

Обзор теоретического материала

гос магистры

Множественная линейная регрессия: общая формулировка, методы оценки

- Зависимая (эндогенная, объясняемая) переменная y представляется в виде функции $y=f(x_1, x_2, \dots, x_k, \beta_1, \dots, \beta_p)$ от независимых (экзогенных, объясняющих) переменных x_1, x_2, \dots, x_k .
 β_1, \dots, β_p – параметры модели, которые подлежат определению
- В зависимости от *вида функции* модели делятся на
 - *линейные*
 - *нелинейные*
- В зависимости от количества экзогенных переменных
 - *на модели парной регрессии*
 - *на модели множественной регрессии*

Классическая линейная регрессионная модель . Общий вид

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \dots + \beta_k x_{ik} + \varepsilon_i$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{12} & \dots & x_{1k} \\ 1 & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n2} & \dots & x_{nk} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_k \end{bmatrix}$$

Соотношение предполагается справедливым для всех возможных наблюдений, но мы наблюдаем только выборку из n наблюдений.

Y – n -мерный вектор, X – матрица размерности $n \times k$ (по строкам – значения каждого из факторов в i -м наблюдении, по столбцам – значения j -го фактора в каждом из наблюдений).

Метод наименьших квадратов

$$RSS = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - b_1 - b_2 x_i)^2 \rightarrow \min$$

$$RSS = (\mathbf{Y} - \mathbf{X}\mathbf{b})^T (\mathbf{Y} - \mathbf{X}\mathbf{b}) \rightarrow \min$$

Метод наименьших квадратов

$$\frac{\partial RSS}{\partial \mathbf{b}} = -2\mathbf{X}^T \mathbf{Y} + 2\mathbf{X}^T \mathbf{X} \mathbf{b} = 0$$

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

Другие методы оценки

- Метод максимального правдоподобия
 - Является альтернативой МНК
 - Применим для обобщенных моделей регрессии
- Нелинейный метод наименьших квадратов
 - Для оценки нелинейных моделей
- Взвешенный метод наименьших квадратов
 - Для оценки моделей с гетероскедастичностью

*Гипотезы классической
регрессионной модели с
нестохастическими регрессорами*

- (A.1) Модель линейна по **параметрам** и правильно специфицирована
 - Примеры линейных по параметрам моделей:
- $y = \alpha + \beta x + \varepsilon$
- $\ln(y) = \alpha + \beta \ln(x) + \varepsilon$
- $y = \alpha + \beta/x + \varepsilon$

Предпосылки регрессионной модели с нестохастическими регрессорами

- (A.2) Матрица X – матрица размерности $n \times k$ имеет ранг $k \leq n$ (отсутствие совершенной мультиколлинеарности)
 - Математически это означает, что для $X^T X$ имеется обратная матрица
 - Для простой (парной) регрессии это условие означает, что объясняющая переменная в выборке имеет ненулевую дисперсию

*Предпосылки классической
регрессионной модели с
нестохастическими регрессорами*

- (A.3) $E(\varepsilon) = 0$ – математическое ожидание случайного возмущения равно нулю и в среднем линия регрессии должна быть истинной
 - Ожидаемое значение случайного возмущения равно нулю в каждом наблюдении
 - Случайный член не имеет систематического смещения ни в положительную, ни в отрицательную сторону

Предпосылки классической регрессионной модели с

нестохастическими регрессорами

- (A.4) $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \sigma^2 \mathbf{I}$ – регрессионные остатки гомоскедастичны и нет автокорреляции
- Т.е. ковариационная матрица вектора регрессионных остатков пропорциональна единичной матрице, а значит

$$\text{Var}(\varepsilon_i) = \text{Const} = \sigma^2$$

Все регрессионные остатки имеют одну и ту же дисперсию

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$$

Разные регрессионные остатки не коррелируют

Гомоскедастичность и отсутствие автокорреляции

Ковариационная матрица вектора регрессионных остатков

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \begin{pmatrix} \text{Var}(\varepsilon_1) & \text{Cov}(\varepsilon_1, \varepsilon_2) \dots & \text{Cov}(\varepsilon_1, \varepsilon_n) \\ \text{Cov}(\varepsilon_1, \varepsilon_2) & \text{Var}(\varepsilon_2) \dots & \text{Cov}(\varepsilon_2, \varepsilon_n) \\ \dots & \dots & \dots \\ \text{Cov}(\varepsilon_n, \varepsilon_1) & \text{Cov}(\varepsilon_n, \varepsilon_2) \dots & \text{Var}(\varepsilon_n) \end{pmatrix} = \sigma^2 \mathbf{I}$$

