Статистический анализ регрессионной модели. О выборе допустимой модели регрессии

Статистический анализ *модели регрессии* (7.9), построенной на основе параметризации искомой функции регрессии $f(x)$ в виде (7.2) и на основе *МНК-оценок* параметров, состоит из следующих трех этапов:

– проверка адекватности модели регрессии;

– проверка значимости модели регрессии и ее параметров;

– анализ точности результатов, полученных с использованием регрессионной модели.

Для проведения статистического анализа требуется дополнить исходные предположения *метода наименьших квадратов* еще одним. Будем считать, что *случайные ошибки* ε_i, $i=\overline{1,n}$, в модели (7.3) не только независимы, но и распределены по нормальному закону: $\varepsilon_i \sim N(0,\sigma^2)$, $i=\overline{1,n}$, т.е. случайная составляющая $\vec{\varepsilon}=(\varepsilon_1,\ \ldots,\ \varepsilon_n)^{\mathrm{T}}$ *линейной регрессионной модели* (7.6) имеет n-мерный нормальный закон распределения с нулевым средним значением и ковариационной матрицей $\sigma^2 I_n$.

Это предположение в силу (7.3) эквивалентно тому, что наблюдения Y_i, $i=\overline{1,n}$, являются независимыми нормально распределенными случайными величинами, т.е.

$$Y_i \sim N\left(f_a(\vec{x}^{\,i}),\sigma^2\right), \tag{7.22}$$

где

$$f_a(\vec{x}^{\,i})=\sum_{k=0}^{m-1}\beta_k\psi_k(\vec{x}^{\,i}),\quad i=\overline{1,n}.$$

Проверку рассматриваемого предположения проводят на основе статистического анализа случайных величин

$$\varepsilon_i = Y_i-\widehat{Y}(\vec{x}^{\,i}),\quad i=\overline{1,n},$$

значения которых представляют собой отклонения наблюдаемых значений y_i отклика Y от его значений, предсказанных моделью регрессии

$$\widehat{y}(\vec{x}^{\,i}) = \sum_{k=0}^{m-1} \widehat{\beta}_i \psi_k(\vec{x}^{\,i}).$$

Таким образом, все сводится к проверке *статистической гипотезы* о выполнении исходных предположений: случайные величины ε_i, $i = \overline{1, n}$, являются независимыми и $\varepsilon_i \sim N(0, \sigma^2)$, $i = \overline{1, n}$. Критерии проверки указанных гипотез рассмотрены выше (см. **5**).

Следует отметить, что, когда каждая случайная величина ε_i имеет единственную *реализацию* (нет повторных наблюдений), мы не можем проверить гипотезу о независимости случайных величин ε_i, $i = \overline{1, n}$. Однако, если у исследователя есть основания считать, что случайные величины ε_i, $i = \overline{1, n}$, независимы и одинаково распределены, можно ограничиться проверкой гипотезы о том, что $\widehat{\varepsilon}_i$, $i = \overline{1, n}$ — реализация случайной величины ε_i распределенной по нормальному закону.

Считая, что исходные предположения метода наименьших квадратов выполнены, перейдем к рассмотрению этапов статистического анализа регрессионной модели.

Проверка адекватности построенной модели регрессии. ***Линейную регрессионную модель*** называют ***адекватной***, если предсказанные по ней значения отклика Y согласуются с результатами наблюдений.

В основе процедуры проверки адекватности модели лежат предположения, что случайные ошибки наблюдений ε_i, $i = \overline{1, n}$, являются независимыми, нормально распределенными случайными величинами с нулевыми средними значениями и одинаковыми дисперсиями σ^2.

Пусть для каждого или некоторых значений переменного $x = (x_1, \ldots, x_p)$ имеется несколько (r_i, $i = \overline{1, n}$) повторных наблюдений отклика Y (т.е. исходные данные представлены матрицей D — см. (**7.1**)). Тогда для проверки адекватности модели можно использовать следующую процедуру.

Итак, повторные наблюдения получены при различных значениях $\vec{x}^{\,1}, \ldots, \vec{x}^{\,n}$ переменного x, причем в точке $\vec{x} = \vec{x}^{\,i}$ про-

изведено r_i наблюдений $y_{i1}, \ldots, y_{ir_i}$ отклика Y, а $\sum\limits_{i=1}^{n} r_i = N$ — *объем выборки.* Введем обозначение

$$\overline{y}_i = \frac{1}{r_i}\sum_{j=1}^{r_i} y_{ij}.$$

Если линейная регрессионная модель адекватна, то значения $\overline{y}_i$ должны быть близки к значениям $\widehat{y}_i = \widehat{y}(\vec{x}^i)$, $i = \overline{1, n}$. Следовательно, сумму квадратов

$$Q_n = \sum_{i=1}^{n} r_i\left(\overline{y}_i - \widehat{y}(\vec{x}^i)\right)^2$$

можно рассматривать как меру неадекватности рассматриваемой модели.

