ЛЕКЦИЯ 10. Простая регрессия

Простая регрессионная модель

Для многих практических приложений найти уравнение регрессии и построить соответствующую линейную зависимость недостаточно. Чтобы решить задачу, требуется построить регрессионную модель. В чем же отличие простой линейной регрессии от регрессионной модели? Дело в том, что уравнение регрессии в большинстве случаев строится на основе выборочных данных. При этом из множества наблюдений, называемого генеральной совокупностью, выделяется некоторая выборка, на основе свойств которой предполагается делать выводы о свойствах всей совокупности. Для данной выборки и строится уравнение регрессии. Затем производится попытка обобщить полученную линейную зависимость на всю генеральную совокупность.

То есть простая линейная регрессия строится на основе выборки наблюдений, каждое из которых содержит соответствие $x \to y$, а регрессионная модель предполагает, что обнаруженная линейная зависимость имеет место для всех возможных пар $x \to y$, содержащихся в генеральной совокупности. Однако при построении регрессионной модели возникает ряд проблем, выходящих за рамки построения уравнения регрессии. Во-первых, нужно определить, справедливо ли предположение о том, что линейная зависимость, обнаруженная для выборочных наблюдений, будет истинна для всей совокупности. Во-вторых, ситуация осложняется тем, что регрессия, вообще говоря, не является детерминированной задачей, поскольку для реальных данных зависимость изменения выходной переменной от изменения входной носит случайный характер. Следовательно, коэффициенты регрессии могут рассматриваться как статистики, описывающие распределение выходной переменной.

Простая линейная регрессионная модель задается следующим образом. Пусть имеется выборка данных, содержащая п наблюдений, в каждом из которых значению независимой переменной x_i соответствует значение зависимой переменной y_i, связанных с помощью линейной зависимости:

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

где β_0 и β_1 – параметры модели, определяющие точку пересечения линии регрессии с осью y и наклон линии регрессии соответственно;

ε – член, определяющий ошибку отклонения реального наблюдения от оценки, полученной с помощью данной модели.

Поскольку для реальных данных отношение между входной переменной x и выходной переменной y носит случайный характер, любое его линейное приближение будет характеризоваться некоторой ошибкой. Для учета этой

ошибки в модель необходимо ввести соответствующий элемент, также представляющий собой случайную переменную. При этом допускается несколько предположений.

- Предположение о нулевом среднем. Член, учитывающий ошибку, является случайной переменной с математическим ожиданием, равным 0, то есть $m(\varepsilon) = 0$.
- Предположение о постоянной дисперсии. Дисперсия ε (обозначим ее σ_ε^2) постоянна.
- Предположение о независимости. Отдельные значения ε являются независимыми.
- Предположение о нормальности. Ошибка ε является нормально распределенной случайной переменной.

Иными словами, ошибка ε_i является независимой нормально распределенной случайной переменной с нулевым математическим ожиданием и дисперсией σ_ε^2. На основе этих предположений можно сделать выводы о поведении зависимой переменной. Статистическая модель простой линейной регрессии предполагает, что для каждого значения входной переменной x наблюдаемое значение выходной переменной y является нормально распределенной случайной величиной со средним $E(y) = \beta_0 + \beta_1 x$ и постоянной дисперсией σ_ε^2. Данное предположение иллюстрируется на рис. 10.1 для случаев $x = 5$, $x = 10$ и $x = 15$. Видно, что все кривые нормального распределения имеют одну и ту же форму, из чего следует, что дисперсия постоянна для всех x.

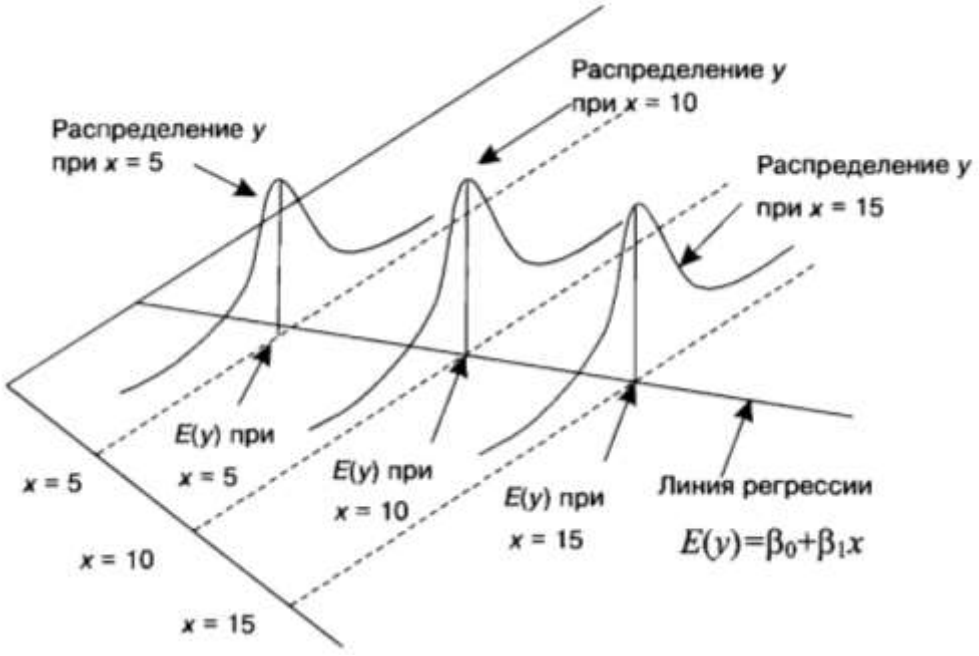


Рис. 10.1. Для каждого значения x значение y есть нормально распределенная случайная величина

Если при решении задачи регрессии строить модель не требуется, то проверять состоятельность данных предположений не обязательно, поскольку они касаются только ошибки ε. Если ошибка не учитывается, то и предположения не нужны. Однако если нужно построить модель, то предположения должны быть проверены.

