_i - b_0 - b_i x_i)^2 = 0$$

Опустив некоторые промежуточные выкладки, сразу запишем результат:

$$b_1 = \frac{\frac{\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\frac{\left(\sum_{i=1}^{n} {x_i}^2\right) - (\sum_{i=1}^{n} x_i)^2}{n}}$$

$$b_0 = \frac{1}{n} \sum_{i=1}^{n} y_i + \frac{b_i}{n} \sum_{i=1}^{n} x = \bar{y} - b_i \bar{x} \quad (4.6)$$

где n – общее число наблюдений;

$\bar{y}$ – среднее значение выходной переменной;

$\bar{x}$ – среднее значение входной переменной,

Уравнения (4.6) это полученные методом МНК для значений b_0 и b_1 оценки, которые минимизируют сумму квадратов ошибок

Разности между наблюдаемыми значениями выходной переменной и значениями, оцененными с помощью регрессии, называются остатками. Справедливо:

$$\text{наблюдение} = \text{оценка} + \text{остаток}$$

Используя МНК, вычислим оценки коэффициентов регрессии для данных из табл. 4.1.

$$b_1 = \frac{1493000 - 1612800}{21560 - 20735} = -145{,}4;\ b_0 = 1120 + 2093{,}6 = 3213{,}6$$

Соответствующее уравнение регрессии будет иметь вид:

$$y = 3213{,}6 - 145{,}4x$$

Смысл коэффициентов уравнения регрессии следующий: b_0 – это значение выходной переменной y при значении входной переменной $x = 0$. Значит, при цене картофеля, равной нулю, оценка объемов продаж составит 3213,6 кг. Однако данная формальная интерпретация явно противоречит здравому смыслу,

поскольку если раздавать картофель бесплатно, то купят любое его доступное количество. Такая ситуация возникла из-за того, что в исходной выборке наблюдений отсутствуют значения x, близкие к нулю. Отсюда вытекает одно из ограничений линейной регрессии: линию регрессии (и, соответственно, описывающее ее уравнение) следует считать подходящей аппроксимацией некоторой реальной функции только в том диапазоне изменений входной переменной x, в котором распределены исходные наблюдения. В противном случае результаты могут оказаться непредсказуемыми.

Значение коэффициента наклона линии регрессии b_1 можно интерпретировать как среднюю величину изменения значения выходной переменной при изменении значения входной переменной на единицу. В нашем примере это означает, что при увеличении цены за один килограмм картофеля на одну денежную единицу можно ожидать уменьшения спроса в среднем на 145,4 кг. Линия регрессии для найденного нами уравнения представлена на рис. 4.2.

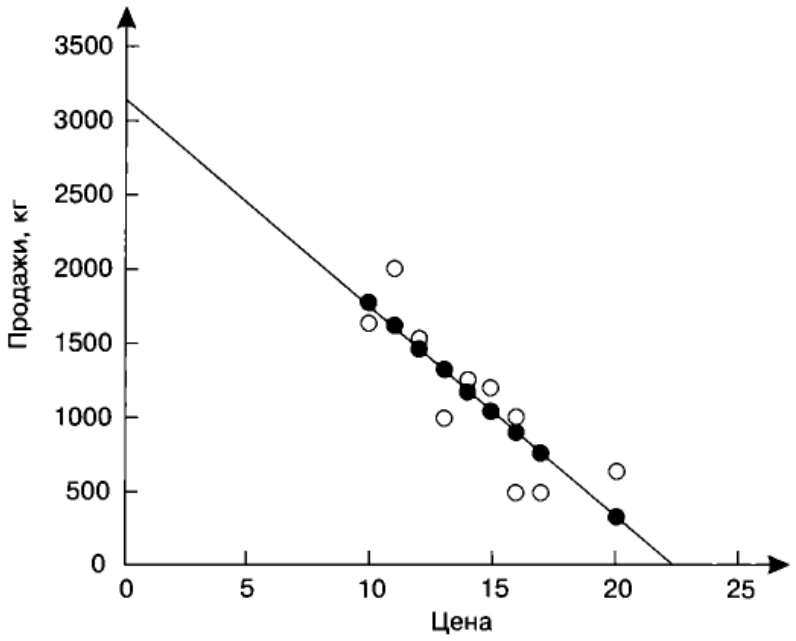

Рис. 4.2. Линия регрессии для примера из табл. 4.1

Для линии регрессии сумма квадратов вертикальных расстояний между точками данных (светлые точки) и линией должна быть меньше, чем аналогичная сумма квадратов для любой другой прямой,

Оценка соответствия простой линейной регрессии реальным данным

Линия регрессии должна аппроксимировать линейную зависимость между входной и выходной переменными модели, Однако при этом возникает вопрос, насколько линейная аппроксимация соответствует наблюдаемым данным. Чтобы определить это, введем в рассмотрение два показателя стандартную ошибку оценивания $\widehat{E_{\text{ст}}}$ и коэффициент детерминации, обозначаемый r^2.