Можно показать, что *статистики*

$$Q_n(\vec{Y}_N) = \sum_{i=1}^{n} r_i\left(\overline{Y}_i - \widehat{Y}(\vec{x}^i)\right)^2,$$

$$Q_p(\vec{Y}_N) = \sum_{i=1}^{n}\sum_{j=1}^{r_i}\left(Y_{ij} - \overline{Y}(\vec{x}^i)\right)^2$$

являются независимыми случайными величинами. Статистика $Q_p(\vec{Y}_N)/\sigma^2$ имеет *χ^2-распределение* с числом степеней свободы $\sum\limits_{i=1}^{n}(r_i - 1)$, а отношение

$$S_y^2(\vec{Y}_N) = \frac{Q_p(\vec{Y}_N)}{\sum\limits_{i=1}^{n}(r_i - 1)}$$

является несмещенной оценкой *остаточной дисперсии.* Эта статистика не связана с ошибкой в выборе модели. Статистика $Q_n(\vec{Y}_N)/\sigma^2$ имеет распределение χ^2 с числом степеней свободы $n - m$, если гипотеза H_0: $\mathbf{M}Y = F\vec{\beta}$ верна (здесь m — число неизвестных параметров в модели (7.2)). При этом $S^2_{\text{ад}} = Q_n(\vec{Y}_N)/(n - m)$ — несмещенная оценка σ^2.

Следовательно (см. **Д.3.1**), статистика имеет *распределение*

Фишера со степенями свободы $n-m$ и $\sum\limits_{i=1}^{n}(r_i-1)$:

$$F=\frac{S^2_{\text{ад}}(\vec{Y}_N)}{S^2_y(\vec{Y}_N)}=\frac{Q_n(\vec{Y}_N)}{n-m}\,\frac{\sum\limits_{i=1}^{n}(r_i-1)}{Q_p(\vec{Y}_N)}\sim F\Big(n-m,\sum_{i=1}^{n}(r_i-1)\Big).$$

Поэтому проверка гипотезы H_0 осуществляется стандартным образом по *критерию Фишера.*

Если *выборочное значение* $f_{\text{в}}$ статистики F не превышает критического $f_{\text{кр}}$, т.е.

$$f_{\text{в}}\leqslant f_{\text{кр}}=f_{1-\alpha}(r_n,r_p),$$

то гипотезу H_0 принимают (точнее, не отклоняют) на *уровне значимости* α, т.е. модель признается адекватной.

В противном случае модель признается неадекватной и нужно пытаться построить более сложную модель, увеличив, например, число *базисных функций* или выбрав другие базисные функции.

Пример 7.5. Найдем МНК-оценки параметров *простой линейной регрессии*

$$f_a(x)=\beta_0+\beta_1 x$$

по данным табл. 7.3 и проверим адекватность модели регрессии на уровне значимости $\alpha=0{,}05$.

Таблица 7.3

x_i	1	2	3	2,7	4,3	5,0
y_{ij}	0,5; 0,1	0,5; 1,2	1,2; 1,7	0,9; 2,2	1,1; 1,7; 2,5	2,0; 2,2
r_i	2	2	2	2	3	2

Имеем $\sum\limits_{i=1}^{n}r_i=N=13$, $n=6$, $m=2$,

$$Q_p=\sum_{i=1}^{6}\sum_{j=1}^{r_i}(y_{ij}-\overline{y}_i)=2{,}29.$$

По формулам (7.20) находим

$$\widehat{\beta}_1 = \frac{9{,}68}{23{,}12} = 0{,}419, \qquad \widehat{\beta}_0 = \frac{17{,}8 - 0{,}419 \cdot 40{,}3}{13} = 0{,}07.$$

Итак, $\widehat{y}(x) = 0{,}07 + 0{,}419x$. Далее вычисляем

$$Q_n = \sum_{i=1}^{6} r_i \left(\overline{y}_i - \widehat{y}(x_i)\right)^2 \approx 0{,}39$$

и рассчитываем выборочное значение

$$f_{\text{в}} = \frac{0{,}39/(6-2)}{2{,}29/(13-6)} \approx 0{,}3$$

статистики

$$F = \frac{Q_n(\vec{Y}_N)}{n-m} \frac{\sum_{i=1}^{n}(r_i - 1)}{Q_p(\vec{Y}_N)}.$$

Поскольку критическое значение $f_{\text{кр}} = f_{0{,}95}(4,7) = 4{,}14$ (см. табл. П.5) существенно больше $f_{\text{в}}$, то построенную модель регрессии можно считать адекватной результатам наблюдений.