Поясним смысл предположений о поведении зависимой переменной y.

- На основе предположения о нулевом среднем имеем:

$E(y) = E(\beta_0 + \beta_1 x + \varepsilon) = E(\beta_0) + E(\beta_1 x) + E(\varepsilon) = \beta_0 + \beta_1 x$

Для каждого значения x среднее значение y лежит на линии регрессии.

- На основе предположения о постоянстве дисперсии $D(y)$ имеем

$$D(y) = D(\beta_0 + \beta_1 x + \varepsilon) = D(\varepsilon) = \sigma^2$$

Независимо от того, какое значение принимает переменная x, дисперсия переменной y всегда является постоянной величиной.

- Из предположения о независимости следует, что для любого x значения y также будут независимыми.

- Из предположения о нормальности следует, что y также является нормально распределенной случайной переменной.

Таким образом, выходная переменная y является независимой нормально распределённой случайной переменной со средним, равным $\beta_0 + \beta_1$ и дисперсией σ^2.

Гипотезы в регрессии

Пусть для уравнения линейной регрессии, построенного на некоторой выборке, определен коэффициент детерминации $r^2 = 0{,}2$. Это значение показывает, что соответствующая модель вряд ли является значимой, то есть линейная зависимость между входной и выходной переменными отсутствует. Но отсутствует ли она полностью? Возможно, линейная зависимость между переменными имеет место даже при очень малых значениях коэффициента детерминации. Существует ли системный подход, позволяющий определить наличие линейной зависимости между входной и выходной переменными? Методическую основу для оценки значимости таких зависимостей создают гипотезы линейной регрессии,

Рассмотрим модель простой линейной регрессии [2]:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Данная модель утверждает, что между выходной переменной y и входной переменной x существует линейная зависимость. Здесь β_1 – параметр модели,

константа, значение которой неизвестно. Существует ли такое значение β_1, при котором линейная зависимость между переменными x и y будет отсутствовать?

Рассмотрим случай $\beta_1 = 0$. Тогда регрессионная модель примет вид

$$y = \beta_0 + 0x + \varepsilon = \beta_0 + \varepsilon$$

Следовательно, линейная зависимость между переменными x и y перестанет существовать. Но если β_1 примет любое отличное от нуля значение, то она будет наблюдаться. Отсюда следует важный вывод: линейная зависимость между переменными x и y зависит от параметра β_1.

Возможна ситуация, кода регрессия, построенная на основе выборки, обнаружит линейную зависимость между входной и выходной переменными, в то время как регрессионная модель, построенная для всей совокупности, нет. Таким образом, если между переменными некоторой выборки существует линейная зависимость из этого не обязательно следует существование линейной зависимости для всей совокупности. Это значит, что даже если коэффициент уравнения регрессии $b_1 = 0$, то коэффициент регрессионной модели β_1 может оказаться равным 0 и линия регрессии, вычисленная с использованием всех наблюдений совокупности, окажется горизонтальный. Поясним данную ситуацию (рис. 1.5).

На рис. 10.2 светлыми кругами обозначены точки, соответствующие наблюдениям генеральной совокупности. Визуальный анализ показывает, что такой набор точек даст горизонтальною линию регрессии ($\beta_1 = 0$), представленную на рисунке сплошной линией. В то же время линия регрессии, построенная по выборочным наблюдениям (темные круги), может иметь некоторый наклон (штрихпунктирная линия), и в этом случае $b_1 \neq 0$, что указывает на наличие положительной линейной зависимости между переменными x и y.

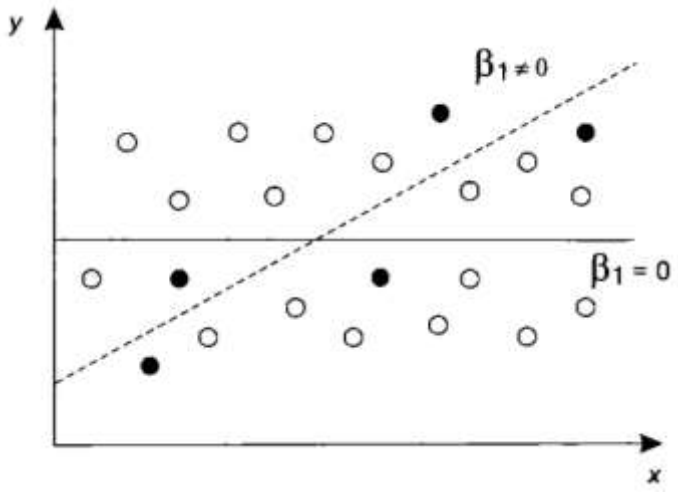


Рис. 10.2. Отсутствие линейной зависимости во всей совокупности

Поскольку зависимость выходной переменной от входной носит случайный характер, то и оценки параметра наклона линии регрессии b являются статистическими. Как и все такие оценки, они имеют распределение с некоторым средним значением и стандартным отклонением. Стандартное отклонение для оценки b_1 определяется следующим образом:

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\sum x^2 - (\sum x)^2/n}}$$

где σ – стандартное отклонение, вычисленное для всей совокупности.

