Статистический анализ регрессионной модели

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.

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

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

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

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 $\hat{\varepsilon}_i$, $i = \overline{1, n}$ — 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, $i = 1, n$, 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 $\sum_{i=1}^{n} r_i = N$ — *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 $\sum_{i=1}^{n} (r_i - 1)$, а отношение

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

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 $\sum_{i=1}^{n}(r_i-1)$:

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

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 $\sum_{i=1} r_i = N = 13$, $n = 6$, $m = 2$,

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

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.

Проверка значимости параметров модели регрессии. Напомним, что регрессионную модель мы выбрали в виде (7.3), т.е. неизвестную функцию регрессии $f(x)$ ищем в виде

$$f_a(x) = \sum_{k=0}^{m-1} \beta_k \psi_k(x), \tag{7.23}$$

где некоторые из базисных функций $\psi_k(x)$ могли быть включены в модель регрессии ошибочно, т.е. на самом деле *отклик* Y от этих $\psi_k(x)$ не зависит и потому соответствующие коэффи-

циенты β_k должны быть равны нулю. Однако может оказаться, что полученные по формуле (7.7) значения МНК-оценок $\widehat{\beta}_k$ отличны от нуля, хотя обычно к нулю и близки.

Проверка значимости коэффициента β_k означает проверку гипотезы H_0: $\beta_k = 0$ против *альтернативной статистической гипотезы* H_1: $\beta_k \neq 0$. ***Коэффициент*** β_k считают ***значимым***, если верна гипотеза H_1.

В общем случае могут возникать более сложные гипотезы, например гипотеза H_0: $\beta_1 = -\beta_2 = \beta$, означающая, что $\beta_1 + \beta_2 = 0$. Такая гипотеза уместна, когда есть подозрение, что действует не каждый из *факторов* X_1 и X_2 по отдельности, а только их разность, т.е. вместо комбинации $\beta_1 X_1 + \beta_2 X_2$ в модель нужно включить выражение $\beta(X_1 - X_2)$.

Статистические гипотезы, которые включают утверждение о линейной комбинации параметров β_j, $j = 0, m-1$, называют ***линейными гипотезами***. Они обычно вытекают из знаний экспериментатора или его предположений относительно возможных моделей. Под проверкой значимости параметров модели регрессии в этом случае понимают проверку всех воз-можных линейных гипотез.

Мы ограничимся здесь проверкой линейных гипотез двух типов:

1) гипотезы H_0: $\beta_0 = \beta_1 = \ldots = \beta_{m-1} = 0$ против альтерна-тивной гипотезы H_1, согласно которой $\beta_k \neq 0$ хотя бы для одного номера k, $k = 0, m-1$;

2) гипотезы H_{0k}: $\beta_k = 0$ против альтернативной гипотезы H_{1k}: $\beta_k \neq 0$, рассматриваемых для некоторого фиксированного номера k, $k = 0, m-1$.

Если гипотеза H_0 верна, то ***модель регрессии*** называют ***незначимой***, т.е. условное математическое ожидание отклика M (Y) $x = \overline{y}(x) = \beta_0$ постоянно и не меняется с изменением x. В противном случае ***модель регрессии*** называют ***значимой***.

Гипотезы второго типа связаны с анализом конкретного коэффициента β_k. Если гипотеза H_{0k} принимается, то коэф-

фициент β_K незначим и может быть удален из модели.

Рассмотрим критерий проверки гипотез первого типа. Исходя из предположений о случайных величинах Y_i, $i = 1, n$, можно показать, что статистики

$$Q_l(\vec{Y}_n) = (Y - \hat{Y})^{\mathrm{T}}(Y - \hat{Y}) \text{ (}\textit{остаточная сумма квадратов}\text{)}$$

и $Q_f(\vec{Y}_n) = (\hat{Y} - I\overline{Y})^{\mathrm{T}}(\hat{Y} - I\overline{Y})$ являются независимыми случайными величинами. Здесь Y — *матрица отклика линейной регрессионной модели* (7.6), $\hat{Y}$ — *матрица МНК-оценок средних значений отклика* и $\overline{Y}$ — *выборочное среднее* отклика.

