

Эконометрика

Полковников Александр Александрович

Волжский политехнический институт (филиал)
ФГБОУ ВПО "Волгоградский государственный технический университет"

Конспект лекций для студентов направления
"Экономика"

Проверка стандартных предположений

Статистические выводы для моделей парной или множественной регрессии опираются на определенные предположения о модели наблюдений (линейная модель, независимые нормально распределенные ошибки с нулевыми математическими ожиданиями и одинаковыми дисперсиями).

Однако отклонения от стандартных предположений могут существенно исказить эти выводы. Поэтому необходимо иметь инструментарий для диагностики модели на предмет обнаружения отклонений от стандартных предположений и для коррекции выявленных отклонений.

Процедуры, рассмотренные ниже, направлены на выявление следующих типов нарушений стандартных предположений:

- отличие распределения ошибок от нормального,
- неодинаковые дисперсии ошибок,
- статистическая зависимость ошибок в наблюдениях, проводимых в последовательные моменты времени.

Эффективным средством обнаружения отклонений от стандартных предположений о линейной модели наблюдений

$$y = X\theta + e$$

является анализ остатков

$$e_k = y_k - \hat{y}_k, \quad k = 1, \dots, n.$$

Во многих статистических пакетах, в частности в Excel, рассматриваются **стандартизированные остатки**

$$c_k = \frac{e_k}{\sqrt{RMS}},$$

где, как обычно, $RMS = RSS/(n - m - 1)$.

Графики стандартизированных остатков позволяют выявлять типичные отклонения от стандартных предположений о модели. При этом имеется в виду, что ожидается поведение остатков похожее на поведение последовательности независимых в совокупности случайных величин, имеющих одинаковое стандартное нормальное распределение.

Наиболее часто используют график зависимости стандартизированных остатков s_k от оцененных значений \hat{y}_k . Анализ этого графика позволяет выявить следующие нарушения стандартных предположений:

- выделяющиеся наблюдения,
- неоднородность дисперсий (гетероскедастичность),
- неправильная спецификация модели ($E(e_k) \neq 0$).

Помимо графических существуют процедуры проверки стандартных предположений, использующие статистические критерии проверки гипотез. В качестве нулевой берется гипотеза $H_0 : e_1, \dots, e_n$ — независимые нормально распределенные величины с параметрами $(0; \sigma^2)$.

Критерий Голдфелда–Квандта

Если графический анализ остатков указывает на возможную неоднородность дисперсий ошибок, то:

- сначала наблюдения упорядочиваются по предполагаемому возрастанию дисперсий случайных ошибок;
- затем отбрасывается r центральных наблюдений, так что для дальнейшего анализа остается $n - r$ наблюдений;
- производится оценивание выбранной модели отдельно по первым $(n - r)/2$ и по последним $(n - r)/2$ наблюдениям;

Критерий Голдфелда–Квандта

- вычисляется отношение $F = RSS_2 / RSS_1$ остаточной суммы квадратов RSS_2 , полученных при подборе модели по последним $(n - r)/2$ наблюдениям, к сумме квадратов RSS_1 , полученной по первым $(n - r)/2$ наблюдениям;
- гипотеза однородности дисперсий (гомоскедастичности) отвергается, если вычисленное значение F -отношения превышает критический уровень

$F_{1-\alpha} \left(\frac{n-r}{2} - m - 1; \frac{n-r}{2} - m - 1 \right)$, соответствующий выбранному уровню значимости α .

Тест ранговой корреляции Спирмена предназначен для проверки гипотезы о независимости дисперсии остатков от значений какого-либо фактора. Критерий предусматривает упорядочивание модулей остатков и значений факторов, например, по возрастанию, а затем, вычисление **коэффициента корреляции рангов Спирмена**

$$r_S = 1 - \frac{6 \sum_{k=1}^n d_k^2}{n(n^2 - 1)},$$

где d_k — разность между рангами k -го остатка и k -го значения фактора.

Полученное значение коэффициента корреляции проверяют на значимость, рассчитывая фактическое значение t -критерия Стьюдента

$$t_S = \frac{r_S \sqrt{n-2}}{\sqrt{1-r_S^2}}$$

и сравнивая его с квантилью распределения Стьюдента $t_{1-\alpha}[n-2]$ уровня $(1-\alpha)$ с $(n-2)$ степенями свободы, где α — уровень значимости.