Как предположение о среднем значении генеральной совокупности может быть сделано на основе выборочного среднего $\bar{x}$, так и предположение о значении β_1 может быть сделано на основе выборочной оценки b_1.

На основе σ_{b_1}, можно получить показатель (обозначим его s_{b_1}), отражающий меру изменчивости b_1. Определим его следующим образом:

$$s_{b_1} = \frac{E_{\text{ст}}}{\sqrt{\sum x^2 - (\sum x)^2/n}}$$

где $E_{\text{ст}}$ – стандартная ошибка оценивания. Большие значения s_{b_1} показывают, что оценка параметра неустойчива, нестабильна.

Во многих задачах интерес представляет не оценка значения параметра модели β_1 для генеральной совокупности, а некоторое утверждение о нем. В этом случае задача сводится к проверке гипотезы, по результатам которой гипотеза будет либо подтверждена, либо отклонена.

Введем в рассмотрение две гипотезы – нулевую (основную) и альтернативную. Нулевая гипотеза H_0 предполагает, что линейные связи между переменными отсутствуют, в то время как альтернативная гипотеза H_a утверждает, что такие связи имеют место. Таким образом:

- $H_0: \beta_1 = 0$ – линейные связи между переменными отсутствуют;
- $H_a: \beta_1 \neq 0$ – линейные связи между переменными имеют место.

Если верна гипотеза H_0, то линия регрессии будет расположена горизонтально и построенная регрессионная модель бесполезна. С таким же успехом можно использовать для оценки простое среднее значение выходной переменной, вычисленное по всем имеющимся наблюдениям. Наоборот, если удастся отвергнуть нулевую гипотезу, подтвердится значимость модели, то есть наличие линейной зависимости между входной и выходной переменными. Но чтобы сделать окончательное заключение о пригодности модели к решению той или иной задачи, необходимо определить степень ее значимости. Для этого можно использовать силу отклонения нулевой гипотезы. Для подтверждения или

отклонения гипотезы используется некоторый статистический показатель, значение которого должно быть вычислено на основе имеющихся наблюдений. Например, если значение, принятое показателем, очень маловероятно в предположении истинности гипотезы, то истинность гипотезы также маловероятна и ее следует отклонить. При этом уверенность в отклонении гипотезы тем сильнее, чем меньше вероятность появления данного значения показателя. Такое проверочное значение часто называют p-значением. Чем меньше p-значением, тем выше ожидаемая значимость модели и сильнее линейная зависимость между входными и выходной переменными.

Для проверки истинности нулевой и альтернативной гипотез существуют различные критерии. Например, гипотеза H_0 может утверждать, что наблюдаемое выборочное распределение значений переменной является нормальным, а альтернативная гипотеза – что оно таковым не является. Тогда вероятность значения проверочной статистики либо подтвердит, что данные имеют нормальное распределение, либо покажет степень его отклонения от нормального.

Одними из наиболее популярных критериев оценки значимости регрессионных моделей являются t-критерий, построенный на основе t-распределения Стьюдента и F-критерий, использующий F-распределение Фишера.

Оценка значимости регрессионной модели: t – критерий и F – критерий

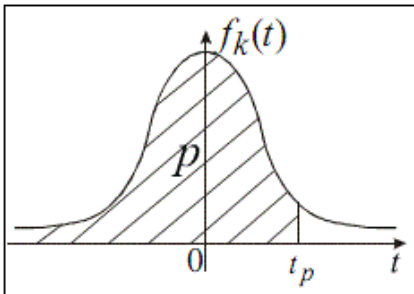
Для проверки истинности гипотезы H_0 t-критерий использует статистику $t = b_1/s_{b_1}$ подчиняющуюся t-распределению с $(n-2)$ степенями свободы. При проверке гипотез возможно появление двух видов ошибок, представленных в табл. 10.1.

Таблица 10.1. Виды ошибок, возникающих при проверке гипотез

Действительное состояние	H_0 принимается	H_0 отвергается
H_0 справедлива	Верное решение	Ошибка I рода: вероятность α
H_0 несправедлива	Ошибка II рода: вероятность β	Верное решение

Ошибка I рода заключается в отклонении верной нулевой гипотезы, а ошибка II рода в принятии неверной нулевой. Вероятности данных событий обозначим α и β соответственно, причем вероятность α называется уровнем значимости критерия. Он определяет вероятность того, что справедливая гипотеза H_0 будет ошибочно отвергнута. Задав значение, α и вычислив значение проверочной статистики t, по таблице t-распределения можно установить,

насколько вероятно появление полученного значения статистики t, если нулевая гипотеза будет ошибочно отвергнута (рис. 10.3).



Число степеней свободы k	Вероятность, p					
	0,9	**0,95**	**0,975**	**0,99**	**0,995**	**0,9995**
1	2	3	4	5	6	7
1	3,078	6,314	12,706	31,821	63,657	636,619
2	1,886	2,920	4,303	6,965	9,925	31,598
3	1,638	2,353	3,182	4,541	5,841	12,941
4	1,533	2,132	2,776	3,747	4,604	8,610
5	1,476	2,015	2,571	3,365	4,032	6,869
6	1,440	1,943	2,447	3,143	3,707	5,959
7	1,415	1,895	2,365	2,998	3,499	5,405
8	1,397	1,860	2,306	2,896	3,355	5,041
9	1,383	1,833	2,262	2,821	3,250	4,781
10	1,372	1,812	2,228	2,764	3,169	4,587
12	1,356	1,782	2,179	2,681	3,055	4,318
14	1,345	1,761	2,145	2,625	2,977	4,140
16	1,337	1,746	2,120	2,584	2,921	4,015
18	1,330	1,734	2,101	2,552	2,878	3,922
20	1,325	1,725	2,086	2,528	2,845	3,850
22	1,321	1,717	2,074	2,508	2,819	3,792
24	1,318	1,711	2,064	2,492	2,797	3,745
26	1,315	1,706	2,056	2,479	2,779	3,707
28	1,313	1,701	2,048	2,467	2,763	3,674
30	1,310	1,697	2,042	2,457	2,750	3,646
60	1,296	1,671	2,000	2,390	2,660	3,460
120	1,289	1,658	1,980	2,358	2,617	3,373
∞	1,282	1,645	1,960	2,326	2,576	3,291

