Data Science.
Lectures. Week 1.
Введение в эконометрику.

Матевосова Анастасия

7 октября 2022 г.

Contents

1 Что такое эконометрика?

- **Эконометрика** - это наука, которая изучает количественные и качественные экономические взаимосвязи с помощью математических и статистических методов и моделей.
- **Эконометрика** - самостоятельная научная дисциплина, объединяющая совокупность теоретических результатов, методов и приёмов, позволяющих на базе экономической теории, экономической информации и математико-статистического инструментария придавать конкретное количественное выражение общим (качественным) закономерностям, обусловленным экономической теорией.

С. Айвазян

- **3 составные части эконометрики:**
  1. *Теоретико-методологическая* - Экономическая теория
  2. *Информационная* - Экономические данные
  3. *Инструментальная* - Методы обработки данных
- *Зачем нужна эконометрика?*

  Основные цели и задачи эконометрических методов:
  1. Задача прогнозирования
     Пример: прогноз основных экономических показателей
  2. Задача построения моделей для экономических систем
     Пример: имитация различных возможных сценариев социально-экономического развития страны.
  3. Дескриптивный анализ
     Пример: конкретный статистический анализ рынка

2 Типы данных в эконометрике

- **3 типа данных в эконометрике:**
  1. Пространственные данные
  2. Временные ряды
  3. Панельные данные
- **Показатели (переменные):**
  1. объясняющие переменные (регрессоры)
  2. зависимые или результирующие переменные
- **Пространственные данные:** в один и тот же момент времени (или промежуток времени) данные снимаются с случайно выбранных объектов. Т.е. объекты рассматриваются в пространстве.

  Есть генеральная совокупность объектов. Каждый объект характеризуется набором показателей (переменных). Из генеральной совокупности извлекаются объекты случайным образом и получается случайная выборка $\xi_1,...,\xi_n$

  С i-го объекта снимаются значения объясняющих переменных и результирующей переменной. i=1,...,n

  X = $\begin{pmatrix} x_i^{(1)} \\ x_i^{(2)} \\ \vdots \\ x_i^{(k)} \end{pmatrix}$ - набор показателей, который играет роль объясняющих переменных

y_i - результирующая переменная (случай одной результирующей переменной)

Результатом является массив данных M=$\{(X_i, y_i),\ i = 1, .., n\}$

- **Временные ряды:** имеем только один объект, снимаем значения показателей, характеризующих этот объект в последовательные моменты времени.

  Результатом является массив данных M=$\{(X_t, y_t),\ t = 1, .., T\}$ - этот массив представляет собой совокупность временных рядов.

- **Принципиальная разница между пространственными данными и временными рядами:**

- В случае пространственных данных как правило считаем, что данные собраны независимо друг от друга. Это случайная выборка. Результаты наблюдений за следующими объектами никак связаны с результатами, полученными в предыдущих наблюдениях. Т.е. в пространственных выборках были независимые одинаково распределённые случайные величины.

- В случае временного ряда: существенная связь между предыдущими и следующими наблюдениями. Случайные величины зависимы и их законы распределения меняются во времени.

- **Панельные данные данные:** Смесь первых двух типов данных. Имеем n случайно выбранных объектов, каждый из которых наблюдается в течение некоторой последовательности моментов времени.

  С i-го объекта в t-ый момент времени снимаем значения показателей: M = $\{(X_{it}, y_{it}),\ i = 1, .., n;\ t = 1, ..., T\}$

3 Пространственные данные

3.1 Модель парной регрессии

M=$\{(x_i, y_i),\ i = 1, .., n\}$ - массив - двумерный вектор, значения которого сняты со случайно взятых объектов

x_i - единственная объясняющая переменная, снятая с i-го объекта

y_i - результирующая переменная, снятая с i-го объекта

Предполагаем: y связан с x какой-то зависимостью $\boxed{y_i = f(x_i, \beta) + \varepsilon_i}$

β-список параметров

ε - случайная ошибка (аддитивная случайная компонента, в которой отражены все дополнительные переменные, влияющие на результрующую переменную).