Если фактическое значение критерия больше квантили, то гипотеза о гомоскедастичности остатков отклоняется.

Критерий Дарбина–Уотсона применяется, когда наблюдения проводятся последовательно во времени, с равными интервалами, и график изменения остатков во времени указывает на наличие автокоррелированности случайных составляющих e_k . Предположим, что структура автокоррелированности задается соотношением:

$$e_k = \rho \cdot e_{k-1} + \delta_k, \quad k = 2, \dots, n,$$

где $|\rho| < 1$, а δ_k , $k = 1, \dots, n$ — независимые в совокупности нормально распределенные случайные величины с параметрами $(0; \sigma_\delta^2)$.

Статистика Дарбина–Уотсона определяется соотношением:

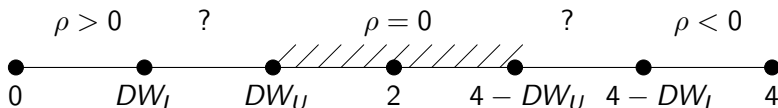
$$DW = \frac{\sum_{k=2}^n (e_k - e_{k-1})^2}{\sum_{k=1}^n e_k^2},$$

где e_1, \dots, e_n — остатки, получаемые при оценивании линейной модели наблюдений.

Фактическое значение статистики DW принадлежит интервалу $[0; 4]$.

Для заданного уровня значимости α по специальным таблицам вычисляются нижняя DW_L и верхняя граница DW_U критерия Дарбина–Уотсона.

При DW “не слишком отличных” от 2, делается вывод об отсутствии автокорреляции.



Коррекция нарушений стандартных предположений

Если нарушены стандартные предположения модели, а именно присутствует гетероскедастичность остатков или остатки коррелируют между собой, то применяется **обобщенный метод наименьших квадратов**. Применение обычного метода наименьших квадратов к модели с нарушенными стандартными предположениями ведет к тому, что оценки параметров модели не будут являться эффективными. Кроме того, дисперсии параметров будут вычислены со смещением, что приведет к ложным выводам об оценке качества модели.

Формула для расчета коэффициентов в модели множественной регрессии имеет вид:

$$\hat{\theta} = (X^T X)^{-1} X^T Y.$$

Обозначим через $\sigma^2 \cdot V$ ковариационную матрицу остатков. Скорректируем расчеты параметров уравнения регрессии с учетом значений матрицы V :

$$\hat{\theta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y.$$

При этом минимизируется сумма

$$\sum_{i=1}^n \sum_{k=1}^n w_{ik} (y_i - \theta_1 x_{i1} - \dots - \theta_m x_{im}) (y_k - \theta_1 x_{k1} - \dots - \theta_m x_{km}),$$

где $w_{ik} = v_{ik}^{(-1)}$ — элементы матрицы V^{-1} .

Основная проблема использования обобщенного метода наименьших квадратов состоит в том, что значения элементов матрицы V неизвестны. Поэтому для применения метода используют оценки элементов матрицы.

В частности, если матрица V диагональная,
 $V = \text{diag}(h_1^2, \dots, h_n^2)$, $h_1^2, \dots, h_n^2 > 0$ и не все одинаковы, т. е.
имеет место гетероскедастичность остатков, но отсутствует
автокорреляция, то

$$V^{-1} = \text{diag}(1/h_1^2, \dots, 1/h_n^2).$$

При этом минимизируется сумма

$$\sum_{i=1}^n \frac{1}{h_i^2} (y_i - \theta_1 x_{i1} - \dots - \theta_m x_{im})^2 = \sum_{i=1}^n \left(\frac{y_i - \theta_1 x_{i1} - \dots - \theta_m x_{im}}{h_i} \right)^2,$$

т. е. минимизируется взвешенная сумма квадратов отклонений.
Поэтому в этом случае обобщенный метод наименьших
квадратов называется **взвешенным методом наименьших
квадратов**.

