

ЛЕКЦИЯ 7. МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

РЕГРЕССИОННЫЙ АНАЛИЗ. ПОСТРОЕНИЕ ЛИНЕЙНОЙ И КВАДРАТИЧНОЙ РЕГРЕССИОННЫХ МОДЕЛЕЙ

1. Оценка коэффициентов регрессии.

Важной задачей математической статистики является получение функциональной зависимости одной величины (y) от другой (x) по результатам эксперимента. Будем считать, что функциональная зависимость между величинами, называемая в дальнейшем *моделью*, известна из предварительных сведений с точностью до параметров $\beta_1, \beta_2, \dots, \beta_m$ и имеет вид

$$y = f(x, \beta_1, \beta_2, \dots, \beta_m). \quad (1.58)$$

Для отыскания неизвестных параметров проведено n наблюдений $(x_i, Y_i) \quad i=1,2,\dots,n$.

Но так как результаты наблюдений не свободны от погрешностей измерений, которые мы будем рассматривать как случайные ошибки, то по ним нельзя точно найти искомые параметры. Поэтому приходится ставить задачу об отыскании не значений параметров, а их оценок по результатам эксперимента.

Будем предполагать, что значения аргументов x_i известны точно, а значения функции Y_i – взаимно независимые случайные величины, включающие случайные ошибки Z_i , то есть $Y_i = f(x_i, \beta_1, \beta_2, \dots, \beta_m) + Z_i$, где

$$M(Z_i) = 0; \quad D(Z_i) = D(Y_i) = \sigma^2. \quad (1.59)$$

Здесь мы предполагаем, что измерения *равноточны*. Для оценок параметров $\beta_1, \beta_2, \dots, \beta_m$ используется *метод наименьших квадратов*. В качестве оценок этих параметров принимают значения $\tilde{\beta}_1, \tilde{\beta}_2, \dots, \tilde{\beta}_m$, при которых имеет минимум функция (МНК-оценки)

$$Q(\beta_1, \beta_2, \dots, \beta_m) = \sum_{i=1}^n (Y_i - f(x_i, \beta_1, \beta_2, \dots, \beta_m))^2. \quad (1.60)$$

Уравнение (1.58) называют *уравнением регрессии*, а отыскание оценок параметров и исследование получаемых моделей – *регрессионным анализом*.

Будем рассматривать уравнения регрессии, линейные относительно оцениваемых параметров $\beta_1, \beta_2, \dots, \beta_m$:

$$y = f(x, \beta_1, \beta_2, \dots, \beta_m) = \beta_1 \varphi_1(x) + \beta_2 \varphi_2(x) + \dots + \beta_m \varphi_m(x). \quad (1.61)$$

Функции $\varphi_1(x), \varphi_2(x), \dots, \varphi_m(x)$ называют *базисными функциями*, их рассматривают на множестве точек $\{x_1, x_2, \dots, x_n\}$, где n – число экспериментов. Функция Q (1.60) в этом случае запишется в виде:

$$Q(\beta_1, \beta_2, \dots, \beta_m) = \sum_{i=1}^n (Y_i - \beta_1 \varphi_1(x_i) - \beta_2 \varphi_2(x_i) - \dots - \beta_m \varphi_m(x_i))^2. \quad (1.62)$$

Для нахождения минимума найдем частные производные функции $Q(\beta_1, \beta_2, \dots, \beta_m)$ по переменным $\beta_1, \beta_2, \dots, \beta_m$ и приравняем их к нулю (необходимые условия минимума функции). Получим систему уравнений:

$$\left\{ \begin{array}{l} \frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^n (Y_i - \beta_1 \varphi_1(x_i) - \beta_2 \varphi_2(x_i) - \dots - \beta_m \varphi_m(x_i)) \varphi_1(x_i) = 0 \\ \frac{\partial Q}{\partial \beta_2} = -2 \sum_{i=1}^n (Y_i - \beta_1 \varphi_1(x_i) - \beta_2 \varphi_2(x_i) - \dots - \beta_m \varphi_m(x_i)) \varphi_2(x_i) = 0 \\ \dots \\ \frac{\partial Q}{\partial \beta_m} = -2 \sum_{i=1}^n (Y_i - \beta_1 \varphi_1(x_i) - \beta_2 \varphi_2(x_i) - \dots - \beta_m \varphi_m(x_i)) \varphi_m(x_i) = 0 \end{array} \right.$$