Рис. 10.3. Таблица t-распределения

В качестве примера возьмем данные о продажах картофеля, для которых было получено уравнение регрессии $\hat{y} = 3213{,}6 - 145{,}4x$. Тогда

$$s_{b_1} = \frac{E_{\text{ст}}}{\sqrt{\frac{\sum x^2 - (\sum x)^2}{n}}} = \frac{E_{\text{ст}}}{\sqrt{\sum (x - \bar{x})^2}} = \frac{272{,}5}{\sqrt{82{,}4}} = 30$$

$$t = b_1 / s_{b_1} = -\frac{145{,}5}{30} = -4{,}85$$

Чтобы оценить вероятность справедливости гипотезы H_0 при значении статистики $t = -4{,}85$, воспользуемся методом p-значения, Под p-значением в критерии проверки гипотезы понимается вероятность получения значения статистики t, большего или равного значению выборочной статистики, вычисленной исходя из предположения о справедливости гипотезы H_0. Следовательно, p-значение представляет собой часть выборочного распределения, расположенную правее значения выборочной статистики t, выраженную в процентах. Иными словами, чем меньше p-значение, тем больше вероятность того, что гипотеза несправедлива, и если вычисленное p-значение будет очень малым, то гипотезу следует отвергнуть.

Из таблицы t-распределения (рис. 10.3) можно увидеть, что для $(n-2) = 10-2 = 8$ степеней свободы значение 3,355 соответствует $\mathrm{p} = 0{,}005$. Это указывает на вероятность справедливости гипотезы, равную 0,5%. Значение статистики $|t| = 4{,}85$, полученное на основе регрессии, существенно больше, чем 3,355. Следовательно, для данного значения вероятность справедливости гипотезы еще меньше. Можно сделать вывод о высокой вероятности несостоятельности гипотезы H_0, поскольку появление значения $|t| = 4{,}85$, очень маловероятно, если гипотеза H_0 истинна. Тогда предположение о том, что $\beta_1 = 0$ и линейная зависимость между переменными отсутствует, также несостоятельно.

Дружим возможным методом определения значимости регрессионной модели является F-критерий. В этом случае используются два показателя – средний квадрат регрессионной квадратичной суммы (в случае простой линейной регрессии $m = 1$):

$$S_R^2 = \frac{Q_R}{m} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{1}$$

и средний квадрат ошибок регрессии:

$$S_E^2 = \frac{Q_E}{n-m-1} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{n-1}$$

С их помощью можно вычислить значение, называемое F-статистикой. Отношение $F = S_R^2/S_E^2$ подчиняется F-распределению с $(1, n-m-1)$ степенями свободы. Если гипотеза H_0: $\beta_1 = 0$ истинна и линейная связь между переменными отсутствует, то ошибка оценивания возрастает и ее средний квадрат S_E^2, соответственно, тоже. Поэтому если нулевая гипотеза справедлива, то значение F-статистики будет мало из-за увеличения знаменателя S_E^2. Напротив, если справедлива альтернативная гипотеза H_a: $\beta_1 \neq 0$, то ошибка оценивания падает, а значение F возрастает. Таким образом, большее значение F согласуется с истинностью альтернативной гипотезы.

Например, значение F-статистики, рассчитанное для данных о продажах картофеля, будет $F = 1742032/74280 = 23{,}5$. По таблице F-распределения (рис. 10.4) можно установить, что для чисел степеней свободы числителя и знаменателя, равных соответственно 1 и $n - m - 1 = 10 - 1 - 1 = 8$, рассчитанное по наблюдаемым данным значение $F = 23{,}5$ намного превышает теоретическое $F_{\propto} = F_{0,01} = 11{,}26$. Таким образом, вероятность того, что гипотеза H_0: $\beta_1 = 0$ истинна для рассчитанного значения F-статистики, будет менее 0,01, Поэтому мы вынуждены отвергнуть нулевую гипотезу, что позволяет сделать вывод о высокой значимости нашей регрессионной модели.