Выполнение замены

$$w_i = y_i/h_i, \quad z_{ij} = x_{ij}/h_i, \quad j = 1, \dots, m, \quad i = 1, \dots, n,$$

приводит к минимизации суммы

$$\sum_{i=1}^n (w_i - \theta_1 z_{i1} - \dots - \theta_m z_{im})^2,$$

т. е. к обычному методу наименьших квадратов.

Процедура взвешенного МНК состоит в следующем:

- к исходным данным применяется обычный МНК и вычисляются остатки e_k ;
- делаются предположения относительно зависимости дисперсии остатков от каких-либо факторов

$$\hat{e}^2 = f(x_1, \dots, x_m);$$

- находят параметры модели из предыдущего пункта, взяв в качестве фактических значений зависимой переменной случайные остатки e_k , найденные на первом шаге;
- найденные оценки \hat{e}_k^2 используют вместо диагональных элементов матрицы V ;
- обобщенным МНК находят новые значения параметров регрессии.

Другим частным случаем является наличие автокоррелированности остатков:

$$e_k = \rho e_{k-1} + \delta_k, \quad k = 2, \dots, n,$$

$|\rho| < 1$, $\delta_k \sim N(0; \sigma^2)$ и не зависит от e_1, \dots, e_n .

Если предположим, что величины e_k одинаково распределены, то получим:

$$E(e_1) = 0, \quad D(e_1) = \frac{\sigma^2}{1 - \rho^2}$$

и

$$V = \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{pmatrix}.$$

В этом случае обратная матрица равна

$$V^{-1} = \frac{1}{1 - \rho^2} \cdot \begin{pmatrix} 1 & -\rho & 0 & \dots & 0 & 0 \\ -\rho & 1 + \rho^2 & -\rho & \dots & 0 & 0 \\ 0 & -\rho & 1 + \rho^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 + \rho^2 & -\rho \\ 0 & 0 & 0 & \dots & -\rho & 1 \end{pmatrix}.$$

При этом замена

$$w_1 = \sqrt{1 - \rho^2} y_1, \quad z_{1j} = \sqrt{1 - \rho^2} x_{1j}, \quad j = 1, \dots, m,$$

$$w_i = y_i - \rho y_{i-1}, \quad z_{ij} = x_{ij} - \rho x_{i-1,j}, \quad j = 1, \dots, m, \quad i = 2, \dots, n$$

приводит к обычному методу наименьших квадратов и называется **преобразованием Прайса–Уинстена**.

В качестве оценки коэффициента ρ можно использовать

$$\hat{\rho} = \frac{\sum_{k=2}^n e_k e_{k-1}}{\sum_{k=2}^n e_{k-1}^2},$$

где e_1, \dots, e_n — остатки, полученные при оценивании обычным методом наименьших квадратов исходной модели наблюдений.

Фиктивные переменные

При изучении связей результирующий показатель может зависеть не только от количественных, но и от неколичественных факторных признаков. Для включения неколичественной переменной в уравнение регрессии необходимо заменить ее категории числами. Это можно сделать с помощью переменных бинарного типа, так называемых **фиктивных переменных**:

$$z = \begin{cases} 1, \\ 0. \end{cases}$$

Для учета влияния одной неколичественной переменной, имеющей k уровней, используют $(k - 1)$ фиктивную переменную.

Один из уровней принимается в качестве базового. Ему соответствует равенство нулю всех фиктивных переменных. Каждому из остальных $(k - 1)$ уровней соответствует по одной фиктивной переменной, которую приравниваем к единице, а остальные переменные — к нулю.

Рассмотрим уравнение регрессии с одним количественным фактором x и фиктивной переменной z :

$$y = a + bx + cz + e.$$

Изменение значения фиктивной переменной с 0 на 1 приводит к изменению результата на величину c при сохранении коэффициента регрессии b . Поэтому переменную z называют фиктивной переменной сдвига.

Рассмотрим уравнение регрессии с одним количественным фактором x и фиктивной переменной z :

$$y = a + bx + dxz + e.$$

Изменение значения фиктивной переменной с 0 на 1 приводит к изменению коэффициента регрессии с b на $b + d$. Поэтому переменную z называют **фиктивной переменной наклона**.

СПАСИБО ЗА ВНИМАНИЕ!