Дополнительные факторы (переменные) - переменные, влияющие на результрующую переменную y помимо объясняющей переменной x, являющиеся менее существенными. Дополнительные переменные ненаблюдаемы, в отличие от x и y.

3.1.1 Линейная модель парной регрессии (ЛМПР)

f является линейной относительно x.

$\boxed{y_i = \beta_1 + \beta_2 x_i + \varepsilon_i}$ - ЛМПР

$(i = 1, ..., n)$

β_1 - свободный коэффициент

β_2 - угловой коэффициент

Надо оценить параметры β_1, β_2. Наиболее распространённый метод-метод наименьших квадратов.

3.1.2 Метод наименьших квадратов

$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$

β_1, β_2 -неизвестны, хотим оценить. Вместо β_1 будем подставлять произвольное число b_1, вместо β_2 подставляем b_2.

Нас интересуют значения b_1, b_2, дающие наиболее хорошую подгонку для неизвестной линейной связи.

Ищем b_1^*, b_2^*-оптимальные значения, которые минимизируют сумму квадратов отклонений

$$\boxed{\sum_{i=1}^{n}(y_i - (b_1 + b_2 x_i))^2 = Q(b_1, b_2) \to \min_{b_1, b_2}}$$

b_1^*, b_2^* - функции от всего массива

$b_1^* = \hat{\beta}_{1,MHK} = \overline{y} - \hat{\beta}_2 \cdot \overline{x}$

$b_2^* = \hat{\beta}_{2,MHK} = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\overline{x^2} - (\overline{x})^2}$

- Считаем, что работаем с данными, которые можно аппроксимировать линейной функцией.

  В модели необходимо учесть все существенные факторы, поэтому нельзя ограничиваться моделью парной регрессии.

3.2 Линейная модель множественной регрессии

M=$\{(X_i, y_i),\ i = 1, .., n\}$ - массив из нескольких объясняющих факторов и одного результирующего. X_i-объясняющие переменные;

y_i-результирующая переменная.

Предполагаем, что связь линейная: $\boxed{y_i = \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + ... + \beta_k x_i^{(k)} + \varepsilon_i}$

Правильность модели выражается в том, что:

- В среднем влияние дополнительных переменных при фиксировванных x отсутствует.

$$\boxed{E(\varepsilon_i | x_i^{(1)}, ..., x_i^{(k)}) = 0}$$

$$\Downarrow$$

$$\boxed{E(y_i | X_i) = \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + ... + \beta_k x_i^{(k)}}$$

- Выбранная модель хороша в том смысле, что в среднем мы угадали, что y линейная функция от $x^{(1)}, \ldots, x^{(k)}$ с ошибками, которые невелируются.

Обозначения:

$E(\varepsilon_i | x_i^{(1)}, ..., x_i^{(k)})$-*условное мат.ожидание* ε_i *при фиксированных значениях всех переменных* $x_i^{(1)}, ..., x_i^{(k)}$

$E(y_i | X_i)$-*условное мат.ожидание переменной* y_i *при условии, что все переменные* $X_i(x_i^{(1)}, ..., x_i^{(k)})$ *зафиксированы на каких-то уровнях.*

4 Объясняющие переменные (регрессоры)

- **Объясняющие переменные:**

  1. детерминированные
  2. случайные

- Детерминированные: Контролируемый эксперимент, в котором сами задаём значения x (объясняющей переменной). Т.е. заказываем очередное наблюдение, в котором объясняющая переменная фиксированна так, как мы хотим.

  Структура данных, которые таким образом будут собираться - системы вертикальных точек.

- Случайные: Получаем конкретные данные и не можем управлять значениями x (объясняющей переменной).

  Структура данных - хаотичная, нет вертикальной структуры.

  В этом случае интерпретируем x, как случайную величину (случайный регрессор).

4.1 Детерминированные объясняющие переменные

4.1.1 Свойства оценок $\hat{\beta}$, полученных с помощью МНК для ