УДК 519.3

Д. В. Шлепаков (Ин-т математики НАН Украины, Киев)

СТАТИСТИЧЕСКИЕ КРИТЕРИИ ДЛЯ ПОРЯДКА РЕГРЕССИОННОЙ МОДЕЛИ ПО ЕДИНСТВЕННОЙ СЕРИИ НАБЛЮДЕНИЙ

We construct an order definition test for a linear regression model on the basis of unique series of observations in the case of normally distributed noise with zero mean and limited dispersion. We propose a new statistics of equilibrium for the model with infinite (but unique) series of observations and, by using this statistics, construct an order definition test for this model.

Побудовано критерій визначення порядку лінійної регресійної моделі за єдиною серією спостережень для випадку нормально розподіленого шуму з нульовим середнім та обмеженою дисперсією. Запропоновано нову статистику рівноваги для моделі з нескінченною (але єдиною) серією спостережень і на її основі побудовано критерій визначення порядку моделі.

1. Введение. Определение порядка модели регрессии в случае единственной серии наблюдений — довольно сложная задача. При единственной серии остаточные дисперсии необходимо являются зависимыми. Более того, оказывается, при большой выборке коэффициент корреляции стремится к единице. Однако при определенных условиях можно построить критерий для определения порядка такой модели.

В статье рассматривается модель линейной регрессии, ошибки которой имеют нормальное распределение с нулевым математическим ожиданием и положительно определенной матрицей ковариаций. Рассматриваем нулевую гипотезу о порядке регрессионной модели и альтернативную ей простую гипотезу. Для данной модели построен критерий проверки гипотез и предельное распределение этой статистики. Конструктивный метод, примененный в статье, основан на свойствах известных распределений.

2. Постановка задачи и проблематика. Приведем некоторые предположения относительно рассматриваемой модели: имеем единственную (!) серию n наблюдений:

$$(\mathbf{x}_i, \mathbf{y}_i), \qquad i = \overline{1, n}, \tag{1}$$

в модели линейной регрессии

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}; \tag{2}$$

ошибки $\boldsymbol{\varepsilon}$ имеют нормальное распределение с нулевым математическим ожиданием и положительно определенной матрицей ковариаций

$$\begin{gathered} \mathrm{E}(\boldsymbol{\varepsilon}) = 0, \\ \mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{\Omega}, \qquad \sigma^2 < \infty. \end{gathered} \tag{3}$$

Будем рассматривать гипотезу о порядке модели (1)–(3). За основную гипотезу принимаем гипотезу H_p: модель (1)–(3) имеет порядок p, где $p < n$. Предполагаем, что $\mathrm{rank}(\mathbf{X}) = p$.

Известно, что для любой положительно определенной матрицы $\boldsymbol{\Omega}$ существует такая невырожденная матрица $\mathbf{P}$ размера $n \times n$, что

$$\mathbf{P}\mathbf{P}' = \boldsymbol{\Omega}.$$

Умножая (2) слева на $\mathbf{P}^{-1}$, получаем $\mathbf{P}^{-1}\mathbf{y} = \mathbf{P}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{P}^{-1}\boldsymbol{\varepsilon}$. Обозначим $\mathbf{y} = \mathbf{P}^{-1}\mathbf{y}$, $\mathbf{Q} = \mathbf{P}^{-1}\mathbf{X}$, $\mathbf{u} = \mathbf{P}^{-1}\boldsymbol{\varepsilon}$. Тогда (2) можно представить в виде

$$\mathbf{z} = \mathbf{Q}\boldsymbol{\beta} + \mathbf{u}. \tag{4}$$

Ошибки модели (4) имеют нулевое математическое ожидание

$$\mathrm{E}(\mathbf{u}) = \mathbf{P}^{-1}\mathrm{E}(\boldsymbol{\varepsilon}) = 0 \tag{5}$$

и ковариационную матрицу

$$\mathrm{Cov}(\mathbf{u}) = \mathrm{E}(\mathbf{u}\mathbf{u}') = \mathrm{E}(\mathbf{P}^{-1}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'(\mathbf{P}^{-1})') = \mathbf{P}^{-1}\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')(\mathbf{P}')^{-1} =$$

$$= \mathbf{P}^{-1}\sigma^2\mathbf{P}\mathbf{P}'(\mathbf{P}')^{-1} = \sigma^2\mathbf{I}. \tag{6}$$

По методу наименьших квадратов получаем оценку Гаусса – Маркова вектора параметров $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}_p = (\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{z},$$

или в обозначениях исходной модели

$$\hat{\boldsymbol{\beta}}_p = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y},$$

где матрица $\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}$ невырождена [1, с. 379].