Раскрывая матричное представление статистик $Q_l(\vec{Y}_n)$ и $Q_f(\vec{Y}_n)$, заключаем, что

$$Q_l(\vec{Y}_n) = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2, \qquad Q_f(\vec{Y}_n) = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2.$$

Статистика $Q_l(\vec{Y}_n)/\sigma^2$ имеет χ^2-распределение с числом степеней свободы $n-m$, а статистика $Q_f(\vec{Y}_n)/\sigma^2$ — χ^2-распределение с числом степеней свободы $m-1$, если H_0 верна. Тогда статистика

$$F = \frac{Q_f(\vec{Y}_n)}{m-1}\,\frac{n-m}{Q_l(\vec{Y}_n)} \sim F(m-1, n-m),$$

т.е. имеет распределение Фишера со степенями свободы $m-1$ и $n-m$.

Статистика $Q_l(\vec{Y}_n)/(n-m)$ является несмещенной оценкой остаточной дисперсии (см. теорему 7.3), обусловленной как случайными ошибками измерений значений функции регрессии, так и неучтенными в регрессии факторами; статистика $Q_f(\vec{Y}_n)/(m-1)$ — несмещенная оценка дисперсии случайных ошибок при использовании функции регрессии (т.е. дисперсии случайных ошибок измерений значений функции регрессии). Поэтому статистика F может быть использована при проверке рассматриваемой гипотезы.

Таким образом, гипотеза H_0: $\beta_1 = \ldots = \beta_{m-1} = 0$ отклоняется на уровне значимости α (а следовательно, регрессия признается значимой), если вычисленное значение статистики F

$$f_{\text{в}} > f_{\text{кр}} = f_{1-\alpha}(m-1, n-m). \qquad (7.24)$$

Замечание 7.4. Полезной характеристикой линейной регрессионной модели является *коэффициент детерминации* R^2 (или квадрат *множественного коэффициента корреляции*).

Оценка

$$\widehat{R}^2 = 1 - \frac{Q_l}{Q_y} = \frac{Q_f}{Q_y}$$

коэффициента детерминации показывает, какая доля в сумме квадратов отклонений отклика Y от его среднего значения, т.е. в $Q_Y(\vec{Y}_n) = (Y - I\overline{Y})^{\mathrm{T}}(Y - I\overline{Y})$, обусловлена регрессией (т.е. показывает, насколько значимы параметры модели регрессии). Величина $\widehat{R}(\vec{Y}_n)$ является оценкой коэффициента корреляции (мерой линейной связи) между случайными величинами Y и $\widehat{Y}(\vec{x})$. #

Перейдем к проверке линейных гипотез второго типа. Эти гипотезы проверяют после того, как обоснована значимость регрессии. Такая проверка позволяет более детально проанализировать структуру модели регрессии на уровне отдельных коэффициентов. Ясно, что возможна ситуация, когда вектор параметров $\vec{\beta}$ модели регрессии является значимым, в то время как отдельные коэффициенты модели незначимы (и, следовательно, их надо принять равными нулю).

Проверку любой из m гипотез H_{0k}, $0 \leqslant k \leqslant m-1$, против гипотезы H_{1k} проводят по *критерию Стьюдента.*

Напомним, что МНК-оценка $\widehat{\beta}_k(\vec{Y}_n)$ параметра β_k линейно зависит от матрицы отклика Y. Следовательно, в силу (7.22) эта оценка имеет нормальный закон распределения с математическим ожиданием β_k (ибо оценка $\widehat{\beta}_k(\vec{Y}_n)$ несмещенная) и дисперсией $\sigma_y^2 c_{kk}$ (см. следствие 7.1). Здесь c_{kk} — k-й диагональный элемент *дисперсионной матрицы Фишера* $C = (F^{\mathrm{T}}F)^{-1}$. Поэтому

$$Z = \frac{\widehat{\beta}_k(\vec{Y}_n) - \beta_k}{\sigma\sqrt{c_{kk}}} \sim N(0,1).$$