7.4. О выборе допустимой модели регрессии

Как уже отмечалось выше, при решении задач *регрессионного анализа* исследователь в первую очередь сталкивается с необходимостью выбора класса $\mathcal{F}$ *допустимых моделей регрессии.* Мы не останавливаемся на этой проблеме* и еще раз отметим, что при ее решении, как правило, исследователь исходит из преследуемых целей, собственного опыта, результатов предварительного анализа, имеющегося экспериментального материала и т.д.

Если класс $\mathcal{F}$ содержит, например, две допустимые модели регрессии, то возникает проблема выбора наилучшей (в каком-то смысле) *допустимой модели регрессии.* Обсуждение этой проблемы можно найти в специальной литературе**, а мы ограничимся рассмотрением *линейной регрессионной модели* (см. (7.6)). При этом будем предполагать, что выполнены основные допущения регрессионного анализа: независимость и нормальное распределение случайных величин ε_i, $i = \overline{1,n}$

Пусть имеем две допустимые модели регрессии

$$\sum_{k=0}^{m_1-1} \beta_k \psi_k(\vec{x}) \quad \text{и} \quad \sum_{k=0}^{m_2-1} \beta_k \psi_k(\vec{x}), \tag{7.30}$$

где $m_2 > m_1$ и объем выборки равен n. Проверим *гипотезу*

$$H_0\colon\ \beta_{m_1} = \beta_{m_1+1} = \ldots = \beta_{m_2-1} = 0$$

против *альтернативной гипотезы*

$$H_1\colon \sum_{k=m_1}^{m_2-1} \beta_k^2 \neq 0.$$

Для проверки гипотезы H_0 можно применить *статистику*

$$F = \frac{Q_{l1}(\vec{Y}_n) - Q_{l2}(\vec{Y}_n)}{Q_{l2}(\vec{Y}_n)} \frac{n - m_2}{m_2 - m_1}, \tag{7.31}$$

где $Q_{l1}(\vec{Y}_n)$ и $Q_{l2}(\vec{Y}_n)$ — *остаточные суммы квадратов* соответственно для первой и второй моделей (7.30). Статистика F имеет *распределение Фишера* с числом степеней свободы $m_2 - m_1$ и $n - m_1 - m_2$.

Гипотезу H_0 следует принять на *уровне значимости* α (принять модель $\sum_{k=0}^{m_1-1} \beta_k \psi(\vec{x})$), если значение $f_{\text{в}}$ статистики F, рассчитанное по результатам наблюдений, не превышает $f_{\text{кр}} = f_{1-\alpha}(m_2 - m_1, n - m_1 - m_2)$.

Заметим, что при $\widehat{Q}_{l2} > \widehat{Q}_{l1}$ всегда следует выбирать модель $\sum_{k=0}^{m_1-1} \beta_k \psi(\vec{x})$.

Рассмотренный критерий называют ***критерием отношения остаточных дисперсий***. Смысл его прозрачен: усложнение допустимой модели регрессии статистически оправдано, если это приводит к значимому (на уровне значимости α) уменьшению значения оценки *остаточной дисперсии*.

Пример 7.7. Вернемся к примеру 7.6. Результаты наблюдений дают основание утверждать, что допустимыми моделями регрессии являются

$$\sum_{k=0}^{1}\beta_k\psi_k(\vec{x}) \qquad \text{и} \qquad \sum_{k=0}^{2}\beta_k\psi_k(\vec{x}).$$

С помощью *метода наименьших квадратов* находим значения оценок для параметров β_k, $k=\overline{0,1}$, первой модели регрессии. Для второй модели оценки параметров найдены в примере 7.6. Имеем

$$\widehat{y}_1(x)=6{,}92+2{,}27x \qquad \text{и} \qquad \widehat{y}_2(x)=6{,}92+2{,}27x+0{,}08x^2.$$

Коэффициент β_2 во второй модели незначим (см. пример 7.6).

