КОРРЕКТИРОВКА СПЕЦИФИКАЦИИ МОДЕЛИ МНОЖЕСТВЕННОЙ РЕГРЕССИИ ПРИ НАЛИЧИИ МУЛЬТИКОЛЛИНЕАРНОСТИ ИСХОДНЫХ РЕГРЕССОРОВ

Орлова И.В.

Финансовый университет при Правительстве РФ,
Россия, г. Москва, Ленинградский пр.,49
ivorlova@fa.ru

Аннотация: Работа посвящена задаче построения регрессионных моделей в условиях мультиколлинеарности исходных регрессоров. Предложен метод преобразования переменных, позволяющий уменьшить степень мультиколлинеарности без большого ущерба для содержательного смысла модели и применимого как для пространственных данных, так и для многомерных временных рядов.

Ключевые слова: мультиколлинеарность, тренд, многомерные временные ряды.

Введение

При построении эконометрических моделей наличие мультиколлинеарности регрессоров приводит к нежелательным последствиям и поэтому возникает необходимость избавиться от нее, либо хотя бы ослабить степень мультиколлинеарности [1, 2, 3] .

В случае, когда исходные переменные представляют собой нестационарные многомерные временные ряды, каждый из них зачастую имеет тренд, что приводит к появлению ложной корреляции и, следовательно, регрессоры становятся мультиколлинеарными, даже если по смыслу они независимы друг от друга. Для уменьшения степени мультиколлинеарности предлагается разбить исходные регрессоры на две части – тренд и остаток. Остатки свободны от корреляции, вызванной трендами переменных, а коэффициенты регрессии при них равны коэффициентам регрессии при исходных переменных. При этом ошибка предложенной трендово-факторной модели меньше ошибок как трендовой модели, так и модели по исходным временным рядам.

В случае пространственных переменных предложенный подход приводит к методу неполной ортогонализации исходных переменных с помощью замены некоторых исходных переменных на остатки регрессии этих переменных на другие экзогенные переменные, тесно связанные с ними. При этом в новом уравнении регрессии меняются коэффициенты регрессии только при переменных, выступающих в роли зависимых переменных во вспомогательных уравнениях регрессии. Метод приводит к поддающимся содержательной интерпретации переменным, менее коррелированным по сравнению с исходными переменными. При этом получены также формулы связи ковариационных матриц коэффициентов регрессии по старым и новым переменным.

1 Решение задачи в случае, когда исходные регрессоры являются многомерными временными рядами

Без ограничения общности, предположим, что тренды и регрессоров (j=1,…,m) являются многочленами не выше второй степени. Модель линейной регрессии на имеет вид:

$$\sum \quad _t,\ t=1,\ldots,n.$$

1.1 Построение модели и прогнозирование по ней

Значения $x_j(t)$ представим в виде суммы тренда и отклонений от тренда: $x_j(t) = a_{0,j} + a_{1,j}t + a_{2,j}t^2 + u_j(t)$, при этом никаких предположений относительно остатков $u_j(t)$ не делается. Тогда $y(t)$ примет вид:

$y(t) = \gamma_0 + \gamma_1 t + \gamma_2 t^2 + \sum_{j=1}^{m} \beta_j u_j(t) + \varepsilon_t$, где $\gamma_0 = \beta_0 + \sum_{j=1}^{m} \beta_j a_{0,j}$, $\gamma_1 = \sum_{j=1}^{m} \beta_j a_{1,j}$, $\gamma_2 = \sum_{j=1}^{m} \beta_j a_{2,j}$.

Как видим, коэффициенты регрессии при остатках $u_j(t)$ те же самые, что и коэффициенты регрессии при исходных переменных $x_j(t)$, но получены они по регрессорам, имеющим гораздо меньший уровень мультиколлинеарности.

Для прогнозирования уровня зависимой переменной при t=L надо иметь прогнозные значения $u_j(L)$. Поскольку среднее значение $u_j(t)$ равно нулю, то естественно положить $u_j(L)$ равными нулю. Тогда прогнозное значение $y(L)$ равно $y(L) = \gamma_0 + \gamma_1 L + \gamma_2 L^2$. Полученные значения $y(L)$ отличаются от прогноза, полученного по трендовой модели, поскольку в трендово-факторной модели присутствуют, кроме t, переменные $u_j(t)$. Поэтому же доверительные интервалы прогнозов будут меньше интервалов прогнозов по трендовой модели, в которой учитывается только время t .