Теорема Гаусса-Маркова

- При соблюдении предпосылок (A1)-(A4) оценки МНК будут BLUE оценками, то есть

- наилучшими (наиболее эффективными)

Best

- линейными (комбинациями наблюдаемых Y_i)

Linear

- несмещенными оценками параметров регрессии

Unbiased Estimators

Несмещенность оценки параметров регрессии

$$E(\hat{\beta}) = \beta$$

- Повторяя оценку уравнения регрессии на различных выборках мы ожидаем, что среднее значение полученных оценок каждого параметра равно его истинному (теоретическому) значению
- Нарушение предположения о гомоскедастичности и отсутствии автокорреляции не влияет на несмещенность оценок
 - Это следствие предположения А3

Несмещенность оценки дисперсии случайной составляющей

Оценка дисперсии случайной
составляющей

$$\text{Var}\hat{\boldsymbol{\varepsilon}} = s^2 = (\mathbf{e}^T \mathbf{e}) / (n - k) = \text{RSS} / (n - k)$$

Эта оценка является несмещенной

$$E(\text{Var}\hat{\boldsymbol{\varepsilon}}) = \sigma^2$$

Свойства оценок на конечных выборках

- Оценка дисперсии случайной составляющей имеет χ^2 распределение с $(n-k)$ степенями свободы
- Оценки вектора параметров и оценка дисперсии случайной составляющей не коррелируют и независимы друг от друга
- Свойства оценок важны для проведения тестов относительно неизвестных параметров регрессии
 - Тест Фишера на значимость уравнения в целом
 - Тест Стьюдента на значимость оценки параметров

Понятие автокорреляции, ее причины, последствия, способы выявления и устранения

- Наличие корреляции между остатками текущих и предыдущих наблюдений (нарушение предпосылки A4)

Исходные предпосылки обобщенной модели

Предположения

- ~~(A.4) $E(\varepsilon\varepsilon^T) = \sigma^2 \mathbf{I}$
(классическая регрессия)~~



- (A.4') $E(\varepsilon\varepsilon^T) = \sigma^2 \mathbf{\Omega} \neq \sigma^2 \mathbf{I}$
(обобщенная регрессия)



