Адекватность математической модели регрессии

Проверка статистических гипотез. Коэффициенты регрессии

Проверка гипотез относительно коэффициентов регрессии

1.Гипотеза равенства истинного коэффициента β_k значению β_{k0}
H0: $\beta_k = \beta_{k0}$, против альтернативной гипотезы H1: : β_k не равно β_{k0}.
2.Гипотеза одновременного равенства нулю всех, кроме β_0, регрессионных коэффициентов

Проверяем конкуренцию двух моделей

$$Y = \sum_k \beta_k x_k + \varepsilon \qquad Y = \beta_0 + \varepsilon$$

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

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

распределена по Стьюденту с числом с.с. v = n – q – 1, а квадрат ее имеет F-распределение (распределение Снедокора-Фишера) с числом с.с. v1 = 1, v2 = n – q –1

Проверка статистической гипотезы 1

F-распределение (распределение Снедокора-Фишера) с числом с.с. v1 = 1, v2 = n – q –1

(1)

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

$F > F(\alpha, \nu_1, \nu_2)$ **- гипотеза** $\beta_k = \beta_{k0}$ **отвергается; при этом**:

(2)

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

К проверке статистической гипотезы 2

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:

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

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_{yp}

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

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

Таблица дисперсионного анализа

Формула (3) определяет отношение дисперсии оценки модели к дисперсии остатка $F_{yp} = S_\theta^2/S_e^2$. Статистика F_{yp} имеет F-распределение с числом с.с. v1 = q, v2 = n – q – 1. Если $Fyp \geq Fкр(\alpha, v_1, v_2)$, 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.

Таблица дисперсионного анализа (ANOVA) (табл. 1). «Средний квадрат» получается при делении каждой суммы квадратов на соответствующее ей число с.с.

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	$\nu = n-1$	–	–

Проверка статистических гипотез. Коэффициенты регрессии

Адекватность модели

Оценка модели, найденная по экспериментальным данным

$$\hat{\mathbf{Y}} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}_1 + \ldots + \hat{\beta}_k \mathbf{x}_q \qquad R = \frac{(\mathbf{Y} - \overline{\mathbf{Y}})'(\hat{\mathbf{Y}} - \overline{\mathbf{Y}})}{[(\mathbf{Y} - \overline{\mathbf{Y}})'(\mathbf{Y} - \overline{\mathbf{Y}})(\hat{\mathbf{Y}} - \overline{\mathbf{Y}})'(\hat{\mathbf{Y}} - \overline{\mathbf{Y}})]^{1/2}}$$

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**

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

(3)

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

$$R^2 = \frac{\tilde{\hat{\mathbf{Y}}}'\tilde{\hat{\mathbf{Y}}}}{\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}} = \frac{SSR}{SSR + SSE} = \frac{SSR}{SST}$$

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

- скорректированный коэффициент детерминации (adjusted coefficient of multiple determination)

Классический регрессионный анализ. Проверка статистических гипотез. Коэффициенты регрессии

Адекватность модели

Скорректированный коэффициент детерминации (adjusted coefficient of multiple determination)

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

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

$$F_{ур} = \frac{SSR/q}{SSE/(n-q-1)} = \frac{(SSR/SST)/q}{(1-SSR/SST)/(n-q-1)} =$$

$$\frac{R^2/q}{(1-R^2)/(n-q-1)}$$

Списки использованной литературы и источников:

- А.А.Большаков, Р.Н.Каримов «Методы обработки многомерных данных и временных рядов» Москва 2007 г.
- Электронный учебник StatSoft по анализу данных.