которую после преобразований можно записать в виде:

$$\left\{ \begin{array}{l} \beta_1 \sum_{i=1}^n \varphi_1^2(x_i) + \beta_2 \sum_{i=1}^n \varphi_1(x_i) \varphi_2(x_i) + \dots + \beta_m \sum_{i=1}^n \varphi_1(x_i) \varphi_m(x_i) = \sum_{i=1}^n Y_i \varphi_1(x_i) \\ \beta_1 \sum_{i=1}^n \varphi_1(x_i) \varphi_2(x_i) + \beta_2 \sum_{i=1}^n \varphi_2^2(x_i) + \dots + \beta_m \sum_{i=1}^n \varphi_2(x_i) \varphi_m(x_i) = \sum_{i=1}^n Y_i \varphi_2(x_i) \\ \dots \\ \beta_1 \sum_{i=1}^n \varphi_1(x_i) \varphi_m(x_i) + \beta_2 \sum_{i=1}^n \varphi_2(x_i) \varphi_m(x_i) + \dots + \beta_m \sum_{i=1}^n \varphi_m^2(x_i) = \sum_{i=1}^n Y_i \varphi_m(x_i) \end{array} \right. \quad (1.63)$$

Следовательно, оценки параметров $\tilde{\beta}_1, \tilde{\beta}_2, \dots, \tilde{\beta}_m$ являются решениями линейной алгебраической системы m уравнений (1.63).

Введем обозначения: $\sum_{i=1}^n \varphi_j(x_i) \varphi_k(x_i) = (\varphi_j, \varphi_k)$; $\sum_{i=1}^n Y_i \varphi_k(x_i) = (Y, \varphi_k)$, тогда система

(1.63) запишется в виде:

$$\left\{ \begin{array}{l} \beta_1 (\varphi_1, \varphi_1) + \beta_2 (\varphi_1, \varphi_2) + \dots + \beta_m (\varphi_1, \varphi_m) = (Y, \varphi_1) \\ \beta_1 (\varphi_2, \varphi_1) + \beta_2 (\varphi_2, \varphi_2) + \dots + \beta_m (\varphi_2, \varphi_m) = (Y, \varphi_2) \\ \dots \\ \beta_1 (\varphi_m, \varphi_1) + \beta_2 (\varphi_m, \varphi_2) + \dots + \beta_m (\varphi_m, \varphi_m) = (Y, \varphi_m) \end{array} \right.$$

С использованием следующих матричных обозначений:

$$A = \begin{pmatrix} (\varphi_1, \varphi_1) & (\varphi_1, \varphi_2) & \dots & (\varphi_1, \varphi_m) \\ (\varphi_2, \varphi_1) & (\varphi_2, \varphi_2) & \dots & (\varphi_2, \varphi_m) \\ \dots & \dots & \dots & \dots \\ (\varphi_m, \varphi_1) & (\varphi_m, \varphi_2) & \dots & (\varphi_m, \varphi_m) \end{pmatrix} \quad \text{- матрица коэффициентов при неизвестных,}$$

$$Y = \begin{pmatrix} (Y, \varphi_1) \\ (Y, \varphi_2) \\ \dots \\ (Y, \varphi_m) \end{pmatrix} \quad \text{- вектор правых частей,} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_m \end{pmatrix} \quad \text{- вектор параметров,}$$

Система (1.63) принимает вид

$$A\beta = Y. \quad (1.64)$$

При условии, что A – невырожденная матрица, решение системы (4.6) можно записать в виде

$$\tilde{\beta} = A^{-1}Y, \quad (1.65)$$

где $\tilde{\beta} = \begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \\ \dots \\ \tilde{\beta}_m \end{pmatrix}$ – вектор МНК-оценок параметров регрессионной модели (1.61).

Оценки параметров линейной регрессии, получаемые по методу наименьших квадратов, имеют следующие свойства:

1. Они являются линейными функциями результатов наблюдений $Y_i, i=1,2,\dots,n$, и несмещенными оценками параметров, т.е. $M(\tilde{\beta}_j) = \beta_j, j=1,2,\dots,m$.
2. Они имеют минимальные дисперсии в классе несмещенных оценок, являющихся линейными функциями результатов наблюдений.

2. Проверка гипотезы об адекватности регрессионной модели.

Регрессионная модель называется *адекватной*, если предсказанные по ней значения переменной Y согласуются с результатами эксперимента. Если модель адекватна, то отклонения результатов эксперимента от полученной функции регрессии

$\Delta Y_i = Y_i - \tilde{Y}(x_i)$ являются реализациями случайных ошибок эксперимента Z_i , которые, в силу предположений (1.58), должны быть независимыми нормально распределенными случайными величинами с нулевыми средними и одинаковыми дисперсиями σ^2 .