1.2 Количественная оценка повышения точности прогноза

Ошибка прогноза $\hat{\sigma}_p$ при $t = L$ равна $\hat{\sigma}_p = \hat{\sigma}_e \sqrt{1 + \left(\boldsymbol{v}^T(L), \boldsymbol{u}^T(L)\right)(U^T U)^{-1}(\boldsymbol{v}^T(L), \boldsymbol{u}^T(L))^T}$, где $\hat{\sigma}_e$ – стандартная ошибка остатков модели, знак T означает транспонирование, $\boldsymbol{v^T}(L) = (1, L, L^2)$, $\boldsymbol{u^T}(L) = (u_1(L), u_2(L), \ldots, u_m(L))$, $U -$ блочная матрица, $U = (T, D)$; строки матриц T и D равны соответственно $\boldsymbol{v^T}(t)$ и $\boldsymbol{u^T}(t)$, t=1,…,n. Можно показать, что матрица $(U^T \cdot U)^{-1}$ равна $(U^T \cdot U)^{-1} = \begin{pmatrix} V_{t,t}^{-1} - B_{u,t}C & C^T \\ C & V_{u,u}^{-1} - B_{t,u}C^T \end{pmatrix}$, где $B_{t,u} = (D^T D)^{-1} D^T T$ - матрица коэффициентов регрессии v(t) на u(t), $B_{u,t} = (T^T T)^{-1} T^T D$, $V_{t,t} = T^T T$, $V_{u,u} = DT^T D$, $C = (V_{u,t}B_{u,t} - V_{u,u})^{-1} B_{u,t}^T$, $V_{\text{u,t}} = D^T T$. Подставляя $\boldsymbol{u}(L) = \boldsymbol{0}$ в формулу для $\hat{\sigma}_p$, получаем, что ошибка прогноза $\hat{\sigma}_p$ равна

$\hat{\sigma}_p = \hat{\sigma}_e \sqrt{1 + \boldsymbol{v^T}(L)\left(V_{t,t}^{-1} - \text{B}_{u,\text{t}}\text{C}\right)\boldsymbol{v}(L)}$. Если бы прогнозирование осуществлялось только по тренду, без учёта uj(t), то ошибка прогноза вычислялась бы по формуле

$\hat{\sigma}_{p,trend} = \hat{\sigma}_{e,trend}\sqrt{1 + \boldsymbol{v}^T(L)V_{t,t}^{-1}\boldsymbol{v}(L)}$, где $\hat{\sigma}_{e,trend}$ - оценка стандартного отклонения трендовой модели. Так как остатки U_jt) не зависят от t, то матрицы оценок коэффициентов регрессий t, t^2 на uj, как и регрессий uj на t, t^2, близки к нулевой, $B_{t,u} \approx 0$, $B_{u,t} \approx 0$, и, следовательно, $C \approx 0$. Поскольку матрица $B_{u,t}C$ близка к нулевой, то ошибки прогноза двух моделей фактически отличаются только множителем, равным стандартной ошибке остатков модели. Поскольку $\hat{\sigma}_e < \hat{\sigma}_{e,trend}$, ошибка прогноза трендово-факторной модели будет меньше ошибки прогноза трендовой модели примерно настолько, во сколько раз $\hat{\sigma}_e$ меньше $\hat{\sigma}_{e1}$.

2 Решение задачи в случае, когда исходные регрессоры являются пространственными переменными

Модель линейной регрессии в этом случае имеет вид:

$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_m x_{im} + \varepsilon_i$, $i = 1, \ldots n$,

где x_{ij} – значение X_j в i-ом наблюдении, ε_i – случайное возмущение.