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞
8	5,32 11,26 25,42	4,46 8,65 18,49	4,07 7,59 15,83	3,84 7,10 14,39	3,69 6,63 13,49	3,58 6,37 12,86	3,44 6,03 12,04	3,28 5,67 11,19	3,12 5,28 10,30	2,99 4,86 9,35
9	5,12 10,56 22,86	4,26 8,02 16,39	3,86 6,99 13,90	3,63 6,42 12,56	3,48 6,06 11,71	3,37 5,80 11,13	3,23 5,47 10,37	3,07 5,11 9,57	2,90 4,73 8,72	2,71 4,31 7,81
10	4,96 10,04 21,04	4,10 7,56 14,91	3,71 6,55 12,55	3,48 5,99 11,28	3,33 5,64 10,48	3,22 5,39 9,92	3,07 5,06 9,20	2,91 4,71 8,45	2,74 4,33 7,64	2,54 3,91 6,77
11	4,84 9,65 19,69	3,98 7,20 13,81	3,59 6,22 11,56	3,36 5,67 10,35	3,20 5,32 9,58	3,09 5,07 9,05	2,95 4,74 8,35	2,79 4,40 7,62	2,61 4,02 6,85	2,40 3,60 6,00
12	4,75 9,33 18,64	3,88 6,93 12,98	3,49 5,95 10,81	3,26 5,41 9,63	3,11 5,06 8,89	3,00 4,82 8,38	2,85 4,50 7,71	2,69 4,16 7,00	2,50 3,78 6,25	2,30 3,36 5,42
13	4,67 9,07 17,81	3,80 6,70 12,31	3,41 5,74 10,21	3,18 5,20 9,07	3,02 4,86 8,35	2,92 4,62 7,86	2,77 4,30 7,21	2,60 3,96 6,52	2,42 3,59 5,78	2,21 3,16 4,97
14	4,60 8,86 17,14	3,74 6,51 11,78	3,34 5,56 9,73	3,11 5,03 8,62	2,96 4,69 7,92	2,85 4,46 7,44	2,70 4,14 6,80	2,53 3,80 6,13	2,35 3,43 5,41	2,13 3,00 4,60
15	4,45 8,68 16,59	3,68 6,36 11,34	3,29 5,42 9,34	3,06 4,89 8,25	2,90 4,56 7,57	2,79 4,32 7,09	2,64 4,00 6,47	2,48 3,67 5,81	2,29 3,29 5,10	2,07 2,87 4,31
16	4,41 8,53 16,12	3,63 6,23 10,97	3,24 5,29 9,01	3,01 4,77 7,94	2,85 4,44 7,27	2,74 4,20 6,80	2,59 3,89 6,20	2,42 3,55 5,55	2,24 3,18 4,85	2,01 2,75 4,06

Рис. 10.4. F-распределение

Значения $F_{\text{табл}}$, удовлетворяющие условию $P(F > F_{\text{табл}})$. первое значение соответствует вероятности 0,05, второе - вероятности 0,01 и третье – вероятности 0,001, ν_1 – число степеней свободы числителя, ν_2 – число степеней свободы знаменателя.

Метрики качества линейных регрессионных моделей

Для того чтобы модель линейной регрессии можно было применять на практике необходимо сначала оценить её качество. Для этих целей предложен ряд показателей, каждый из которых предназначен для использования в различных ситуациях и имеет свои особенности применения (линейные и нелинейные, устойчивые к аномалиям, абсолютные и относительные, и т.д.). Корректный выбор меры для оценки качества модели является одним из важных факторов успеха в решении задач анализа данных.

«Хорошая» аналитическая модель должна удовлетворять двум, зачастую противоречивым, требованиям — как можно лучше соответствовать данным и при этом быть удобной для интерпретации пользователем. Действительно, повышение соответствия модели данным как правило связано с её усложнением (в случае регрессии — увеличением числа входных переменных модели). А чем сложнее модель, тем ниже её интерпретируемость.

Поэтому при выборе между простой и сложной моделью последняя должна значимо увеличивать соответствие модели данным чтобы оправдать рост сложности и соответствующее снижение интерпретируемости. Если это условие не выполняется, то следует выбрать более простую модель.

Таким образом, чтобы оценить, насколько повышение сложности модели значимо увеличивает её точность, необходимо использовать аппарат оценки качества регрессионных моделей. Он включает в себя следующие меры:

- Среднеквадратичная ошибка (Mean Squared Error – MSE).
- Корень из среднеквадратичной ошибки (Root Mean Squared Error – RMSE).
- Среднеквадратичная ошибка в процентах (Mean Squared Percentage Error – MSPE).
- Средняя абсолютная ошибка (Mean Absolute Error – MAE).
- Средняя абсолютная ошибка в процентах (Mean Absolute Percentage Error – MAPE).
- Симметричная средняя абсолютная процентная ошибка (Symmetric Mean Absolute Percentage Error – SMAPE).
- Средняя абсолютная масштабированная ошибка (Mean absolute scaled error – MASE)
- Средняя относительная ошибка (Mean Relative Error – MRE).
- Среднеквадратичная логарифмическая ошибка (Root Mean Squared Logarithmic Error – RMSLE).

- Коэффициент детерминации R-квадрат.
- Скорректированный коэффициент детерминации.

Прежде чем перейти к изучению метрик качества, введём некоторые базовые понятия, которые нам в этом помогут. Для этого рассмотрим рисунок 10.5.