Оценка $\hat{\boldsymbol{\beta}}_p$ не смещена:

$$\mathrm{E}(\hat{\boldsymbol{\beta}}_p) = (\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathrm{E}(\mathbf{z}) = (\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{Q}\boldsymbol{\beta} + (\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathrm{E}(\mathbf{z}) = \boldsymbol{\beta},$$

а ее ковариационная матрица имеет вид

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}_p) = \mathrm{E}\left(\left(\hat{\boldsymbol{\beta}}_p - \mathrm{E}\left(\hat{\boldsymbol{\beta}}_p\right)\right)\left(\hat{\boldsymbol{\beta}}_p - \mathrm{E}\left(\hat{\boldsymbol{\beta}}_p\right)\right)'\right) =$$

$$= (\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathrm{E}(\mathbf{u}\mathbf{u}')\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1} = \sigma^2(\mathbf{Q}'\mathbf{Q})^{-1} = \sigma^2(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}.$$

Учитывая предположение (3), а также используя (5) и (6), получаем $\mathbf{u} \sim$ $\sim N(0, \sigma^2\mathbf{I})$. Обозначим остаточную сумму квадратов S_p:

$$S_p = (\mathbf{z} - \mathbf{Q}\hat{\boldsymbol{\beta}}_p)'(\mathbf{z} - \mathbf{Q}\hat{\boldsymbol{\beta}}_p),$$

или в обозначениях исходной модели:

$$S_p = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_p)'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_p).$$

Тогда несмещенной оценкой для σ^2 будет остаточная дисперсия:

$$\hat{S}_p^2 = \frac{S_p}{\nu}, \quad \text{где} \quad \nu = n - p.$$

Рассмотрим альтернативную гипотезу о порядке модели — гипотезу H_{p+k}. Оценку методов наименьших квадратов вектора $\boldsymbol{\beta}$ для этой гипотезы обозначим $\hat{\boldsymbol{\beta}}_{p+k}$. Соответственно обозначим остаточную сумму квадратов через S_{p+k} и остаточную дисперсию —

$$\hat{S}_{p+k}^2 = \frac{S_{p+k}}{\nu - k}.$$

Тогда справедлива следующая теорема.

Теорема 1. *Остаточные дисперсии альтернативных гипотез* $\hat{S}_p^2$ *и* $\hat{S}_{p+k}^2$ *модели* (1) – (3) *коррелированы с коэффициентом корреляции*

$$r = \sqrt{1 - k/\nu}\,. \tag{7}$$

Доказательство. Легко убедиться, что $\hat{\beta}_p \sim N(\beta, \sigma^2(\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1})$. Величина $\hat{S}_p^2 \sim \frac{\sigma^2}{\nu}\chi_\nu^2$. $\hat{\beta}_p$ не зависит от $\hat{S}_p^2$. Используя свойства χ^2-распределения, получаем $\mathrm{E}(\hat{S}_p^2) = \sigma^2$, $\mathrm{D}(\hat{S}_p^2) = 2\sigma^4/\nu$. Аналогично $\mathrm{E}(\hat{S}_{p+k}^2) = \sigma^2$, $\mathrm{D}(\hat{S}_p^2) = 2\sigma^4/(\nu-k)$. Вычислим величину $\mathrm{E}(\hat{S}_p^2\hat{S}_{p+k}^2)$:

$$\mathrm{E}(\hat{S}_p^2\hat{S}_{p+k}^2) = \frac{2\sigma^4}{\nu(\nu-k)}\mathrm{E}\left(\chi_\nu^2\chi_{\nu-k}^2\right) = \frac{2\sigma^4}{\nu(\nu-k)}\mathrm{E}\left(\sum_{j=p+1}^{n}\xi_j^2\sum_{j=p+k+1}^{n}\xi_j^2\right) =$$