В статистике мерой разброса случайной величины относительно среднего значения является стандартное отклонение. Аналогично в качестве меры разброса точек наблюдений относительно линии регрессии можно использовать стандартную ошибку оценивания, которая показывает среднюю величину отклонения точек исходных данных от линии регрессии вдоль оси y. Стандартная ошибка равна корню квадратному *среднеквадратической ошибки* ($E_{СКО}$), то есть сумме квадратов разностей между реальным и оцененным значениями, вычисленной по всем наблюдениям и отнесенной к числу степеней свободы выборки:

$$E_{СКО} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - m - 1}$$

где m – количество независимых переменных, которое для простой линейной регрессии равно 1.

$E_{СКО}$ можно рассматривать как меру изменчивости выходной переменной, объясняемую регрессией. Тогда стандартная ошибка оценивания определится следующим образом:

$$E_{\text{ст}} = \sqrt{E_{СКО}}$$

Значение стандартной ошибки $E_{\text{ст}}$ позволяет оценить степень рассогласования оценок, полученных с помощью регрессии, и реальных наблюдений аналогично тому, как стандартное отклонение позволяет оценить в статистическом анализе степень разброса случайной величины относительно среднего. Чем меньше стандартная ошибка оценивания, тем лучше работает модель.

Рассмотрим пример. Пусть требуется оптимизировать закупку бензина для сети автозаправочных станций (АЗС). Для этого была исследована выборка, в которой представлены данные, описывающие объемы продаж бензина определенной марки десятью АЗС, В ней представлено количество бензина в тоннах, проданное АЗС за определенное число дней. На основе наблюдений было получено уравнение регрессии $\hat{y} = 6 + 2x$. Объем продаж оценивается как 6 т плюс удвоенное количество дней, в течение которых осуществлялись продажи. Уравнение позволяет оценить количество тонн бензина, которое может быть продано за заданный интервал времени. Полученные оценки представлены в табл. 4.2.

Таблица 4.2. Расчет среднеквадратической ошибки для примера о продажах бензина

№	Время, x,	Объем	Оцененное	Ошибка	$(y - \hat{y})^2$

АЗС	дней	продаж, y, т.	количество, $\hat{y} = 6 + 2x$	оценивания, $y - \hat{y}$	
1	2	10	10	0	0
2	2	11	10	1	1
3	3	12	12	0	0
4	4	13	14	-1	1
5	4	14	14	0	0
6	5	15	16	-1	1
7	6	20	18	2	4
8	7	18	20	-2	4
9	8	22	22	0	0
10	9	25	24	1	1
$\sum(y - \hat{y})^2$					12

Из табл. 4.2 видно, что сумма квадратов ошибок оценивания $\sum(y - \hat{y})^2 = 12$. Эта величина представляет собой общую меру ошибки оценивания значения выходной переменной с помощью данного уравнения регрессии. Если она велика, то модель работает неудовлетворительно. Является ли полученное значение суммы квадратов ошибок, равное 12, большим? Этого достоверно сказать нельзя, поскольку на данном этапе мы не имеем других мер для сравнения.

Для примера, который представлен в табл. 4.2, стандартная ошибка будет $E_{\text{ст}} = \sqrt{\frac{12}{10-1-1}} \cong 1{,}2$. Следовательно, при оценке объема продаж АЗС с помощью уравнения $\hat{y} = 6 + 2x$ ожидаемая ошибка равна 1,2 т,

Теперь предположим, что информация о времени, в течение которого осуществлялись продажи, отсутствует, то есть использовать переменную x для оценивания переменной y невозможно. Полученные в этом случае оценки объемов продаж окажутся менее точными, поскольку количество исходной информации уменьшится, Тогда единственно возможной оценкой для y будет простое среднее значение $\bar{y} = 16$. Рассмотрим рис. 4.3, где оценка объемов продаж, не учитывающая информацию о времени, за которое они были сделаны, представлена горизонтальной линией . $\bar{y} = 16$.

