Анализ данных. Лекция. Неделя 1-2. Линейная модель со стохастическим регрессором.

Филиппов Роман

5 октября 2022г.

Contents

1 Предпосылки **2**

2 Тестирование гипотез и построение доверительных интервалов **3**

3 Проблемы спецификации **4**

4 Инструментальные переменные. **5**

1 Предпосылки

Модель ЛМСР предполагает следующие пять предпосылок:

1. Модель линейна по своим параметрам:
$$y_i = \beta_0 + \beta_1 * x_i^{(1)} + ... + \beta_k * x_i^{(k)} + \varepsilon_i, i = 1, ..., n; \quad (1)$$
2. Наблюдения $\{(x_i^1, ..., x_i^k, y_i), i = 1, ..., n\}$ независимы и одинаково распределены;
3. $x_i^1, ..., x_i^k$ и y_i имеют конечные ненулевые четвертые моменты распределения
4. Случайнные ошибки имеют нулевое математическое ожидание при заданном x_i: $E(\varepsilon_i|x_i) = 0$;
5. Между переменными отсутствует чистая мультиколлинеарность;

- В классической линейной модели мы предполагали неслучайность(детерминированность) регрессоров, гомоскедастичность и нормальность случайных ошибок;
- Теперь мы отказались от всех этих предположений, чтобы получить реалистичную модель. В реальности гомоскедастичности в данных обычно нет. Используйте состоятельные в условиях гетероскедастичности (робастные) стандартные ошибки. Также в реальности обычно трудно гарантировать нормальность случайных ошибок. Для тестирования гипотез используйте асимптотические результаты. Детерминированность регрессоров тоже являлась упрощающей предпосылкой, которая не всегда удобна на практике;
- Вторая предпосылка обычно верна для пространственных данных но нарушается для временных рядов;
- Третья предпосылка подразумевает отсутствие очень больших выбросов, она требуется для того, чтобы оценки коэффициентов имели асимптотически нормальное распределение(это позволит тестировать гипотезы). Ее выполнение сложно проверять, но она слабая, поэтому оно пустулируется;
- Четвертая предпосылка является ключевой: если она не выполняется, то нельзя говорить о причинно-следственных связях. Она говорит о том, что прочие факторы, содержащиеся в ошибке, не связаны с регрессорами. Также, из предпосылки следует:
  1. $E(\varepsilon_i) = 0$
  2. $cov(x_i, \varepsilon_i) = 0$

Опр. Регрессор называется экзогенным, если $cov(x_i, \varepsilon_i) = 0$. В противном случае, он называется эндогенным.

Теорема Если выполнены предпосылки 1-5, то МНК оценки коэф-ов $\beta_i, i = 1, ..., n$ состоятельны и асимптотически нормальны.

Состоятельность, по определению, говорит нам о том, что при увеличении числа наблюдений мы будем приближаться к верному ответу. Асимптотическая нормальность позволяет строит доверительные интервалы и тестировать гипотезы.

2 Тестирование гипотез и построение доверительных интервалов

Тестирование гипотезы $H_0 : \beta_k = A$ при $n \to \infty$: $t = \frac{\hat{\beta}_k - A}{s.e.(\hat{\beta}_k)} \sim N(0,1)$ Тестирование осуществляется стандартным образом, при этом из свойств нормального распределения известно, что:

- критическое значение для уровня значимости 5% равно 1,96;
- $P_{value} = 2 * \Phi(-|\frac{\hat{\beta}_j}{s.e.\hat{\beta}_j}|)$, где Ф-функция нормального стандартного распределения;

Тогда 95%-процентный доверительный интервал для $\hat{\beta}_k$ определяется следующим образом: $(\hat{\beta}_k - 1.96 * s.e.\hat{\beta}_j; \hat{\beta}_k + 1.96 * s.e.\hat{\beta}_j)$.

Тестирование линейных ограничений: F-test.

- В условиях гомоскедастичности можно использовать все стандартные варианты F-теста, которые мы рассматривали ранее;
- Единственное отличие состоит в том, что если верна нулевая гипотеза данного теста, то расчетное значение тестовой статистики имеет распределение Фишера с q, ∞ степеней свободы $F(q, \infty)$, так как $n \to \infty$;
- Поэтому для получения критических значений следует использовать таблицу распределения Фишера $F^{\alpha}(q, \infty)$ или таблицу распределения Хи-квадрат,так как случайная величина с распределением $\chi^2(q)$ в q раз больше случайной величины с распределением $F(q, \infty)$.

Тестирование линейных ограничений: Тест Вальда.