$$= \frac{2\sigma^4}{\nu(\nu-k)}\mathrm{E}\left(\left(\xi_{p+1}^2+\xi_{p+2}^2+\cdots+\xi_{p+k}^2+\sum_{j=p+k+1}^{n}\xi_j^2\right)\sum_{j=p+k+1}^{n}\xi_j^2\right) =$$

$$= \frac{2\sigma^4}{\nu(\nu-k)}\left(k(\nu-k)+3(\nu-k)+(\nu-k)^2-(\nu-k)\right) = \sigma^2\left(1+\frac{2}{\nu}\right).$$

Теперь вычислим коэффициент корреляции r:

$$r = \frac{\mathrm{E}\left(\hat{S}_p^2\hat{S}_{p+k}^2\right)-\mathrm{E}\left(\hat{S}_p^2\right)\mathrm{E}\left(\hat{S}_{p+k}^2\right)}{\sqrt{D\left(\hat{S}_p^2\right)D\left(\hat{S}_{p+k}^2\right)}} = \frac{\sigma^4\left(1+\frac{2}{\nu}\right)-\sigma^4}{\sqrt{\frac{2\sigma^4}{\nu}\frac{2\sigma^4}{\nu-k}}} =$$

$$= \frac{1}{\nu}\sqrt{\nu(\nu-k)} = \sqrt{1-k/\nu}\,.$$

Теорема доказана.

Из теоремы 1 следует, что остаточные дисперсии асимптотически линейно зависимы.

3. Распределение t-статистики. Критерий для определения порядка модели.

Определение 1. *Статистику отношения остаточных дисперсий будем называть t-статистикой модели* (1)–(3):

$$T_{p,k}^{(n)} = \frac{\hat{S}_p^2}{\hat{S}_{p+k}^2}. \tag{8}$$

Теорема 1 явным образом отрицает использование критерия для проверки гипотезы о порядке модели [2]. Попробуем найти распределение статистики (8).

Определение 2. *Обратной нормализованной t-статистикой назовем величину*

$$\zeta_{p,k}^{(n)} = \frac{r^2}{T_{p,k}^{(n)}}. \tag{9}$$

Лемма 1. *Обратная нормализованная t-статистика $\zeta_{p,k}^{(n)}$ модели* (1)–(3) *имеет β-распределение с параметрами $(\nu-k)/2$ и $k/2$.*

Доказательство. На основании теоремы 1 получаем

$$\zeta_{p,k}^{(n)} = \frac{r^2}{T_{p,k}^{(n)}} = r^2\frac{\nu}{\nu-k}\frac{S_p}{S_{p+k}} = \frac{\sum_{j=p+k+1}^{n}\xi_j^2}{\sum_{j=p+1}^{n}\xi_j^2} =$$

$$= \frac{\sum\limits_{j=p+k+1}^{n} \xi_j^2}{\sum\limits_{j=p+k+1}^{n} \xi_j^2 + \sum\limits_{j=p+1}^{p+k} \xi_j^2} = \frac{\chi^2_{\nu-k}}{\chi^2_{\nu-k} + \chi^2_k},$$

причем случайные величины в знаменателе независимы, а отсюда из свойств β-распределения имеем в правой части случайную величину с распределением $\beta((\nu-k)/2, k/2)$.

Теорема 2. *Плотность распределения t-статистики имеет вид*

$$f_{T^{(n)}_{p,k}}(x) = \frac{\Gamma\left(\frac{\nu}{2}\right)}{\Gamma\left(\frac{\nu-k}{2}\right)\Gamma\left(\frac{k}{2}\right)} r^{\nu-2}\left(\frac{x}{r^2}-1\right)^{\frac{k}{2}-1} x^{-\frac{\nu}{2}}, \quad \text{где} \quad x \geq r^2.$$

Доказательство. Используем лемму 1. Производя замену переменных $u = r^2/x$, получаем

$$f_{\zeta^{(n)}_{p,k}}(x) = \begin{cases} \dfrac{\Gamma(\nu_1+\nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} u^{\nu_1-1}(1-u)^{\nu_2-1}, & u \in (0,1); \\ 0, & u \notin (0,1). \end{cases} \tag{10}$$

В рассматриваемом случае $\nu_1 = (\nu-k)/2$, $\nu_2 = k/2$. Из (10) получаем утверждение теоремы.