$\mathbf{\Omega}$ – положительно определенная матрица

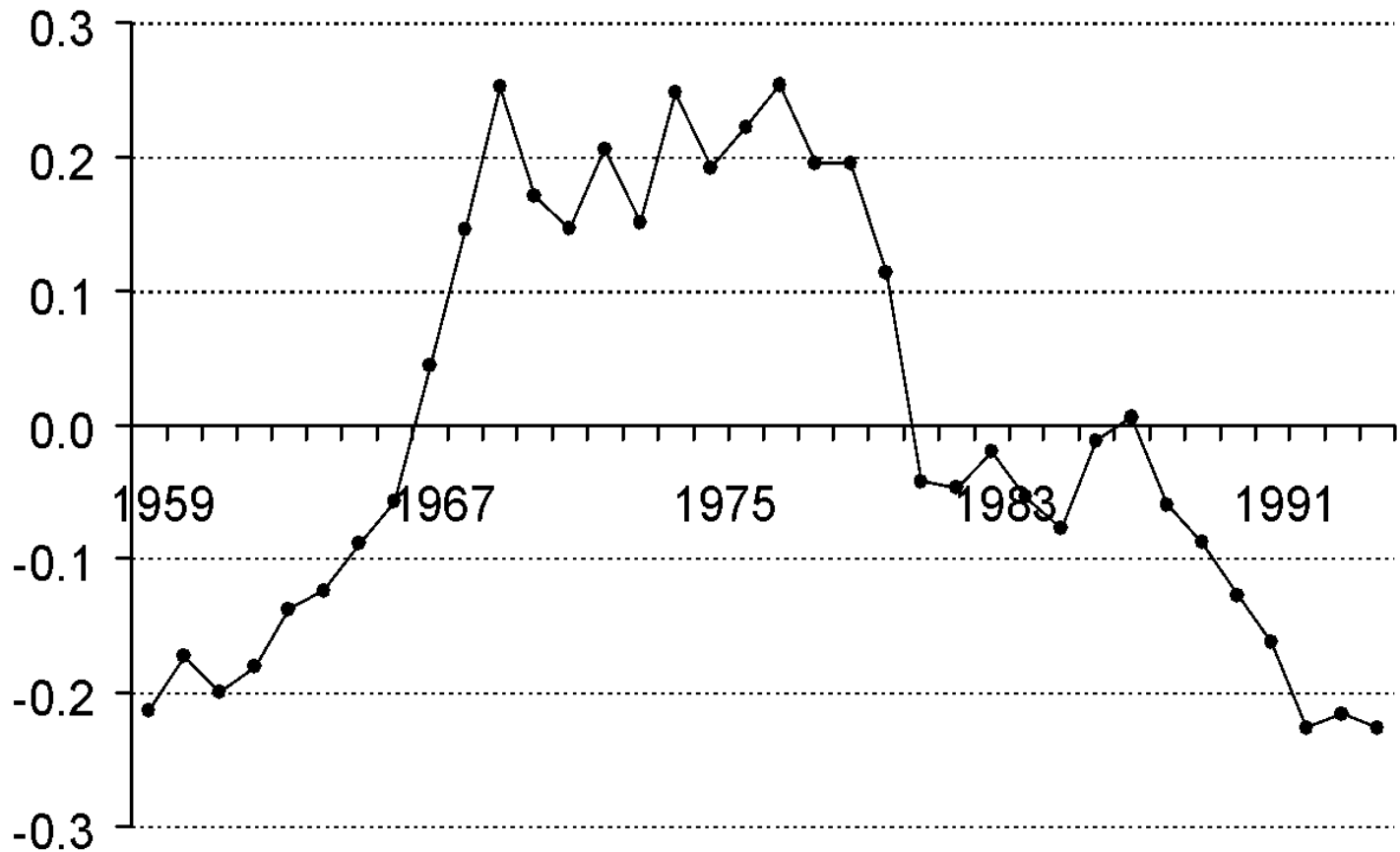
Следствия

- ~~МНК-оценка эффективная~~
- МНК-оценка остается несмещенной и состоятельной, но становится **неэффективной**
- Формулы для вычисления стандартных ошибок изменяются

Ковариационная матрица вектора регрессионных остатков в случае авторегрессии первого порядка

$$\sigma^2 \mathbf{\Omega} = \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{T-2} & \rho^{T-1} \\ \rho & 1 & \rho & & \dots & \rho^{T-2} \\ \rho^2 & \rho & 1 & & \dots & \rho^{T-3} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \rho^{T-1} & \dots & \dots & \dots & \rho & 1 \end{pmatrix}$$

График остатков в случае сильной положительной автокорреляции



Причины автокорреляции

- Ошибки спецификации модели (пропуск важной объясняющей переменной, использование ошибочной функциональной зависимости между переменными и т.д.)
- Ошибки измерений
- Характер наблюдений (например, данные временных рядов)

Последствия автокорреляции

- Оценки параметров остаются несмещенными
- Оценка дисперсии возмущений смещена \square
 - Смещены оценки стандартных ошибок
 - Некорректно определяются доверительные интервалы параметров модели и значений эндогенной переменной

Автокорреляция первого порядка

Простейший вид АК

$$\varepsilon_t = \rho \varepsilon_{t-1} + v_t$$

$-1 < \rho < 1$ – коэф. автокорреляции 1-го
порядка

v_t – свободная от АК случайная
составляющая

Тест Дарбина - Уотсона

$H_0: \rho=0$ – нет АК

$H_1: \rho>0$ – есть положительная АК

Проверка гипотезы:

1. Рассчитывается тестовая статистика

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{RSS}$$

$$d \rightarrow 2 - 2\rho$$

2. По таблице Дарбина – Уотсона находятся значения d_L , d_U (верхняя и нижняя границы для критических значений, определяется для выбранного уровня значимости по числу наблюдений)