В то же время

$$V = \frac{Q_l(\vec{Y}_n)}{\sigma} = \frac{(n-m)S_y^2(\vec{Y}_n)}{\sigma^2} \sim \chi^2(n-m).$$

Таким образом, если гипотеза H_{0k}: $\beta_k = 0$ верна, то

$$T_k = \frac{\widehat{\beta}_k(\vec{Y}_n)}{S_y\sqrt{c_{kk}}} \sim S(n-m), \quad k = \overline{0, m-1}. \tag{7.25}$$

Если модуль вычисленного значения t_k статистики T_k превысит критический уровень $t_k^{\text{кр}} = t_{1-\alpha/2}(n-m)$, то гипотезу H_{0k} следует отклонить на уровне значимости α и признать коэффициент β_k значимым.

Замечание 7.5. Проверку значимости коэффициента β_k модели регрессии (7.23) можно проводить также с помощью *доверительного интервала* $J_\gamma(\beta_k) = (\underline{\beta}_k(\vec{Y}_n), \overline{\beta}_k(\vec{Y}_n))$, значения границ которого в силу (7.25) имеют вид (см. **3.3**)

$$\widehat{\beta}_k \pm t_{1-\alpha/2}(n-m)S_y\sqrt{c_{kk}}. \tag{7.26}$$

Гипотеза H_{0k}: $\beta_k = 0$ принимается, если интервал с границами (7.26) накрывает нуль, и отклоняется в противном случае.

Замечание 7.6. Для простой линейной регрессии

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

(см. пример 7.3) число параметров $m = 2$, а дисперсионная матрица Фишера имеет вид

$$\Sigma\widehat{\beta} = \begin{pmatrix} c_{00} & c_{01} \\ c_{10} & c_{11} \end{pmatrix},$$

где

$$c_{00} = \frac{\sum\limits_{i=1}^{n} x_i^2}{n\sum\limits_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum\limits_{i=1}^{n} x_i^2}{nQ_x}, \qquad c_{11} = \frac{1}{Q_x}.$$

Поэтому из (7.25) следует, что

$$T_0 = \frac{\widehat{\beta}_0(\vec{Y}_n)}{S_y}\sqrt{\frac{nQ_x}{\sum\limits_{i=1}^{n} x_i^2}} \sim S(n-2), \qquad T_1 = \frac{\widehat{\beta}_1(\vec{Y}_n)}{S_y}\sqrt{Q_x} \sim S(n-2),$$

а значения (7.26) границ доверительных интервалов для параметров β_0 и β_1 принимают соответственно вид

$$\widehat{\beta}_0 \pm t_{1-\alpha/2}(n-2)S_y\sqrt{\frac{\sum\limits_{i=1}^{n} x_i^2}{nQ_x}}, \qquad \widehat{\beta}_1 \pm t_{1-\alpha/2}(n-2)S_y\sqrt{\frac{1}{Q_x}}.$$

Пример 7.6. Результаты y_i, $i = 1, n$, наблюдений, проведенных над откликом Y при значениях x_i фактора X, представлены в табл. 7.4.

Таблица 7.4

x_i	0	1	2	3	4	5	6	7	8	9	10
y_i	8,98	8,82	9,09	11,94	24,63	14,06	14,00	24,93	33,22	15,7	35,92

Рассмотрим в качестве *допустимой модели регрессии* функцию

$$f_a(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

и найдем МНК-оценки неизвестных параметров модели регрессии: $\widehat{\beta}_0 = 6{,}92$; $\widehat{\beta}_1 = 2{,}27$; $\widehat{\beta}_2 = 0{,}08$. Таким образом, имеем

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

Есть основания предполагать, что $\beta_2 = 0$. Для проверки гипотезы H_0: $\beta_2 = 0$ (значимости коэффициента β_2) против альтернативной гипотезы H_1: $\beta_2 \neq 0$ находим значение $t_2 = 0{,}20$ статистики T_2 (7.25).