Допустим, что среди регрессоров есть группы тесно связанных между собой переменных (эти группы можно определить, например, с помощью корреляционной матрицы или с помощью метода Belsley-Kuh-Welsch [4]). В каждой группе выберем одну или несколько переменных и с помощью метода наименьших квадратов (МНК) получим оценки коэффициентов регрессий остальных переменных группы на выбранные переменные. При этом никаких допущений относительно остатков

регрессии не делается. Остатки этих регрессий обозначим через U_j. Назовём регрессоры X_j, выступавшие в качестве зависимых переменных в дополнительных регрессиях, “выбранными” переменными, а остальные – “не выбранными” перевменными. Уравнение регрессии “не выбранных” на “выбранные” переменные группы имеет вид:

= ,

Из последнего уравнения получаем

=x im·

Как видим, остатки U_j являются линейными комбинациями исходных регрессоров X_k и константы. Выберем в качестве регрессоров эндогенной переменной Y остатки вместо X_j, если выступает в роли зависимой переменной вспомогательной регрессии, остальные равны исходным переменным X_j. Есть основания считать, что уровень мультиколлинеарности переменных меньше уровня мультиколлинеарности переменных (к примеру, для парной регрессии остатки от регрессии ортогональны зависимой переменной).

Обозначим через U матрицу значений новых регрессоров размерности n×). Так как коррелируют между собой меньше, чем , то вычисление коэффициентов регрессии и их интерпретация не связаны с затруднениями, вызванными наличием мультиколлинеарности регрессоров.

Обозначим через A матрицу линейного преобразования исходных переменных X_j. Тогда U=X·A^T. Матрица A является невырожденной (если X_j не являются строго мультиколлинеарными), следовательно, преобразование переменных является взаимно-однозначным.

Пусть γ и β - векторы коэффициентов регрессии Y на переменные U_j и X_j соответственно. Взаимосвязь векторов коэффициентов регрессии описывается формулой β= A· γ. В силу того, что преобразование, определяемое матрицей A, является взаимно-однозначным, остатки обеих регрессий, т. е., Y-X·β и Y-U·γ, совпадают и тогда совпадают и коэффициенты детерминации.

Если – «не выбранная» переменая, то коэффициент регрессии при равен коэффициенту регрессии β_j при X_j, если же X_j - «выбранная» переменая, то γ_j равен плюс сумма членов по всем таким k, для которых входит регрессором во вспомогательную регрессию на связанные с ним регрессоры. В этом случае может интерпретироваться как приращение Y при изменении на единицу с учётом изменений других регрессоров, связанных с корреляционной зависимостью.

Для прогнозирования значения Y при заданных значениях исходных регрессоров

,

надо вычислить прогнозное значение u , по формуле и затем воспользоваться уравнением регрессии Y по $_j$.

Заключение

Предложенный метод неполной ортогонализации пространственных исходных переменных позволяет, в отличие от ортогонализации с помощью метода главных компонент, получать поддающиеся содержательной интерпретации результаты моделирования. Получены формулы, позволяющие при необходимости перейти к уравнению регрессии по первоначальным переменным и получить все характеристики этого уравнения.

Рассмотренная в работе трендово-факторная модель позволяет частично решить проблему мультиколлинеарности при моделировании многомерных временных рядов. При этом ошибка прогноза трендово-факторной модели меньше ошибок других моделей.

Литература

1. Péter Kovács Examination of Multicollinearity in Linear Regression Models Examination of PETRES’ Red. Theses of PhD Dissertation Szeged 2008 http://docplayer.hu/3607097-Examination-of-multicollinearity-in-linear-regression-models-examination-of-petres-red.html (дата обращения 12.03.2019)

2. *Gordinsky, A.* (2016) New Facts in Regression Estimation under Conditions of Multicollinearity. Open Journal of Statistics, 6, 842-861. doi: 10.4236/ojs.2016.65070

3. *Бабешко Л.О., Орлова И.В.* Модификация трендово-факторной модели при прогнозировании по многомерным временным рядам// "Фундаментальные исследования"—2019. —№ 3, С. 5-10.

4. Regression Diagnostics - Identifying Influential Data and Sources of Collinearity (David A. Belsley, Edwin Kuh, Roy E. Welsch) John Wiley & Sons, New York. 1980. — P. 297.