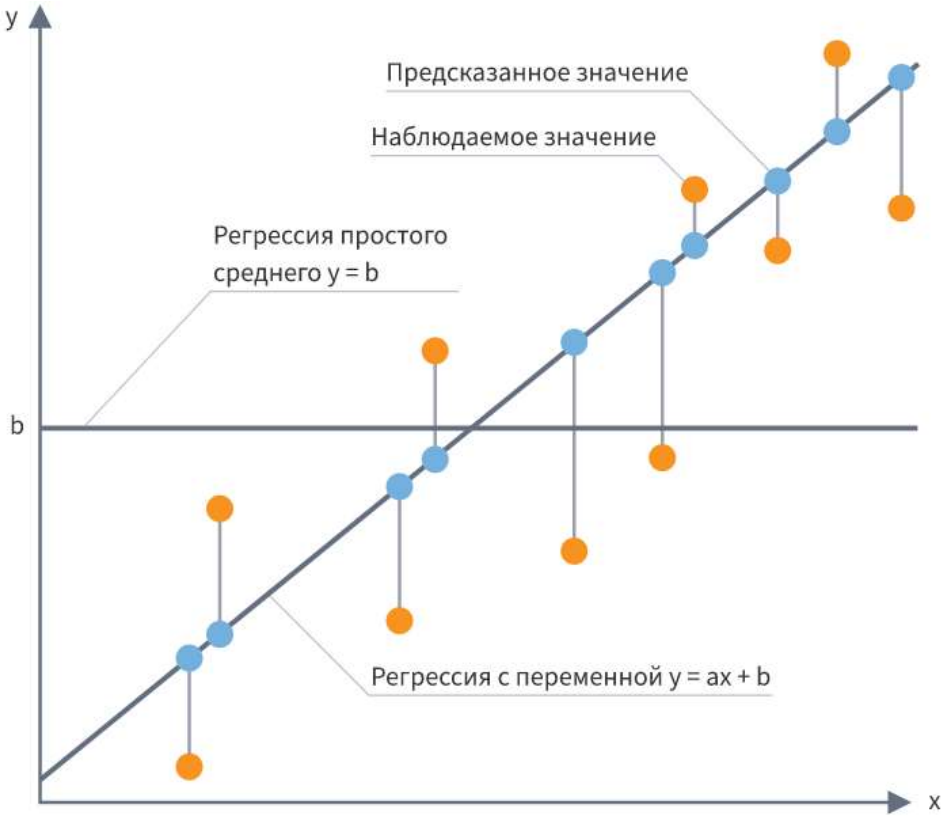


Рис. 10.5 Линейная регрессия

Наклонная прямая представляет собой линию регрессии с переменной, на которой расположены точки, соответствующие предсказанным значениям выходной переменной $\hat{y}$ (кружки синего цвета). Оранжевые кружки представляют фактические (наблюдаемые) значения y. Расстояния между ними и линией регрессии — это ошибка предсказания модели $y - \hat{y}$ (невязка, остатки). Именно с её использованием вычисляются все приведённые меры качества.

Горизонтальная линия представляет собой модель простого среднего, где коэффициент при независимой переменной x равен нулю, и остаётся только свободный член b, который становится равным среднему арифметическому фактических значений выходной переменной, т.е. $b = \bar{y}$. Очевидно, что такая модель для любого значения входной переменной будет выдавать одно и то же значение выходной – $\bar{y}$

В линейной регрессии такая модель рассматривается как «бесполезная», хуже которой работает только «случайный угадыватель». Однако, она

используется для оценки, насколько дисперсия фактических значений y относительно линии среднего, больше, чем относительно линии регрессии с переменной, т.е. насколько модель с переменной лучше «бесполезной».

MSE

Среднеквадратичная ошибка (Mean Squared Error) применяется в случаях, когда требуется подчеркнуть большие ошибки и выбрать модель, которая дает меньше именно больших ошибок. Большие значения ошибок становятся заметнее за счет квадратичной зависимости.

Действительно, допустим модель допустила на двух примерах ошибки 5 и 10. В абсолютном выражении они отличаются в два раза, но если их возвести в квадрат, получив 25 и 100 соответственно, то отличие будет уже в четыре раза. Таким образом модель, которая обеспечивает меньшее значение MSE допускает меньше именно больших ошибок.

MSE рассчитывается по формуле:

$$\text{MSE} = \frac{1}{n}(y_i - \hat{y}_i)^2$$

где n – количество наблюдений по которым строится модель и количество прогнозов, y_i – фактические значение зависимой переменной для i-го наблюдения, $\hat{y}_i$ – значение зависимой переменной, предсказанное моделью.

Таким образом, можно сделать вывод, что MSE настроена на отражение влияния именно больших ошибок на качество модели.

Недостатком использования MSE является то, что если на одном или нескольких неудачных примерах, возможно, содержащих аномальные значения будет допущена значительная ошибка, то возведение в квадрат приведёт к ложному выводу, что вся модель работает плохо. С другой стороны, если модель даст небольшие ошибки на большом числе примеров, то может возникнуть обратный эффект — недооценка слабости модели.

RMSE

Корень из среднеквадратичной ошибки (Root Mean Squared Error) вычисляется просто как квадратный корень из MSE (среднеквадратичная ошибка):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

MSE и RMSE могут минимизироваться с помощью одного и того же функционала, поскольку квадратный корень является неубывающей функцией.

Например, если у нас есть два набора результатов работы модели, A и B, и MSE для A больше, чем MSE для B, то мы можем быть уверены, что RMSE для A больше RMSE для B. Справедливо и обратное: если $MSE(A) < MSE(B)$, то и $RMSE(A) < RMSE(B)$.

Следовательно, сравнение моделей с помощью RMSE даст такой же результат, что и для MSE. Однако с MSE работать несколько проще, поэтому она более популярна у аналитиков. Кроме этого, имеется небольшая разница между этими двумя ошибками при оптимизации с использованием градиента:

$$\frac{\partial RMSE}{\partial \hat{y}_i} = \frac{1}{2\sqrt{MSE}} \frac{\partial MSE}{\partial \hat{y}_i}$$

Это означает, что перемещение по градиенту MSE эквивалентно перемещению по градиенту RMSE, но с другой скоростью, и скорость зависит от самой оценки MSE. Таким образом, хотя RMSE и MSE близки с точки зрения оценки моделей, они не являются взаимозаменяемыми при использовании градиента для оптимизации.