Воспользовавшись таблицей квантилей распределения Стьюдента (см. табл. П.4), на уровне значимости $\alpha = 0{,}1$ находим $t_{\text{кр}} = t_{1-\alpha/2}(n-m) = t_{0{,}95}(8) = 2{,}31$. Коэффициент β_2 незначим, так как $t_2 = 0{,}20 < t_{\text{кр}} = 2{,}31$.

Значение оценки коэффициента детерминации

$$\widehat{R}^2 = 1 - \frac{Q_l}{Q_y} = 1 - \frac{1060{,}51}{2214{,}24} \approx 0{,}52.$$

Полученный результат указывает на 52 %-ный разброс результатов наблюдений относительно горизонтальной прямой $\overline{y} = 18{,}29$.

Анализ точности результатов, полученных с использованием регрессионной модели. Если модель регрессии прошла проверку на значимость, то ее можно использовать для решения различных практических задач. Основными из них являются:

– определение значения отклика Y в той части факторного пространства, где эксперимент не проводился, т.е. либо интерполяция, либо экстраполяция (прогнозирование) отклика;

– определение экстремальных условий протекания процесса, модель которого построена, т.е. отыскание такой точки $x^* = (x_1^*, \ldots, x_n^*)$, в которой $\widehat{y}(x)$ имеет экстремум; эту задачу решают методами математического анализа [V].

В обоих случаях с помощью построенной модели

$$\widehat{Y}(x) = \sum_{k=0}^{m-1} \widehat{\beta}_k \psi_k(x)$$

требуется оценить точность предсказания в рассматриваемой точке $x = x_0$ либо среднего значения отклика $\mathrm{M}(Y \mid x) = \overline{y}(x)$, либо ожидаемого значения отклика $Y = Y_0$.

Для решения первой задачи нужно для величины $\overline{y}(x)$ построить доверительный интервал J_γ с заданным уровнем доверия γ, а для решения второй — так называемый прогнозирующий интервал $\widetilde{J}_\gamma$, в который случайная величина Y при $x = x^0$ попадает с заданной доверительной вероятностью γ.

При нахождении доверительного интервала J_γ важно то, что МНК-оценки $\widehat{\beta}_k(\vec{Y}_n)$ имеют нормальный закон распределения, а следовательно, оценка $\widehat{Y}(x)$ также распределена по нормальному закону со средним $\mathrm{M}\widehat{Y}(x) = \overline{y}(x)$ и дисперсией

$$\mathbf{D}\,\hat{Y}(x) = \sigma^2 \psi^{\mathrm{T}}(x) C \psi(x).$$

Значит,

$$Z = \frac{\hat{Y}(x) - \overline{y}(x)}{\sigma\sqrt{\psi^{\mathrm{T}}(x) C \psi(x)}} \sim N(0,1).$$

С другой стороны, несмещенная оценка дисперсии отклика σ^2, определяемая по формуле (7.17), не зависит от Z и

$$V = \frac{Q_l}{\sigma^2} = \frac{(n-m) S_y^2(\vec{Y}_n)}{\sigma^2} \sim \chi^2(n-m),$$

т.е. имеет χ^2-распределение с числом степеней свободы $n-m$.