Отсутствие информации о времени приводит к оценке объемов продаж 16т для всех АЗС независимо от отработанного времени. Очевидно, что ценность такой оценки сомнительна. Действительно, одни АЗС расположены на оживленных трассах, а другие на второстепенных, плотность транспортного потока меняется под действием различных факторов. Поэтому ожидать, что АЗС,

отработавшие 1-2 дня и 5-7 дней, продадут одинаковое количество бензина, наивно.

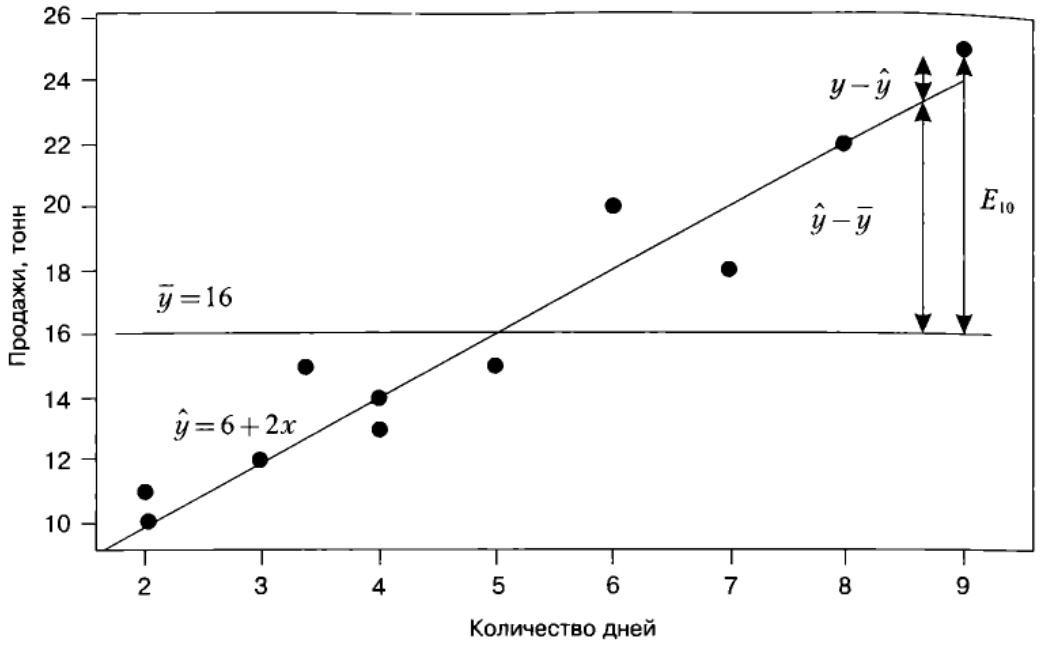

Рис. 4.3. Геометрическая интерпретация линий регрессии

Точки на рис. 4.3 сосредоточены вдоль линии регрессии, а не вдоль линии $\bar{y} = 16$. То есть предполагается, что суммарная ошибка оценивания будет меньше, если использовать информацию о времени. Например, рассмотрим АЗС №10, на которой было продано $y = 25$т бензина за 9 дней. Если игнорировать информацию о времени и предположить, что на ней, как и на остальных АЗС, продано только 16т, то ошибка оценивания составит $E_{10} = y - \bar{y} = 9$т.

Сравним стандартную ошибку $E_{\text{ст}} = 1{,}2$ и ошибку относительно среднего значения $\bar{y} = 16$, когда информация о предсказывающей переменной игнорируется:

$$E_{\text{ст}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} = 5{,}0$$

Таким образом, типичной ошибкой предсказания для случая, когда информация о значениях предсказывающей переменной не используется, будет 5,0 т.