Применяя статистику (7.31), проверим гипотезу H_0: $\beta_2=0$ против альтернативной гипотезы H_1: $\beta_2\neq 0$.

В нашем случае $n=11$, $m_1=2{,}28$, $m_2=0{,}08$. Рассчитываем остаточные суммы квадратов $Q_{l1}=393{,}84$ и $Q_{l2}=455{,}21$. Значения оценок *остаточных дисперсий* соответственно равны 43,76 и 56,90. Поскольку $56{,}90>43{,}76$, то следует выбрать модель $\widehat{y}_1(x)=6{,}91+2{,}28x$.

Пример 7.11. Считая, что зависимость между x и y имеет вид $y=\beta_0+\beta_1x+\beta_2x^2$, найдем значения оценок параметров и проверим значимость *модели регрессии* на уровне $\alpha=0{,}1$ по выборке, представленной в табл. 7.9.

Таблица 7.9

x_i	26	30	34	38	42	46	50
y_i	3,94	4,60	5,67	6,93	7,73	8,25	9,56

По данным выборки запишем матрицы

$$F=\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 26 & 30 & 34 & 38 & 42 & 46 & 50 \\ 676 & 900 & 1156 & 1444 & 1764 & 2116 & 2500 \end{pmatrix}^{\mathrm{T}}$$

и

$$Y=(3{,}94\ \ 4{,}60\ \ 5{,}67\ \ 6{,}93\ \ 7{,}73\ \ 8{,}25\ \ 9{,}56)^{\mathrm{T}}.$$

Пример 7.11. Считая, что зависимость между x и y имеет вид $y = \beta_0 + \beta_1 x + \beta_2 x^2$, найдем значения оценок параметров и проверим значимость *модели регрессии* на уровне $\alpha = 0{,}1$ по выборке, представленной в табл. 7.9.

Таблица 7.9

x_i	26	30	34	38	42	46	50
y_i	3,94	4,60	5,67	6,93	7,73	8,25	9,56

По данным выборки запишем матрицы

$$F = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 26 & 30 & 34 & 38 & 42 & 46 & 50 \\ 676 & 900 & 1156 & 1444 & 1764 & 2116 & 2500 \end{pmatrix}^{\mathrm{T}}$$

и

$$Y = (3{,}94 \;\; 4{,}60 \;\; 5{,}67 \;\; 6{,}93 \;\; 7{,}73 \;\; 8{,}25 \;\; 9{,}56)^{\mathrm{T}}.$$

Используя матрицу F, находим

$$M = F^{\mathrm{T}}F = \begin{pmatrix} 7 & 266 & 10556 \\ 266 & 10556 & 435176 \\ 10556 & 435176 & 18527600 \end{pmatrix},$$

$$M^{-1} = \begin{pmatrix} 91{,}926 & -4{,}962 & 0{,}064 \\ -4{,}962 & 0{,}271 & -3{,}534 \cdot 10^{-3} \\ 0{,}064 & -3{,}534 \cdot 10^{-3} & 4{,}65 \cdot 10^{-5} \end{pmatrix}.$$

Теперь вычисляем вектор-столбец параметров

$$\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = M^{-1}F^{\mathrm{T}}Y == \begin{pmatrix} -2{,}6589 \\ 0{,}2579 \\ -0{,}0003 \end{pmatrix}.$$

Итак, $\widehat{y}(x) = -2{,}6589 + 0{,}2579x - 0{,}0003x^2$.

Проверим значимость модели регрессии на уровне $\alpha = 0{,}1$. Для этого составим таблицу (табл. 7.10), в которой $\widehat{y}_i = \widehat{y}(x_i)$ и $\overline{y}$ — выборочное среднее показателя y (среднее значение второго столбца табл. 7.9), равное

$$\overline{y} = \frac{1}{7}\Big(3{,}94 + 4{,}6 + 5{,}67 + 6{,}93 + 7{,}73 + 8{,}25 + 9{,}56\Big) = 6{,}67.$$

Таблица 7.10

x_i	y_i	$\widehat{y}_i$	$y_i - \widehat{y}_i$	$(y_i - \widehat{y}_i)^2$	$\widehat{y}_i - \overline{y}$	$(\widehat{y}_i - \overline{y})^2$
26	3,94	3,8343	0,11	0,0121	−2,84	8,0656
30	4,60	4,7958	−0,20	0,0400	−1,87	3,4969
34	5,67	5,7472	−0,08	0,0064	−0,92	0,8464
38	6,93	6,6886	0,24	0,0576	0,02	0,0004
42	7,73	7,6201	0,11	0,0121	0,95	0,9025
46	8,25	8,5415	−0,29	0,0841	1,87	3,4969
50	9,56	9,453	0,11	0,0121	2,78	7,7284