Проверка выполнения этих предположений осуществляется статистическими методами и лежит в основе оценки адекватности модели регрессии.

Для проверки адекватности регрессионной модели вычисляют остаточную дисперсию (так называемую дисперсию адекватности) по формуле

$$S_{\text{ад}}^2 = \frac{\sum_{i=1}^n (\Delta Y_i)^2}{k_{\text{ад}}}; \quad k_{\text{ад}} = n - m, \quad (1.66)$$

где ΔY_i – отклонения средних Y_i от проверяемой модели регрессии; $k_{\text{ад}}$ – число степеней свободы дисперсии адекватности; n – число точек, в которых проводился эксперимент; m – число оцениваемых параметров β_j в проверяемой модели.

Если истинная функция регрессии имеет тот же вид, что и рассматриваемая модель (например, так же, как и модель, представляет собой квадратичную функцию), то дисперсия адекватности служит несмещенной оценкой истинной дисперсии эксперимента и ее можно сравнивать с другими подобными оценками. В частности, может быть проведена независимая серия измерений для получения оценки дисперсии эксперимента $S_{\text{эксп}}^2$. В этом случае $S_{\text{эксп}}^2$ оценивает дисперсию эксперимента $D_{\text{эксп}}$, $S_{\text{ад}}^2$ характеризует степень отклонения экспериментальных точек от регрессионной модели, т.е. оценивает некую дисперсию адекватности $D_{\text{ад}}$. Проверка адекватности модели заключается в проверке гипотезы $H_0: D_{\text{ад}} = D_{\text{эксп}}$ при альтернативной гипотезе $H_1: D_{\text{ад}} > D_{\text{эксп}}$ (если модель неадекватна, отклонения экспериментальных точек от модели будут больше погрешностей эксперимента). Таким образом, задача сводится к проверке гипотезы о равенстве дисперсий, которая решается с помощью критерия Фишера. Вычисляем отношение

$$F = S_{\text{ад}}^2 / S_{\text{эксп}}^2. \quad (1.67)$$

Если при заданном уровне значимости α отношение F окажется меньше квантили $F_{1-\alpha}(k_1, k_2)$, где $k_1 = k_{\text{ад}}$, $k_2 = k_{\text{эксп}}$, то рассматриваемая модель не противоречит результатам эксперимента и принимается; в противоположном случае модель отвергается с уровнем значимости α , как противоречащая результатам эксперимента.

3. Построение доверительных интервалов для коэффициентов регрессии.

Оценки параметров регрессии $\tilde{\beta}_j$, определяемые формулами (1.65), являются точечными оценками истинных значений параметров β_j . Если результаты экспериментов независимы и подчиняются нормальному закону распределения с дисперсией σ^2 , то

доверительный интервал с доверительной вероятностью $P = 1 - \alpha$ для каждого параметра $\tilde{\beta}_j$ можно определить неравенством

$$|B_j - \tilde{B}_j| < \varepsilon_j \quad \text{или} \quad B_j = \tilde{B}_j \pm \varepsilon_j,$$

где $\varepsilon_j = t_{1-\alpha/2}(k)S\sqrt{a_{jj}}$; $(j = 1, 2, \dots, m)$, (1.68)

здесь S^2 – несмещенная оценка дисперсии σ^2 с числом степеней свободы k ; $t_{1-\alpha/2}(k)$ – квантиль распределения Стьюдента; a_{jj} – диагональный элемент матрицы A^{-1} , $\alpha = 1 - P$, P – доверительная вероятность.

Как было указано выше, если рассматриваемая модель регрессии адекватна, то дисперсия адекватности служит несмещенной оценкой истинной дисперсии эксперимента. Следовательно, в случае адекватности модели в формулу (1.69) в качестве оценки среднего квадратического отклонения S можно подставить корень из дисперсии

$$\text{адекватности } S_{\text{ад}} = \sqrt{S_{\text{ад}}^2}.$$

В построенной регрессионной модели (1.61) некоторые коэффициенты могут быть незначимы, т.е. может выполняться гипотеза $H_0: \beta_j = 0$. Для проверки этой гипотезы можно найти доверительный интервал для коэффициента β_j с уровнем значимости α . Если этот интервал «накрывает» значение $\beta_j = 0$, гипотеза H_0 принимается и коэффициент β_j признается незначимым, в противоположном случае коэффициент β_j значим.