Следовательно, применение регрессии вместо оценки на основе простого среднего значения позволяет уменьшить ошибку с 5,0 до 1,2 т, то есть более чем в 4 раза. Это позволяет сделать вывод о значимости полученной регрессионной модели. Если бы значения стандартной ошибки, полученные для оценок регрессии и простого среднего значения, были примерно одинаковы, это говорило

бы о том, что регрессия практически не дает выигрыша в точности оценки по сравнению с обычным средним наблюдаемых значений, то есть о низкой значимости регрессионной модели,

Изменчивость выходной переменной

Чтобы оценить степень соответствия регрессии реальным данным, полезно ввести меры, количественно характеризующие поведение выходной переменной. В качестве таких мер используются три квадратичные суммы: общая (полная) Q, регрессионная Q_R и ошибки (остаточная) Q_E. Чтобы пояснить смысл данных величин, представим наблюдаемое значение выходной переменной в виде $y = \hat{y} + (y - \hat{y})$, то есть наблюдаемое значение равно сумме euo оценки и ошибки оценивания, Данное выражение может быть записано в виде $y = (b_0 + b_1 x) + (y - b_0 - b_1 x)$.Здесь y – наблюдаемое значение, $(b_0 + b_1 x)$ – член, описывающий долю изменчивости выходной переменной, объясняемой регрессией, $(y - b_0 - b_1 x)$ – член, описывающий отклонение выходной переменной от линии регрессии (остаток). Если все точки данных лежат непосредственно на линии регрессии (идеальное соответствие), то остатки будут равны 0. Если из обеих частей предыдущего выражения вычесть среднее значение $\bar{y}$, то есть:

$$y - \bar{y} = (\hat{y} - \bar{y}) + (y - \hat{y})$$

то можно показать, что суммы квадратов складываются следующим образом:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y - \hat{y})^2$$

или, используя введенные выше обозначения для квадратичных сумм:

$$Q = Q_R + Q_E$$

Таким образом,

$$Q = \sum_{i=1}^{n} (y_i - \bar{y})^2 ;\ Q_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 ;\ Q_E = \sum_{i=1}^{n} (y - \hat{y})^2$$

Следовательно, полная изменчивость выходной переменной складывается из части, объясняемой регрессией (линейной зависимостью), и ошибки, то есть части, не объясненной регрессией.

Нужно ли ожидать, что полная квадратичная сумма Q, полученная только при использовании среднего значения $, \bar{y}$ будет больше или меньше, чем остаточная сумма Q_E, полученная при оценивании с учетом входной переменной? Используя данные табл. 4.2, получим, что $Q = 228$ окажется намного больше, чем $Q_E = 12$ (табл.4.3).

Таблица 4.3. Расчет полной квадратичной суммы для примера о продажах бензина

№ АЗС	Время, x, дней	Объем продаж, y, т.	$\bar{y}$	$y-\bar{y}$	$(y-\bar{y})^2$
1	2	10	16	-6	36
2	2	11	16	-5	25
3	3	12	16	-4	16
4	4	13	16	-3	9
5	4	14	16	-2	4
6	5	15	16	-1	1
7	6	20	16	4	16
8	7	18	16	2	4
9	8	22	16	6	36
10	9	25	16	9	81
Q					228

Теперь сравнib две меры точности оценивания выходной переменной. Поскольку Q_E намного меньше, чем Q, можно сделать вывод, что использование информации о входной переменной в задаче регрессии позволяет получить более точные оценки, чем без ее использования, то есть когда в качестве оценки используется простое среднее значение наблюдений выходной переменной.

Определим меру, которая позволит увидеть, насколько улучшились оценки, полученные с помощью уравнения регрессии, Для этого вернемся к рис. 4.3.

Для АЗС №10 ошибка, полученная при использовании уравнения регрессии, составит $y-\hat{y}=25-24=1$. Если же производить оценку только на основе среднего значения $\bar{y}=16$, то полученная ошибка составит $y-\bar{y}=25-16=9$. Таким образом, ошибка оценивания уменьшится на $\hat{y}-\bar{y}=24-16=8$, то есть на разность оценок с помощью регрессии и простого среднего.

Коэффициент детерминации

Введем понятие коэффициента детерминации r^2, который показывает степень согласия регрессии как приближения линейного отношения между входной и выходной переменными с реальными данными:

$$r^2=\frac{Q_R}{Q}$$

Значение коэффициента детерминации максимально, когда имеет место идеальное соответствие: все точки данных лежат точно на прямой регрессии. В

этом случае квадратичная сумма Q_E оценки, полученной с помощью регрессии, равна 0. Тогда $Q = Q_R$ и $r^2 = \frac{Q_R}{Q} = 1$. Максимальное значение коэффициента детерминации, равное 1, имеет место только тогда, когда уравнение регрессии идеально описывает связь между входной и выходной переменными.