Тест Дарбина - Уотсона

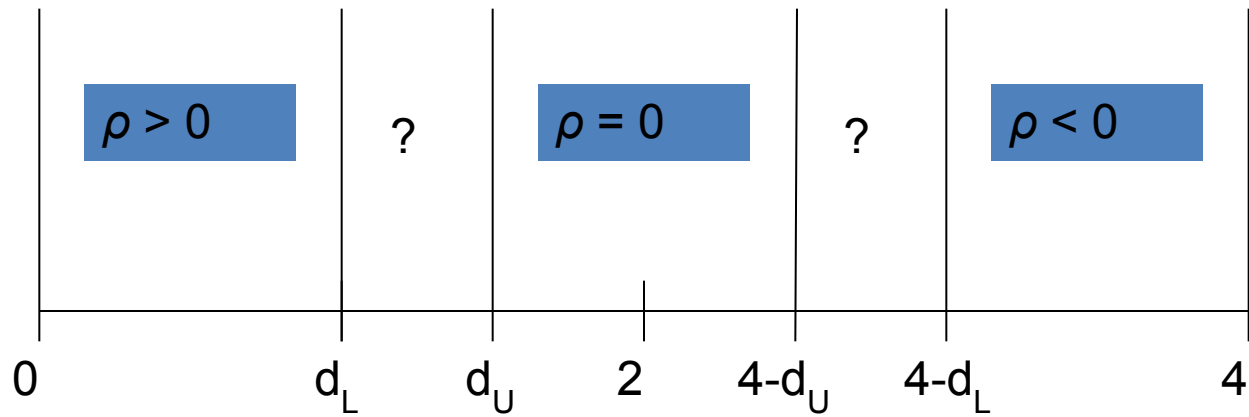
3. Сравниваем расчетное с табличным. Если

$$d \in [0, d_L) \Rightarrow \rho > 0$$

$$d \in (d_L, 2] \Rightarrow \rho = 0$$

$$d \in [d_L, d_U] \Rightarrow ???$$

Тест Дарбина - Уотсона



Способы корректировки автокорреляции

- Использование обычного МНК с коррекцией стандартных ошибок по методу Ньюи-Веста
- Использование обобщенного МНК
- Нелинейный МНК

Гетероскедастичность

- Нарушение независимости дисперсии возмущений от номера (момента) наблюдений (предпосылки A4)
- Причины:
 - Неоднородность исследуемых объектов
 - Характер наблюдений
- Последствия те же, что и в случае автокорреляции

$$\sigma^2 \mathbf{\Omega} = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix}$$

Обнаружение гетероскедастичности

- Тест Голдфельда-Квандта
- Предпосылки теста:
 - Пропорциональность дисперсии случайного возмущения величине некоторого регрессора X_j
 - Случайное возмущение распределено нормально и не подвержено автокорреляции

Тест Голдфельда-Квандта

1. Упорядочиваются выборочные данные по величине модуля регрессора $|X_{jt}|$, относительно которого есть подозрение на гетероскедастичность
2. Отбрасываются r центральных наблюдений
3. По первым и последним n' ($k+1 < n' = (n-r)/2$) данным выборки оцениваются две частные регрессии и вектора остатков e_1 и e_2 соответственно
4. По остаткам частных регрессии вычисляются RSS_1 и RSS_2
5. Вычисляются
 $F_{\text{стат}} = RSS_1 / RSS_2$ и $F_{\text{стат}}^{-1} = RSS_2 / RSS_1$

Тест Голдфельда-Квандта

5. По таблице определяется

$$F_{кр}(\gamma, n-k-1, n-k-1)$$

6. $H_0: \text{Var}(\varepsilon_i) = \sigma^2$ (гомоскедастичность) не отвергается, если выполняются оба неравенства

$$F_{\text{стат}} \leq F_{кр}$$

$$F_{\text{стат}}^{-1} \leq F_{кр}$$

В противном случае делается вывод о гетероскедастичности случайных возмущений

Методы устранения последствий гетероскедастичности

- Вывод **альтернативной** оценки, которая при предположении (A4') является наилучшей несмещенной оценкой
- Сохранение МНК-оценки, но с **коррекцией** стандартных ошибок
- Пересмотр **спецификации** исходной модели

Взвешенный метод наименьших квадратов

1. Выявляется гетероскедастичность (непостоянство дисперсии остатков)
 - a) графически
 - b) с помощью различных тестов
2. В соответствии с поведением остатков подбираются «веса»
3. Каждое наблюдение «взвешивается» и оценивается преобразованная модель