Влияние каждой ошибки на RMSE пропорционально величине квадрата ошибки. Поэтому большие ошибки оказывают непропорционально большое влияние на RMSE. Следовательно, RMSE можно считать чувствительной к аномальным значениям.

MSPE

Среднеквадратичная ошибка в процентах (Mean Squared Percentage Error) представляет собой относительную ошибку, где разность между наблюдаемым и фактическим значениями делится на наблюдаемое значение и выражается в процентах:

$$MSPE = \frac{100}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{y_i} \right)^2$$

Проблемой при использовании MSPE является то, что, если наблюдаемое значение выходной переменной равно 0, значение ошибки становится неопределённым.

MSPE можно рассматривать как взвешенную версию MSE (среднеквадратичной ошибки), где вес обратно пропорционален квадрату наблюдаемого значения. Таким образом, при возрастании наблюдаемых значений ошибка имеет тенденцию уменьшаться.

MAE

Средняя абсолютная ошибка (Mean Absolute Error) вычисляется следующим образом:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Т.е. MAE рассчитывается как среднее абсолютных разностей между наблюдаемым и предсказанным значениями. В отличие от MSE и RMSE она является линейной оценкой, а это значит, что все ошибки в среднем взвешены одинаково. Например, разница между 0 и 10 будет вдвое больше разницы между 0 и 5. Для MSE (среднеквадратичная ошибка) и RMSE (корень из среднеквадратичной ошибки), как отмечено выше, это не так.

Поэтому MAE широко используется, например, в финансовой сфере, где ошибка в 10 долларов должна интерпретироваться как в два раза худшая, чем ошибка в 5 долларов.

MAPE

Средняя абсолютная процентная ошибка (Mean Absolute Percentage Error) вычисляется следующим образом:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i|}$$

Эта ошибка не имеет размерности и очень проста в интерпретации. Её можно выражать как в долях, так и в процентах. Если получилось, например, что $MAPE = 11{,}4$, то это говорит о том, что ошибка составила 11,4% от фактического значения.

SMAPE

Симметричная средняя абсолютная процентная ошибка (Symmetric Mean Absolute Percentage Error) – это мера точности, основанная на процентных (или относительных) ошибках. Обычно определяется следующим образом:

$$SMAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{\frac{(|y_i| + |\hat{y}_i|)}{2}}$$

т.е. абсолютная разность между наблюдаемым и предсказанным значениями делится на полусумму их модулей. В отличие от обычной MAPE (средней абсолютной процентной ошибки), симметричная имеет ограничение на диапазон значений. В приведённой формуле он составляет от 0 до 200%. Однако, поскольку диапазон от 0 до 100% гораздо удобнее интерпретировать, часто используют формулу, где отсутствует деление знаменателя на 2.

Одной из возможных проблем SMAPE является неполная симметрия, поскольку в разных диапазонах ошибка вычисляется неодинаково. Это

иллюстрируется следующим примером: если $y_i = 100$ и $\hat{y}_i = 110$, то $SMAPE = 4{,}76$ $(n = 2)$, а если $y_i = 100$ и $\hat{y}_i = 90$, то $SMAPE = 5{,}26$ $(n = 2)$

Ограничение SMAPE заключается в том, что, если наблюдаемое или предсказанное значение равно 0, ошибка резко возрастет до верхнего предела (200% или 100%).

MASE

Средняя абсолютная масштабированная ошибка (Mean absolute scaled error) — это показатель, который позволяет сравнивать две модели. Если поместить MAE (среднюю абсолютную ошибку) для новой модели в числитель, а MAE для исходной модели в знаменатель, то полученное отношение и будет равно MASE. Если значение MASE меньше 1, то новая модель работает лучше, если MASE равно 1, то модели работают одинаково, а если значение MASE больше 1, то исходная модель работает лучше, чем новая модель. Формула для расчета MASE имеет вид:

$$MASE = \frac{MAE_i}{MAE_j}$$

MASE симметрична и устойчива к выбросам.

MRE

Средняя относительная ошибка (Mean Relative Error) вычисляется по формуле:

$$MRE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i|}$$

Несложно увидеть, что данная мера показывает величину абсолютной ошибки относительно фактического значения выходной переменной (поэтому иногда эту ошибку называют также средней относительной абсолютной ошибкой, MRAE). Действительно, если значение абсолютной ошибки, скажем, равно 10, то сложно сказать много это или мало. Например, относительно значения выходной переменной, равного 20, это составляет 50%, что достаточно много. Однако относительно значения выходной переменной, равного 100, это будет уже 10%, что является вполне нормальным результатом.

Очевидно, что при вычислении MRE нельзя применять наблюдения, в которых $y_i = 0$.

Таким образом, MRE позволяет более адекватно оценить величину ошибки, чем абсолютные ошибки. Кроме этого она является безразмерной величиной, что упрощает интерпретацию.

RMSLE

Среднеквадратичная логарифмическая ошибка (Root Mean Squared Logarithmic Error) представляет собой RMSE (rорень из среднеквадратичной ошибки), вычисленную в логарифмическом масштабе:

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(\hat{y}_i + 1) - \log(y_i + 1))^2}$$

Константы, равные 1, добавляемые в скобках, необходимы чтобы не допустить обращения в 0 выражения под логарифмом, поскольку логарифм нуля не существует.

Известно, что логарифмирование приводит к сжатию исходного диапазона изменения значений переменной. Поэтому применение RMSLE целесообразно, если предсказанное и фактическое значения выходной переменной различаются на порядок и больше.

