АДЕКВАТНОСТЬ МАТЕМАТИЧЕСКОЙ МОДЕЛИ РЕГРЕССИИ

Слайд 2

Проверка статистических гипотез

Поверка гипотез относительно коэффициентов регрессии

В регрессионном анализе проверяются две нулевые гипотезы относительно коэффициентов уравнения.

1. Гипотеза равенства истинного коэффициента β_k значению β_{k0}

H_0: $\beta_k = \beta_{k0}$, против альтернативной гипотезы H_1: $\beta_k \neq \beta_{k0}$.

2. Гипотеза одновременного равенства нулю всех, кроме β_0, регрессионных коэффициентов

$$H_0 : \beta_1 = \mathrm{K} = \beta_q = 0.$$

Проверяем конкуренцию двух моделей

$$\mathbf{Y} = \sum_k \beta_k \mathbf{x}_k + \boldsymbol{\varepsilon} \text{ и } \mathbf{Y} = \beta_0 + \boldsymbol{\varepsilon}$$

или, что тоже самое, проверяем эффект от введения переменных $\mathbf{x}_1, \ldots, \mathbf{x}_q$ в модель регрессии.

Первая гипотеза позволяет проверить значимость отдельно взятого коэффициента. Если $\hat{\beta}_k : N_n(\beta_k, \sigma_\varepsilon^2 \mathbf{C}^{-2})$, то статистика

$$\frac{\hat{\beta}_k - \beta_{k0}}{S_{\hat{\beta}_k}} = \frac{\hat{\beta}_k - \beta_{k0}}{S_e \sqrt{\mathbf{C}_{kk}^{-1}}} = t$$

распределена по Стьюденту с числом с.с. $\nu = n - q - 1$, а квадрат ее имеет F-распределение (*распределение Снедокора-Фишера*) с числом с.с. $\nu_1 = 1$, $\nu_2 = n - q - 1$:

Слайд 3

$$F = \frac{(\hat{\beta}_k - \beta_{0k})^2}{S_e^2 (C^{-1})_{kk}} \tag{1}$$

Если окажется, что вычисленное значение F связано с табличным для заданного уровня значимости α неравенством

$$F > F(\alpha, \nu_1, \nu_2), \ \nu_1 = 1, \ \nu_2 = n - q - 1,$$

то гипотеза $\beta_k = \beta_{k0}$ отвергается. Обычно значения β_{k0} неизвестны, поэтому проверяют гипотезу H_0: $\beta_k = 0$. В этом случае получаем

$$F = \frac{\hat{\beta}_k^{\,2}}{S_e^{\,2} (\mathbf{C}^{-1})_{kk}} \tag{2}$$

с числом с.с. $\nu_1 = 1$, $\nu_2 = n - q - 1$. Проверка этой гипотезы имеет важное значение, так как позволяет ответить на вопрос: можно ли считать $\beta_k = 0$? Если ответ положительный, то k-й регрессор можно удалить из рассматриваемой модели.

Слайд 4

Для проверки второй гипотезы сначала выражение (1) перепишем в виде взвешенного расстояния между $\hat{\beta}_k = 0$ и β_{k0}:

$$d(\hat{\beta}_k, \beta_{k0}) = (\hat{\beta}_k - \beta_{k0})' (S_{\beta_K}^2)^{-1} (\hat{\beta}_k - \beta_{k0}) = F(\nu_1, \nu_2).$$

Рассмотрим случай с центрированными данными

$\tilde{\mathbf{x}}_j = \mathbf{x}_j - \bar{x}_j$, $\tilde{\mathbf{Y}} = \mathbf{Y} - \bar{Y}$.

Для центрированных данных оценка вектора коэффициентов $\hat{\boldsymbol{\beta}}^* = (\hat{\beta}_1, \hat{\beta}_2, \text{K}, \hat{\beta}_q)'$ не содержит свободного члена $\hat{\beta}_0$, Оценку $\hat{\beta}_0$ находим по формуле:

$\hat{\beta}_0 = \bar{Y} - \sum_{j=1}^{q} \hat{\beta}_j \bar{x}_j$.

