Data Science.
Lectures. Week 1.
Экономическое моделирование

Зудин Антон

28 сентября 2022 г.

Содержание

1 Определения

- ЛММР - линейная модель множественной регрессии
- ОМНК(GLS) - обощённый МНК
- γ - уровень доверия, надёжности; α - уровень значимости, $\alpha = 1 - \gamma$
- $E\xi$ - мат. ожидание ξ
- $\mathcal{D}\xi$ - дисперсия ξ
- $\mathcal{N}(a, \sigma^2)$ - нормальное распределение с параметрами (a, σ^2)
- $\mathcal{F}(q, n-k)$ - распределение Фишера с $(q, n-k)$ степенями свободы
- $V(\xi, \eta)$ - ковариационная матрица для случайных величин ξ и η
- $s.e.$[1]$(\hat{\beta}_{j,MHK}) = \sqrt[2]{\widehat{\mathcal{D}(\hat{\beta}_{j,MHK})}}$

2 Статистические выводы ЛММР

- Построение доверительных интервалов
- Проверка гипотез

УТВ

Если выполнены условия теоремы Гаусса Маркова и $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$,
то

1. $\hat{\beta}_{j,MHK} \sim \mathcal{N}\ (\beta_j, \mathcal{D}\hat{\beta}_{j,MHK})$
2. $\frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} = t_j \sim t(n-k)$[a]

[a] $t(n-k)$ - распределение Стьюдента с $n-k$ степенями

$s.e.V(\hat{\beta}_{MHK}) = \sigma^2 (X^T X)^{-1}$
Ковариационная матрица $V(\hat{\beta}_{MHK})$ это такая матрица, что $\left[V(\hat{\beta}_{MHK})\right]_{ij} = cov(\hat{\beta}_i, \hat{\beta}_j)$
Нам нужно оценить σ^2, чтобы найти ковариационную матрицу.
Тогда $\hat{V}(\hat{\beta}_{MHK}) = \hat{\sigma^2}(X^T X)^{-1}$.

$$\hat{\sigma^2} = \frac{1}{n-k} \sum_{i=1}^{n} e_i^2 = \frac{ESS}{n-k}$$

Мы взяли имеено $\frac{ESS}{n-k}$, так как эта оценк является несмещённой, то есть $E(\frac{ESS}{n-k}) = \sigma^2$
На диагонали $\hat{V}(\hat{\beta}_{MHK})$ стоят оценённые дисперсии, т. е. $\left[\hat{V}(\hat{\beta}_{MHK})\right]_{jj} = \widehat{\mathcal{D}}(\hat{\beta}_{j,MHK}) = \hat{\sigma^2}\left[(X^T X)^{-1}\right]_{jj}$
$s.e.(\hat{\beta}_{j,MHK}) = \sqrt[2]{\widehat{\mathcal{D}}(\hat{\beta}_{j,MHK})}$
Тогда $\frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} = t_j \sim t(n-k)$.
Пусть мы хотим построить доверительный интервал с уровнем доверия γ.

[1] s.e. - standard error

$\left|\frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)}\right| < t_{Table}$, где t_{Table} - критическая тчока статистики, можно найти в таблице распределения

$\hat{\beta}_j - t_{Table} \cdot s.e.(\hat{\beta}_j) < \beta_j < \hat{\beta}_j + t_{Table} \cdot s.e.(\hat{\beta}_j)$

3 Проверка гипотез

$H_0 : \beta_j = \beta^0$(какое-то известное число)

!Самая главная гипотеза: $\beta_j = 0$.

Если $\beta_j = 0$, то x_j не оказывает значимого влияния на результирующую переменную y.

Критическая статистика: $\left|\frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)}\right| > t_{Table}$

УТВ

Если $\left|\frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)}\right| > t_{Table}$,

то мы отвергаем гипотезу H_0: произошло маловероятное событие.