Линейные и нелинейные модели

1. Линейная зависимость – Y линейна и по переменной X и по параметрам β_1 и β_2
2. Нелинейная зависимость 1 типа - Y нелинейна по переменной X , но линейна по параметрам β_1 и β_2
3. Нелинейная зависимость 2 типа - Y нелинейна по переменной X и нелинейна по параметрам β_1 и β_2

Линейные и приводимые к
линейным (нелинейные 1-го типа)

$$Y = \beta_1 + \beta_2 X$$

$$Y = \beta_1 + \beta_2 \frac{1}{X}$$

$$Y = \beta_1 X^{\beta_2}$$

Линеаризация производственной функции (эконометрическая модель модель)

$$Y_t = \alpha_0 \cdot K_t^\alpha \cdot L_t^{1-\alpha} \cdot v_t$$

$$\ln\left(\frac{Y_t}{L_t}\right) = \gamma + \alpha \cdot \ln\left(\frac{K_t}{L_t}\right) + \varepsilon_t$$

$$\gamma = \ln(\alpha_0), \quad \varepsilon_t = \ln(v_t)$$

Случайное возмущение ε_t удовлетворяет условиям Гаусса-Маркова
Получено уравнение для выпуска на одного работника в зависимости от затрат капитала на одного работника

Параметр α – эластичность выпуска по капиталу

Полулогарифмическая модель

- Показательная функция может быть приведена к линейному виду логарифмическим преобразованием
- Исходная модель

$$Y = \beta_1 e^{X\beta_2} v$$

Здесь v – случайная составляющая

- Преобразованная модель

$$\ln Y = \ln \beta_1 + \beta_2 X + \varepsilon$$

Здесь $\varepsilon = \ln v$

Полулогарифмическая модель

- Коэффициент регрессии β_2 в этом случае интерпретируется как *относительное* изменение Y в расчете на единицу *абсолютного* изменения X
- Если говорится о процентном изменении Y , то оценку коэффициента β_2 нужно умножить на 100
- Такая интерпретация справедлива при малых β_2 ($<0,1$)

ТЕСТЫ НА ФУНКЦИОНАЛЬНУЮ ФОРМУ

- 1. RESET тест Рамсея
- 2. J-тест Дэвидсона и МакКиннона
- 3. PE-тест МакКиннона
- 4. Метод Зарембки
- 5. Тест Кокса-Бокса

Тест Рамсея

- RESET Test – Regression Specification Error Test– тест на спецификацию (Ramsey, 1969)
- Тестируется спецификация

$$Y = X\beta + \varepsilon$$

на ошибки следующих типов:

- Пропущенные переменные: матрица X содержит не все нужные переменные
- Неправильная функциональная форма: переменные должны быть прологарифмированы, или взяты в квадрат и пр.
- Ненулевая корреляция между регрессорами и случайным фактором, вызванная, например,
 - Ошибками измерений
 - Одновременностью (объем продаж и затраты на рекламу)
 - Присутствием в уравнении лагированных значений эндогенной переменной при автокорреляции остатков

Тест Рамсея

$$Y = X\beta + \varepsilon$$

- При таких ошибках спецификации МНК-оценка будет смещенной и несостоятельной
- Рамсей показал, что в этом случае случайный фактор имеет ненулевое математическое ожидание
- Тест Рамсея тестирует нулевую гипотезу

$$H_0 : \varepsilon \sim N(0, \sigma^2 \mathbf{I})$$

Против альтернативной

$$H_1 : \varepsilon \sim N(\mu, \sigma^2 \mathbf{I})$$

Тест Рамсея

- Идея проверки: нелинейные функции $\hat{Y} = X\beta$ не должны улучшать объяснение поведения

- Алгоритм теста

- Оценивается уравнение $Y = X\beta + \varepsilon$

- Определяются расчетные значения \hat{Y}

- Оценивается вспомогательная регрессия

$$Y = X\beta + \alpha_1 \hat{Y}^2 + \alpha_2 \hat{Y}^3 + \alpha_4 \hat{Y}^4 + \dots + \varepsilon'$$

- Проверяется гипотеза, что все коэффициенты при степенях расчетных \hat{Y} равны нулю одновременно