В условиях гетероскедастичности лучше использовать не F-test, а тест Вальда. Тестируемая гипотеза: $H * \beta = r$, где

- β-вектор коэффициентов модели;
- H-матрица размера q на k, r- вектор столбец длины q;
- q-количество тестируемых ограничений, k-число коэффициентов в модели;

Расчтеное значение тестовой статистики теста Вальда:

$(H * \hat{\beta} - r)'(H * (X' * \Omega^{-1} * X)^{-1} * H')^{-1}(H * \hat{\beta} - r)$, где Ω-ковариационная матрица ветора случайных ошибок. Если верна тестируемая гипотеза, то эта величина имеет распределение Хи-квадрат с q степенями свободы.

- На практике ковариационная матрица Ω обычно не известна. Поэтому в формуле тестовой статистики её заменяют оценкой $\hat{\Omega}$;
- В частности, если выполнены предпосылки линейной модели со стохастическими регрессорами о независимости отдельных наблюдений, то $\hat{\Omega}$ будет диагональной матрицей, где на главной диагонали стоят оценки дисперсий случайных ошибок;
- Современные эконометрические пакеты осуществляют все расчеты, необходимые для теста Вальда, автоматически;

3 Проблемы спецификации

Типичные проблемы эндогенности.

Если предпосылка об экзогенности регрессора выполнена, то обычный МНК дает состоятельные результаты. Как понять, выполнена ли она? Для этого стоит знать типичные ситуации, в которых она нарушается. То есть, знать типичные причины эндогенности:

1. Эндогенность из-за пропуска существенной переменной;
2. Эндогенность из-за выбора неверной функциональной зависимости;
3. Эндогенность из-за двусторонней причинно-следственной связи;
4. Эндогенность из-за ошибок измерения;

Несостоятельность из-за пропуска решения.

Опр.Переменная интереса- фактор, влияние которого на зависимиую переменную нас интересует.

Опр.Контрольные переменные- переменные, которые мы включаем в модель для того, чтобы избежать смещения коэффициента при интересующей нас переменной.

При пропуске важной переменной МНК оценки коэффициентов параметров становятся смещенными и в большинстве случаев несостоятельными. Решением является добавление переменной в модель, если она наблюдаема. Но тут возникает вопрос существенности переменной, ведь если добавить несущественную переменную, то:

- Коэффициенты при прочих переменных остаются несмещенными и состоятельными;
- Из-за необходимости оценивать большее количество коэффициентов, а также вероятной мультиколлинеарности увеличивается дисперсия оценок коэффициентов, то есть понижается точность модели.

Поэтому, существуют **критерии для включения переменной в модель:**

- Роль переменной в уравнении опирается на прочные теоретические основания. Ну или хотя бы на здравый смысл;
- Переменная статистически значима;
- Оценки других коэффициентов сильно меняются при включении новой переменной в модель. Это значит, что до этого они страдали от смещения из-за пропуска существенной переменной. Теперь вы эту существенную переменную добавили, и смещение пропало;
- Скорректированный R-квадрат значительно увеличивается в результате включения переменной в модель;

Если переменная ненаблюдаема, то это следует решать следующими методами:

- Используйте замещающие переменные. Замещающей переменной называется переменная, которая сильно коррелирована с ненаблюдаемой существенной переменной, и которая при этом является наблюдаемой;
- Используйте инструментальные переменные;
- Используйте модель с фиксированными эффектами;
- Используйте контролируемый эксперимент или квазиэксперимент;

Несостоятельность из-за неверной формы функциональной связи.
Рассмотрение неверной формы функциональной связи эквивалентено пропуску существенной переменной и приводит к таким же последствиям. Решением является рассмотрение корректной формы связи. Для ее выявление могут быть полезны следующие шаги:

- Осуществите графический анализ исходных данных и графический анализ остатков оцененного уравнения регрессии;
- Опирайтесь экономическую теорию или другие содержательные соображения по поводу природы анализируемых переменных;
- Используйте формальные статистические критерии. Например, тест Рамсея;

Несостоятельность из-за двусторонней причинно-следственной связи.
Здесь возникают те же проблемы, решаемые следующим образом:

- Используйте инструментальные переменные;
- Используйте контролируемый эксперимент или квазиэксперимент;
- Используйте методы, предназначенные для оценивания систем одновременных уравнений, например SVAR и VECM;

Ошибки измерения регрессора.
Возможные пути решения проблемы:

- Используйте инструментальные переменные;
- Если проблема ошибок измерения стоит не слишком остро, в прикладных исследованиях ее иногда игнорируют (действительно, если коэффициент оказался статистически значим даже в условиях ошибок измерения, то после устранения проблемы он тем более должен оказаться значим);

4 Инструментальные переменные.