Тогда

$d(\hat{\boldsymbol{\beta}}^*, \mathbf{0}) = (\hat{\boldsymbol{\beta}}^*)'(\hat{\text{cov}}(\hat{\boldsymbol{\beta}}^*))^{-1}\hat{\boldsymbol{\beta}}^* = F_{yp} \cdot q$,

где $\hat{\text{cov}}(\hat{\boldsymbol{\beta}}^*) = \mathbf{C}^{-1}S_e^2 = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}S_e^2$, а $d(\hat{\boldsymbol{\beta}}', \mathbf{0})$ – расстояние Махаланобиса.

Из этого уравнения находим статистику $F_{\text{ур}}$

$$F_{\text{ур}} = \frac{(\hat{\boldsymbol{\beta}}^*)'\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}^*}{S_e^2 q} = \frac{\hat{\tilde{\mathbf{Y}}}'\hat{\tilde{\mathbf{Y}}}}{S_e^2 q} = \frac{SSR/q}{SSE/(n-q-1)} \qquad (3)$$

где $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y}_i)^2$ – сумма квадратов, объясненная уравнением регрессии (*Sum of Squares about Regression*), $SSE = \sum_{i=1}^{n} e_i^2$ – сумма квадратов остатков.

Слайд 5

Формула (3) определяет отношение дисперсии оценки модели к дисперсии остатка $F_{yp} = S_\theta^2/S_e^2$. Статистика $F_{\text{ур}}$ имеет *F*-распределение с числом с.с. $\nu_1 = q$, $\nu_2 = n - q - 1$. Если $F_{yp} \geq F_{кр}(\alpha, \nu_1, \nu_2)$, то нулевая гипотеза отвергается. Всякая сумма квадратов связана с числом с.с. Например, для *SSE* число с.с. равно числу опытов *n* минус $(q + 1)$ коэффициентов регрессии.

Используя формулу (3) можем построить таблицу *дисперсионного анализа* (*ANOVA*) (табл. 1). «Средний квадрат» получается при делении каждой суммы квадратов на соответствующее ей число с.с.

Таблица дисперсионного анализа Таблица 1.

Источник дисперсии	Сумма квадратов	ч.с.с.	Средний квадрат	F-отношение
Модели	*SSR*	$\nu = q$	$MSE = \frac{SSR}{q}$	$F = \frac{SSR/q}{SSE/(n-q-1)}$
Остатки	*SSE*	$\nu = n - q - 1$	$MSE = S_e^2 = \frac{SSE}{n-q-1}$	–
Полная	*SST*	$\nu = n - 1$	–	–

Слайд 6

Адекватность модели

Оценка постулируемой модели $\mathbf{M}[\mathbf{Y}/\mathbf{x}] = \beta_0 + \beta_1\mathbf{x}_1 + \text{K} + \beta_k\mathbf{x}_q$, найденная по экспериментальным данным, равна

$\hat{\mathbf{Y}} = \hat{\beta}_0 + \hat{\beta}_1\mathbf{x}_1 + \text{K} + \hat{\beta}_k\mathbf{x}_q$.

Адекватность модели означает, что постулируемая модель не противоречит наблюдениям. Она оценивается с помощью множественного *коэффициента детерминации* (*coefficient of multiple determination*), который равен квадрату коэффициента *множественной корреляции R* между **Y** и $\hat{\mathbf{Y}}$

$$R = \frac{(\mathbf{Y} - \overline{\mathbf{Y}})'(\hat{\mathbf{Y}} - \overline{\mathbf{Y}})}{[(\mathbf{Y} - \overline{\mathbf{Y}})'(\mathbf{Y} - \overline{\mathbf{Y}})(\hat{\mathbf{Y}} - \overline{\mathbf{Y}})'(\hat{\mathbf{Y}} - \overline{\mathbf{Y}})]^{1/2}}.$$

Введем обозначения для центрированных данных: $\tilde{\mathbf{Y}} = \mathbf{Y} - \overline{\mathbf{Y}}$ и $\tilde{\hat{\mathbf{Y}}} = \hat{\mathbf{Y}} - \overline{\mathbf{Y}}$. Тогда коэффициент детерминации запишется в виде

$$R^2 = \frac{(\tilde{\mathbf{Y}}'\tilde{\hat{\mathbf{Y}}})^2}{(\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}})(\tilde{\hat{\mathbf{Y}}}'\tilde{\hat{\mathbf{Y}}})} \tag{4}$$