Чтобы определить максимальное значение коэффициента детерминации, предположим, что регрессия совсем не улучшает точность оценки по сравнению с использованием среднего значения, то есть не объясняет изменчивость выходной переменной. В этом случае $Q_R = 0$, а значит, и $r^2 = 0$. Таким образом, коэффициент детерминации может изменяться от 0 до 1 включительно. При этом чем выше значение r^2, тем больше регрессионная модель соответствует реальным данным. Значения r^2, близкие к 1, означают очень хорошее соответствие регрессионной модели реальным данным, а значения, близкие к 0, очень плохое.

Вычислим значение коэффициента детерминации для примера из табл. 8.3:

$$r^2 = \frac{Q_R}{Q} = \frac{216}{228} = 0{,}9474$$

Можно сделать вывод, что регрессионная модель работает хорошо,

Коэффициент корреляции

Еще одной мерой, используемой для количественного описания линейной зависимости между двумя числовыми переменными, является коэффициент корреляции, который определяется следующим образом:

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{(n-1)\sigma_x\sigma_y} \qquad (4.1)$$

где σ_x и σ_y стандартные отклонения соответствующих переменных. Значение коэффициента корреляции всегда расположено в диапазоне от -1 до 1 включительно и может быть интерпретировано следующим образом:

- Если коэффициент корреляции близок к 1, то между переменными имеет место сильная положительная корреляция. Иными словами, наблюдается высокая степень зависимости входной и выходной переменных (если значения входной переменной x возрастают, то и значения выходной переменной y также будут увеличиваться).
- Если коэффициент корреляции близок $k - 1$, это означает, что между переменными наблюдается отрицательная корреляция: поведение выходной переменной будет противоположным поведению входной (когда значение x возрастает y уменьшается, и наоборот).

- Промежуточные значения указывают на слабую корреляцию между переменными и, соответственно, на низкую зависимость между ними: поведение входной переменной x совсем (или почти совсем) не будет влиять на поведение y.

Зависимость степени связи между переменными от конкретных значений коэффициента корреляции в большинстве задач можно определить с помощью приближенной оценки. Но следует отметить, что в технических приложениях, например, при обнаружении целей в радиолокации на фоне шумов и помех необходимо применять более строгие подходы, Для приближенной оценки можно воспользоваться следующей шкалой (табл. 4.4).

Таблица 8.4. Шкала соответствия для коэффициента корреляции

Коэффициент корреляции, r	Корреляция
$0{,}6 < r < 1$	Высокая положительная
$0{,}3 \le r \le 0{,}6$	Средняя положительная
$-0{,}3 < r < 0{,}3$	Корреляция отсутствует
$-0{,}6 \le r \le -0{,}3$	Средняя отрицательная
$1 < r < 0{,}6$	Средняя отрицательная

Вычисление коэффициента корреляции по формуле (4.1) может быть несколько громоздкими, поскольку в знаменателе требуется найти стандартные отклонения. для x и y/ Поэтому при вычислении коэффициента корреляции для примера из табл, 4/2 воспользуемся упрощенным выражением:

$$r = \frac{\sum xy - (\sum x \sum y)/n}{\sqrt{\sum x^2 - (\sum x)^2/n} \times \sqrt{\sum y^2 - (y)^2/n}} = \frac{908 - (50 \times 60)/10}{\sqrt{304 - 50^2/10} \times \sqrt{2788 - 160^2/10}} = 0{,}9733$$

Видно, что между временем, в течение которого производились продажи, и количеством проданного бензина существует сильная корреляция (чего, впрочем, и следовало ожидать): с увеличением времени продаж, количество проданного бензина увеличивается. Чтобы вычислить коэффициент корреляции, можно использовать коэффициент детерминации r^2, то есть $r = \pm\sqrt{r^2}$,

Если коэффициент уравнения регрессии $b_1 > 0$ (линия регрессии возрастает), то и коэффициент корреляции также будет положительным, то есть $r = \sqrt{r^2}$. В противном случае коэффициент корреляции будет отрицательным, то есть $r = -\sqrt{r^2}$

Поскольку в нашем примере $b_1 = 2$, коэффициент корреляции будет положительным и равным $r = \sqrt{r^2} = \sqrt{0{,}9474} = 0{,}9733$