Как мы выяснили, если переменные эндогенны, то МНК оценки несостоятельны. Тогда нужно найти переменную такую z,которая будет с одной стороны экзогенна($cov(z_i, \varepsilon_i) = 0$), с другой релевантна, т.е. кореллирована с эндогенной переменной($cov(z_i, x_i) = 0$). Такую переменную мы называем **валидной**, она позволяет нам использовать двухшаговый

МНК для построение состоятельной оценки коэффициентов. Если мы имеем следующую модель ЛМСР:

$$y_i = \beta_0 + \beta_1 * x_i^{(1)} + ... + \beta_k * x_i^{(k)} + \beta_{k+1} * w_i^{(1)} + ... + \beta_{k+r} * w_i^{(r)} + \varepsilon, \tag{2}$$

где $x_i^{(1)}, ..., x_i^{(k)}$-эндогенные регрессоры, $w_i^{(1)}, ..., w_i^{(r)}$-экзогенные, $z_i^{(1)}, ..., z_i^{(m)}$-инструментальные переменные(причем необходимо $m \geq k$), то двухшаговый МНК состоит из следующих шагов:

1. Сначала для каждой из переменных $x_i^{(1)}, ..., x_i^{(k)}$ мы оцениваем регрессию на константу, инструментальные переменные $z_i^{(1)}, ..., z_i^{(m)}$ и экзогенные переменные $w_i^{(1)}, ..., w_i^{(r)}$, получаем прогнозные значения $\hat{x_i^1}, ..., \hat{x_i^k}$;

2. Затем мы оцениваем для y_i регрессию на константу, предсказанные значения $\hat{x_i^1}, ..., \hat{x_i^k}$ и экзогенные переменные $w_i^{(1)}, ..., w_i^{(r)}$;

3. В итоге, получаем

$$y_i = \beta_0 + \beta_1 * \hat{x_i^{(1)}} + ... + \beta_k * \hat{x_i^{(k)}} + \beta_{k+1} * w_i^{(1)} + ... + \beta_{k+r} * w_i^{(r)} + \varepsilon \tag{3}$$

При этом, стоит понимать, что мы давно забыли об эффективности в нашей модели, хоть оценка и является состоятельной. То есть, при маленькой выборке 2МНК, как правило, дает плохие результаты.

Тестирование гипотез.

В случае использования 2МНК дисперсии оценок коэффициентов будут отличаться от случая обычного МНК. Например, для парной регрессии $y_i = \beta_1 + \beta_2 * x_i + \varepsilon_i$ с единственным инструментом z корректная формула для расчета стандартной ошибки коэффициента при регрессоре имеет вид:

$$s.e.(\hat{\beta_2}) = \sqrt{\frac{S^2}{\Sigma(x_i - \overline{x})^2} * \frac{1}{corr(\hat{x}, z)^2}}, S^2 = \frac{\Sigma e_i^2}{n-2} \tag{4}$$

Опр.Инструмент называется слабым, если он объясняет малую долю дисперсии эндогенной переменной.

Соответственно, если инструменты слабые, то:

- Точность 2МНК становится низкой;
- Результаты на значимость могут быть некорректны, т.к.распределение оценки не является асимптотически нормальным;

Если в нашей модели ЛМСР присутствует ровно одна эндогенная переменная и определенное количество экзогенных:

$$y_i = \beta_0 + \beta_1 * x_i + \beta_2 * w_i^{(1)} + ... + \beta_{1+r} * w_i^{(r)} + \varepsilon, \tag{5}$$

то с помощью разложение этого x_i на инструментальные переменные $z_i^{(1)}, ..., z_i^{(m)}$ и проверки гипотезы о том, что все коэф-ы при этих инструментах равны нулю, можно понять, являются ли инструменты слабыми. Для этого мы считаем соответствующую F-статистику, если она больше 10, то инструменты релевантны(иначе слабые).

Что делать для решения проблемы слабых инструментов?

- Следует попытаться найти инструменты получше;
- Если в модели слишком много инструментов, то следует отказаться от некоторых из них;

Допустим, наши инструменты оказались релевантными, тогда следует их проверить на экзогенность. В этом нам поможет **Sargan test**(необходимо m>k):

Нулевая гипотеза:все инструменты экзогенны. e_i-остатки, полученные в ходе использования 2МНК. Оценим вспомогательную регрессию:

$$e_i = \alpha_0 + \alpha_1 * z_i^{(1)} + ... + \alpha_m * z_i^{(m)} + \alpha_{m+1} * w_i^{(1)} + ... + \alpha_{m+r} * w_i^{(r)} + \varepsilon_i \tag{6}$$

$J_{statistic} = m * F \sim \chi^2(m-k)$, где F- расчетное значение F-статистики для гипотезы $\alpha_0 = ...\alpha_m = 0$. Если нулевая гипотеза отклоняется, то, по-крайней мере, некоторые из инструментов не экзогенны.