4 Линейные ограничения, связывающие β

$H_0 : Q\beta = r$, где $\underset{q\times k}{Q} = (h_{ij}), i = \{1, \ldots, q\}; j = \{1, \ldots, k\}$

То есть, ограничение имеет вид:

$$\begin{cases} h_{11}\beta_1 + h_{12}\beta_2 + \ldots + h_{1k}\beta_k = r_1 \\ \ldots \\ h_{q1}\beta_1 + h_{q2}\beta_2 + \ldots + h_{qk}\beta_k = r_q \end{cases}$$

Если это $(Q\hat{\beta}_{MHK} - r)^T(Q\hat{\beta}_{MHK} - r) > \Delta_{critical}$ выполняется, то гипотезу нужно отвергать, так как сумма квадратов отклонений слишком большая.

Проблема здесь в том, что мы не знаем закон распределение суммы квадратов отклонения.

Смотрим $(Q\hat{\beta}_{MHK} - r)^T\Sigma(\hat{\beta}_{MHK} - r)(Q\hat{\beta}_{MHK} - r) \sim \chi^2(q)$, где $\Sigma(Q\hat{\beta}_{MHK} - r) = V(Q\hat{\beta}_{MHK} - r)^{-1}$

4.1 Как обычно считают

Часто вместо подсчёта вышеуказанной статистики считают следуюущим образом:

Есть модель: $y = X\beta + \varepsilon$

$H_0 : Q\beta = r$

1. без учёта ограничений, получаются остатки ESS_{unres}
2. с учётом ограничений, получаются остатки ESS_{res}

УТВ

$F_{stat} = \frac{(ESS_{res} - ESS_{unres})/q}{ESS_{unres}/n - k} \sim \mathcal{F}(q, n-k)$[a],

где n - объём выборки, k -количество регрессоров в модели, q - число ограничений

[a] $\mathcal{F}(q, n-k)$ - распределение Фишера с числом степеней свободы $(q, n-k)$

$ESS_{unres} < ESS_{res}$, так как без ограничений больше "степеней свободы поэтому можем сделать сумму квадратов меньше.

Если $F_{stat} > F_{threshold}$, то отвергаем H_0.

Любая статистика должна обладать 3-мя свойствами:

1. вычисляемая
2. известнвый закон распределение
3. величина критической статистики должна быть мерой адекватности H_0 и данных

Наша F_{stat} удовлетворяет всем 3-м свойствам. Самое интересное: 3. выполняется, так как, если F_{stat} маленькое, тогда $(Q\hat{\beta}_{MHK} - r)^T \Sigma (\hat{\beta}_{MHK} - r)(Q\hat{\beta}_{MHK} - r)$ небольшая, то есть взвешанная сумма квадратов маленькая $\Rightarrow$Q$\beta - r$ небольшая. То есть, $Q\beta \approx r$

4.2 Практическое применение

Пример №1

Основные этапы:

а) Чтобы проверить стоит ли использовать модель

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i,$$

нужно проверить $H_0 : \begin{cases} \beta_2 = 0 \\ \beta_3 = 0 \end{cases}$

$F_{stat} = \dfrac{(ESS_{res} - ESS_{unres})/2}{ESS_{unres}/(n-3)} \sim \mathcal{F}(2, n-3)$

Здесь ещё $F_{stat} = \dfrac{R^2_{unres}/(k-1)}{(1-R^2_{unres})/(n-k)}$. $R^2_{res} = 0$, т.к. $y_i = \beta_1 + \varepsilon_i$.

Если $F_{stat} > F_{threshold}$, то мы отвергаем H_0. То есть, модель *хоть немного хорошая.*

б) Аналогичным образом проверяем $H_1 : \beta_2 = 0$ и $H_2 : \beta_3 = 0$

Пример №2

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i \tag{1}$$

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \beta_4 w_i + \beta_5 v_i + +\beta_6 d_i + \varepsilon_i \tag{2}$$

Хотим узнать какая из моделей лучше.