R-квадрат

Перечисленные выше ошибки не так просто интерпретировать. Действительно, просто зная значение средней абсолютной ошибки, скажем, равное 10, мы сразу не можем сказать хорошая это ошибка или плохая, и что нужно сделать чтобы улучшить модель.

В этой связи представляет интерес использование для оценки качества регрессионной модели не значения ошибок, а величину показывающую, насколько данная модель работает лучше, чем модель, в которой присутствует только константа, а входные переменные отсутствуют или коэффициенты регрессии при них равны нулю.

Именно такой мерой и является коэффициент детерминации (Coefficient of determination), который показывает долю дисперсии зависимой переменной, объяснённой с помощью регрессионной модели. Наиболее общей формулой для вычисления коэффициента детерминации является следующая:

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

Практически, в числителе данного выражения стоит среднеквадратическая ошибка оцениваемой модели, а в знаменателе — модели, в которой присутствует только константа.

Главным преимуществом коэффициента детерминации перед мерами, основанными на ошибках, является его инвариантность к масштабу данных. Кроме того, он всегда изменяется в диапазоне от $-\infty$ до 1. При этом значения близкие к 1 указывают на высокую степень соответствия модели данным.

Очевидно, что это имеет место, когда отношение в формуле стремится к 0, т.е. ошибка модели с переменными намного меньше ошибки модели с константой. $R^2 = 0$ показывает, что между независимой и зависимой переменными модели имеет место функциональная зависимость.

Когда значение коэффициента близко к 0 (т.е. ошибка модели с переменными примерно равна ошибке модели только с константой), это указывает на низкое соответствие модели данным, когда модель с переменными работает не лучше модели с константой.

Кроме этого, бывают ситуации, когда коэффициент R^2 принимает отрицательные значения (обычно небольшие). Это произойдёт, если ошибка модели среднего становится меньше ошибки модели с переменной. В этом случае оказывается, что добавление в модель с константой некоторой переменной только ухудшает её (т.е. регрессионная модель с переменной работает хуже, чем предсказание с помощью простой средней).

На практике используют следующую шкалу оценок. Модель, для которой $R^2 > 0.5$, является удовлетворительной. Если $R^2 > 0.8$, то модель рассматривается как очень хорошая. Значения, меньшие 0.5 говорят о том, что модель плохая.

Скорректированный R-квадрат

Основной проблемой при использовании коэффициента детерминации является то, что он увеличивается (или, по крайней мере, не уменьшается) при добавлении в модель новых переменных, даже если эти переменные никак не связаны с зависимой переменной.

В связи с этим возникают две проблемы. Первая заключается в том, что не все переменные, добавляемые в модель, могут значимо увеличивать её точность, но при этом всегда увеличивают её сложность. Вторая проблема — с помощью коэффициента детерминации нельзя сравнивать модели с разным числом переменных. Чтобы преодолеть эти проблемы используют альтернативные показатели, одним из которых является скорректированный коэффициент детерминации (Adjasted coefficient of determinftion).

Скорректированный коэффициент детерминации даёт возможность сравнивать модели с разным числом переменных так, чтобы их число не влияло на статистику R^2, и накладывает штраф за дополнительно включённые в модель переменные. Вычисляется по формуле:

$$R^2_{adj} = 1 - \frac{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{(n-k)}}{\sum_{i=1}^{n} \frac{(\bar{y}_i - y_i)^2}{(n-1)}}$$

где n – число наблюдений, на основе которых строится модель, k – количество переменных в модели.

Скорректированный коэффициент детерминации всегда меньше единицы, но теоретически может принимать значения и меньше нуля только при очень малом значении обычного коэффициента детерминации и большом количестве переменных модели.

Сравнение метрик

Мера	Сильные стороны	Слабые стороны
MSE	Позволяет подчеркнуть большие отклонения, простота вычисления.	Имеет тенденцию занижать качество модели, чувствительна к выбросам. Сложность интерпретации из-за квадратичной зависимости.
RMSE	Простота интерпретации, поскольку измеряется в тех же единицах, что и целевая переменная.	Имеет тенденцию занижать качество модели, чувствительна к выбросам.
MSPE	Нечувствительна к выбросам. Хорошо интерпретируема, поскольку имеет линейный характер.	Поскольку вклад всех ошибок отдельных наблюдений взвешивается одинаково, не позволяет подчёркивать большие и малые ошибки.
MAPE	Является безразмерной величиной, поэтому её интерпретация не зависит от предметной области.	Нельзя использовать для наблюдений, в которых значения выходной переменной равны нулю.
SMAPE	Позволяет корректно работать с предсказанными значениями независимо от того больше они фактического, или меньше.	Приближение к нулю фактического или предсказанного значения приводит к резкому росту ошибки, поскольку в знаменателе присутствует как фактическое, так и предсказанное значения.

MASE	Не зависит от масштаба данных, является симметричной: положительные и отрицательные отклонения от фактического значения учитываются одинаково. Устойчива к выбросам. Позволяет сравнивать модели.	Сложность интерпретации.
MRE	Позволяет оценить величину ошибки относительно значения целевой переменной.	Неприменима для наблюдений с нулевым значением выходной переменной.
RMSLE	Логарифмирование позволяет сделать величину ошибки более устойчивой, когда разность между фактическим и предсказанным значениями различается на порядок и выше	Может быть затруднена интерпретация из-за нелинейности.
R-квадрат	Универсальность, простота интерпретации.	Возрастает даже при включении в модель бесполезных переменных. Плохо работает когда входные переменные зависимы.
R-квадрат скорр.	Корректно отражает вклад каждой переменной в модель.	Плохо работает, когда входные переменные зависимы.