По данным составленной таблицы находим

$$Q_l = \sum_{i=1}^{7}(y_i - \widehat{y})^2 = 0{,}2244, \qquad Q_f = \sum_{i=1}^{7}(\widehat{y} - \overline{y})^2 = 24{,}5371,$$

$$f_{\text{в}} = \frac{24{,}5371 \cdot 4}{2 \cdot 0{,}2244} \approx 218{,}69.$$

По таблице квантилей распределения Фишера (см. табл. П.5) находим

$$f_{\text{кр}} = f_{1-\alpha/2}(m-1, n-m) = f_{0{,}95}(2, 4) = 6{,}94.$$

Из неравенства $f_{\text{в}} = 218{,}69 > f_{\text{кр}} = 6{,}94$, согласно критерию (7.24), приходим к заключению, что модель значима.

Пример 7.13. По данным наблюдений (табл. 7.11) найдем оценки параметров модели регрессии $y = \beta_0 + \beta_1 x + \beta_2 x^2$ и проверим адекватность этой модели на уровне значимости $\alpha = 0{,}01$.

Таблица 7.11

x_i	0	0	0	1	2	2	3	3	4	4
y_i	22,8	21,9	22,1	24,5	26,0	26,1	26,8	27,3	28,2	28,5
x_i	5	6	6	6	7	8	8	9	10	
y_i	28,9	30,0	30,3	29,8	30,4	31,4	31,5	31,8	33,1	

По данным из табл. 7.11 запишем матрицы

7.14. Результаты эксперимента представлены таблицей

Значение x	Значения y
1	1; 1; 2
2	1; 2; 2; 3; 3; 3; 4
3	3; 4; 4; 4; 5; 5
4	4; 5; 5; 6
5	5; 5; 6
6	5; 6

Полагая, что переменные y и x связаны линейной зависимостью, найдите значения оценок параметров.

Ответ: $\hat{y} = 0{,}932 + 0{,}906x$.

7.15. Для модели регрессии, построенной в примере 7.3, проверьте ее значимость на уровне значимости $\alpha = 0{,}05$ и значимость ее коэффициентов β_0 и β_1 на уровне значимости $\alpha = 0{,}1$.

Ответ: модель значима; оба коэффициента значимы.

7.16. Зависимость между переменными x и y имеет вид $y = \beta_0 + \beta_1 x + \beta_2 x^2$. По данным выборки

(0,07, 1,34); (0,31, 1,08); (0,61, 0,94); (0,99, 1,06);
(1,29, 1,25); (1,78, 2,01); (2,09, 2,60)

выполните следующее:

а) найдите значения оценок параметров модели регрессии;

б) проверьте значимость модели регрессии на уровне значимости $\alpha = 0{,}05$.

Ответ: а) $\hat{y} = 1{,}40 - 1{,}22x + 0{,}87x^2$; б) модель значима.

7.17. Зависимость между переменными x и y имеет вид $y = \beta_0 + \beta_1 x + \beta_2 x^2$. По данным выборки (табл. 7.14) выполните следующее:

а) найдите значения оценок параметров модели регрессии;

б) проверьте значимость модели регрессии на уровне значимости $\alpha = 0{,}01$.

Таблица 7.14

x_i	26	30	34	38	42	46	50
y_i	3,94	4,60	5,67	6,93	8,25	7,73	10,55

Ответ: а) $\hat{y} = 0{,}175 + 0{,}085x + 0{,}002x^2$; б) модель не является значимой.

7.18. Проведены равноточные измерения некоторой величины y через равные интервалы аргумента x (табл. 7.15). Считая, что зависимость между x и y имеет вид $y = \beta_0 + \beta_1 x + \beta_2 x^2$, выполните следующее:

а) найдите значения оценок параметров модели регрессии;

б) проверьте значимость модели на уровне значимости $\alpha = 0{,}01$;

в) проверьте значимость коэффициентов β_1 и β_2 на уровне значимости $\alpha = 0{,}01$.

Таблица 7.15

x_i	−3	−2	−1	0	1	2	3
y_i	−0,71	−0,01	0,51	0,82	0,88	0,81	0,49