Проверяем гипотезу H_0: $\begin{cases} \beta_4 = 0 \\ \beta_5 = 0 \\ \beta_6 = 0 \end{cases}$

Если мы принимаем $H_0 \Rightarrow$ (1) лучше, чем (2).

Также $F_{stat} = \dfrac{(R^2_{unres} - R^2_{res})/3}{(1-R^2_{un})/(n-6)}$

Если отвергаем H_0, тогда по-отдельности проверяем $\beta_4, \beta_5, \beta_6$ на целесообразность их включения в модель.

5 Нестандартные модели

5.1 Гетероскедастичность

Если $E\varepsilon_i$=0 и $\mathcal{D}\varepsilon_i = \sigma_i^2$, то это случай гетероскедастичности.

В этом случае мы переходим к новой модели:

$$\begin{array}{cccc} x, y & y = X\beta + \varepsilon & \text{неклассическая} & MHK \\ \downarrow & & \downarrow & \\ \tilde{x}, \tilde{y} & \tilde{y} = \tilde{X}\beta + \tilde{\varepsilon} & \text{классическая} & OMHK \end{array}$$

Идея: исходная модель приобразуется так, чтобы гетероскедастические ошибки первой модели переходили в гомоскедастические второй.

$\hat{\beta}_{j,MHK}$ всё ещё хорошие (несмещённые и состоятельные) оценки. Однако возникают проблемы с $s.e.(\hat{\beta}_{j,MHK})$.

Как мы решаем проблему $s.e.(\hat{\beta}_{j,MHK})$:

- Используем поправки Уайта и Ньюи-Уэста для $s.e.\hat{\beta}_{j,MHK}$. Мы всё также считаем $\dfrac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)}$.
- Доступный ОМНК - ассимптотически эффективный

5.2 Стахостические регрессоры

МНК оценки перестают быть состоятельными.

Возникают 2 метода:

- МИП(метод инстументальных переменных)
- 2МНК(2-х шаговый МНК)

6 Временные ряды

Пусть есть временной ряд $y_1, y_2, \ldots, y_T$, нужно найти $\hat{y}_{T+1} = f(y_1, \ldots, y_T)$.

Есть 2 основных подхода к моделированию:

1. Структурное моделирование - можем, используя экономическую теорию, предложить какую-то объясняющю модель: показать, какие переменные объясняющие, как будет объсняться результирующая перменная через них
2. Неструктурное моделирование - просто занимаемся "подгонкой без особой теории

6.1 Авторегрессионная модель

- $AR(1) : y_t = \delta + \theta y_{t-1} + \varepsilon_t$ - модель авторегресси 1-го порядка
- $AR(2) : y_t = \delta + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \varepsilon_t$ - модель авторегресси 2-го порядка

ε_t - белый шум: независимые, одинаково распределённые случ. величины $E\varepsilon_t = 0, \mathcal{D}\varepsilon_t = \sigma_\varepsilon^2, \forall k\ cov(\varepsilon_t, \varepsilon_{t+k}) = 0$

На y_t влияет ε_t - некоторая новость, которая рождена в момент t и y_{t-1} - вчерашнее значение y.

6.2 Модель скользящего среднего

- $MA(1) : y_t = \varepsilon_t + \alpha_1 \varepsilon_{t-1}$
- $MA(q) : y_t = \varepsilon_t + \alpha_1 \varepsilon_{t-1} + \ldots + \alpha_q \varepsilon_{t-q}$

6.3 Стационарность

2 основных вида рядов:

- *Стационарный* - самые главные характеристики во времени не меняются:
  $Ey_t = a, \mathcal{D}y_t = \gamma_0, cov(y_t, y_{t-k}) = \gamma_k$, т.е. ковариация между двумя величинами зависит лишь от того, сколько между величинами времени
- *Нестационарный* - главные характеристики во времени меняются