Отсюда следует, что статистика $Z/\sqrt{V/(n-m)}$ распределена по закону Стьюдента с числом степеней свободы $n-m$ (см. **Д.3.1**):

$$\frac{Z}{\sqrt{V/(n-m)}} = \frac{\hat{Y}(x) - \overline{y}(x)}{S_y(\vec{Y}_n)\sqrt{\psi^{\mathrm{T}}(x) C \psi(x)}} \sim S(n-m).$$

Таким образом, с вероятностью $\gamma = 1-\alpha$ выполняется неравенство

$$\left| \frac{\hat{Y}(x) - \overline{y}(x)}{S_y(\vec{Y}_n)\sqrt{\psi^{\mathrm{T}}(x) C \psi(x)}} \right| < t_{1-\alpha/2}(n-m),$$

где $t_{1-\alpha/2}(n-m)$ — квантиль уровня $1-\alpha/2$ распределения Стьюдента с числом степеней свободы $n-m$.

Это равенство дает границы доверительного интервала с уровнем доверия γ для среднего значения отклика $\overline{y}(x)$ в произвольной точке x факторного пространства в виде

$$\hat{y}(x) \pm t_{1-\alpha/2}(n-m) S_y(\vec{Y}_n)\sqrt{\psi^{\mathrm{T}}(x) C \psi(x)}, \qquad (7.27)$$

где, напомним, $C = (F^{\mathrm{T}}F)^{-1}$.

В частном случае простой линейной регрессии

$$\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

дисперсию $\hat{Y}(x)$ вычисляют по формуле

$$\mathbf{D}\hat{Y}(x) = \sigma^2\left(\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}\right),$$

и формула (7.27) принимает следующий вид:

$$\hat{y}(x) \pm t_{1-\alpha/2}(n-m)S_y(\vec{Y}_n)\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}}. \qquad (7.28)$$

Из выражения (7.28) видно, что наиболее узким интервал J_γ будет в точке $x = \overline{x}$, и по мере удаления x от $\overline{x}$ точность уменьшается (рис. 7.6).

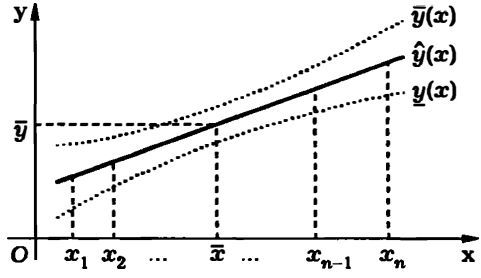

Рис. 7.6

Для отыскания прогнозирующего интервала $\widetilde{J}_\gamma$ с уровнем доверия γ используют тот факт, что разность между откликом Y и оценкой его среднего значения $\hat{Y}(x)$ в любой точке x имеет нормальный закон распределения со средним значением $\mathbf{M}(Y - \hat{Y}(x)) = 0$ и дисперсией (в силу независимости Y и $\hat{Y}(x)$)

$$\mathbf{D}(Y - \hat{Y}(x)) = \mathbf{D}Y + \mathbf{D}\hat{Y}(x) = \sigma^2 + \mathbf{D}\hat{Y}(x) = \sigma^2(1 + \psi^{\mathrm{T}}(x)C\psi(x)),$$

т.е. к дисперсии $\widehat{Y}(x)$ добавляется дисперсия отклика Y.

Повторяя предыдущие рассуждения при построении доверительного интервала, вместо (7.27) получаем окончательный результат в виде

$$\widehat{y}(x) \pm t_{1-\alpha/2}(n-m) S_y(\vec{Y}_n) \sqrt{1+\psi^T(x) C \psi(x)}. \qquad (7.29)$$

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, $i = 1, n$ (см. (7.4)).

Пусть имеем две допустимые модели регрессии

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

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*

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

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 $\sum\limits_{k=0}^{m_1-1} \beta_k \psi(\vec{x})$), если значение $f_в$ статистики F, рассчитанное по результатам наблюдений, не превышает $f_{кр} = f_{1-\alpha}(m_2 - m_1, n - m_1 - m_2)$.

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

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

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

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

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

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 $\widehat{y}_1(x) = 6{,}91 + 2{,}28x$.