Сделаем следующие преобразования:

$$\tilde{\mathbf{Y}}'\tilde{\hat{\mathbf{Y}}} = (\tilde{\mathbf{Y}} - \tilde{\hat{\mathbf{Y}}} + \tilde{\hat{\mathbf{Y}}})'\tilde{\hat{\mathbf{Y}}} = (\tilde{\mathbf{Y}} - \tilde{\hat{\mathbf{Y}}})\tilde{\hat{\mathbf{Y}}} + \tilde{\hat{\mathbf{Y}}}'\tilde{\hat{\mathbf{Y}}} = \tilde{\hat{\mathbf{Y}}}'\tilde{\hat{\mathbf{Y}}} = SSR,$$

где $(\tilde{\mathbf{Y}} - \tilde{\hat{\mathbf{Y}}})\tilde{\hat{\mathbf{Y}}} = 0$ в силу ортогональности $(\tilde{\mathbf{Y}} - \tilde{\hat{\mathbf{Y}}})$ и $\tilde{\hat{\mathbf{Y}}}$;

$$\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}} = (\tilde{\mathbf{Y}} - \tilde{\hat{\mathbf{Y}}} + \tilde{\hat{\mathbf{Y}}})'(\tilde{\mathbf{Y}} - \tilde{\hat{\mathbf{Y}}} + \tilde{\hat{\mathbf{Y}}}) = (\tilde{\mathbf{Y}} - \tilde{\hat{\mathbf{Y}}})'(\tilde{\mathbf{Y}} - \tilde{\hat{\mathbf{Y}}}) + \tilde{\hat{\mathbf{Y}}}'\tilde{\hat{\mathbf{Y}}} =$$

$$= \mathbf{e}'\mathbf{e} + \tilde{\hat{\mathbf{Y}}}'\tilde{\mathbf{Y}} = SSE + SSR = SST.$$

Здесь $SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$ – *полная сумма квадратов* (*Total Sum of Squares*). В результате получаем

$$F_{\text{ур}} = \frac{(\hat{\boldsymbol{\beta}}^*)'\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}^*}{S_{\hat{e}}^2 q} = \frac{\tilde{\hat{\mathbf{Y}}}'\tilde{\hat{\mathbf{Y}}}}{S_{\hat{e}}^2 q} = \frac{SSR/q}{SSE/(n-q-1)} \tag{5}$$

Отсюда видно, что адекватность модели определяется отношением доли дисперсии, объясненной уравнением регрессии вариации откликов *SSR* к общей вариации *SST*. Если окажется, что величина R^2 не меньше табличного $R^2(\alpha, \nu_1, \nu_2)$, где $\nu_1 = q$, $\nu_2 = n - q - 1$, то постулируемая модель адекватна. Так как R есть мера взаимосвязи, то значение коэффициента детерминации всегда находится между нулем и единицей $0 \le R^2 \le 1$. Равенство R^2 единице свидетельствует о том, что модель полная и полнее не может быть ($\mathbf{e}'\mathbf{e} = 0$), тогда, как $R^2 = 0$ означает, что модель абсолютно не способна объяснить вариацию наблюдаемых данных.

Коэффициент детерминации, определяемый выражением (4), обладает одним существенным недостатком. При равенстве числа регрессоров q числу наблюдений n величина R^2 равна 1. Кроме того, по мере добавления регрессоров в уравнение, значение R^2 неизбежно возрастает. Это ведет к неоправданному предпочтению моделей с большим числом регрессоров. Отсюда следует, что необходима поправка к R^2, которая бы учитывала число регрессоров и число наблюдений. В результате получаем *скорректированный коэффициент детерминации* (*adjusted coefficient of multiple determination*) $\overline{R}^2$.

$$\overline{R}^2 = 1 - \frac{n-1}{n-q-1}(1 - R^2).$$

Если в уравнение регрессии добавить некоторую переменную, то $\overline{R}^2$ увеличится тогда и только тогда, когда F-статистика для соответствующего коэффициента переменной будет больше единицы.

Коэффициент детерминации R^2связан с $F_{\text{ур}}$ соотношением

$$F_{\text{ур}} = \frac{SSR/q}{SSE/(n-q-1)} = \frac{(SSR/SST)/q}{(1-SSR/SST)/(n-q-1)} = \frac{R^2/q}{(1-R^2)/(n-q-1)}.$$