Лемма 2. *Критерий проверки гипотезы о порядке регрессионной модели* (1) – (3) *против простой альтернативы имеет вид*

$$T^{(n)}_{p,k} < 1 + \frac{k(F_\alpha(k, \nu-k)-1)}{\nu}, \tag{11}$$

где $F_\alpha(k, \nu-k)$ *— α-квантиль F-распределения с* k *и* $\nu-k$ *степенями свободы.*

Доказательство. Из свойств β-распределения известно, что случайная величина

$$\frac{k_1 F(k_1, k_2)}{k_2 + k_1 F(k_1, k_2)}$$

имеет β-распределение $\beta(k_1/2, k_2/2)$. Согласно лемме 1

$$\zeta^{(n)}_{p,k} \sim \beta\left(\frac{\nu-k}{2}, \frac{k}{2}\right).$$

Из представления (9) получаем

$$T^{(n)}_{p,k} \sim r^2\left[1 + \frac{k}{(\nu-k)F(\nu-k, k)}\right].$$

Учитывая (7) и свойства F-распределения, имеем

$$T^{(n)}_{p,k} \sim \frac{\nu-k}{\nu}\left[1 + \frac{kF(k,\nu-k)}{\nu-k}\right] = 1 + \frac{k(F(k,\nu-k)-1)}{\nu}. \tag{12}$$

Представление (12) дает нам выражение для распределения t-статистики $T^{(n)}_{p,k}$ через F-распределение. При таких условиях можем построить простой в применении критерий проверки гипотезы о порядке регрессионной модели (1) – (3)

против простой альтернативы. Этот критерий задается неравенством (11). Следствие доказано.

4. Предельное распределение t-статистики. Рассмотрим бесконечную серию наблюдений

$$(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \ldots, (\mathbf{x}_n, \mathbf{y}_n), \ldots . \tag{13}$$

Построим критерий для проверки гипотезы о порядке модели (1)–(3), (13).

Определение 3. *Назовем величину*

$$\tau_{p,k}^{(n)} = \nu\left(1-\zeta_{p,k}^{(n)}\right) \tag{14}$$

статистикой равновесия.

Теорема 3. *Статистика равновесия* (14) *модели* (1)–(3), (13) *для произвольного* p *при* $n \to \infty$ *сходится по распределению к случайной величине* χ_k^2 / σ^2.

Доказательство. Используя лемму 1, имеем

$$1-\zeta_{p,k}^{(n)} = 1-\frac{\chi_{\nu-k}^2}{\chi_{\nu-k}^2+\chi_k^2} = 1-\frac{\chi_\nu^2-\chi_k^2}{\chi_\nu^2} = \frac{\chi_k^2}{\chi_\nu^2}.$$

Принимая во внимание, что

$$\frac{\chi_\nu^2}{\nu} \xrightarrow[n\to\infty]{} \sigma^2,$$

получаем

$$\tau_{p,k}^{(n)} \xrightarrow[n\to\infty]{} \frac{\chi_k^2}{\sigma^2}.$$

Теорема доказана.

Следствие. *Для достаточно больших* n *критерий проверки гипотезы о порядке регрессионной модели* (1)–(3), (13) *против простой альтернативы имеет вид*

$$T_{p,k}^{(\infty)} < \frac{1}{1-\chi_k^2(\alpha)/\nu\sigma^2},$$

где $\chi_k^2(\alpha)$ — α*-квантиль* χ^2*-распределения с* k *степенями свободы.*

1. *Себер Дж.* Линейный регрессионный анализ. – М.: Мир, 1980. – 466 с.
2. *Кендал М., Стьюарт А.* Статистические выводы и связи: В 2-х т. – М.: Наука, 1973. – 899 с.

Получено 26.05.99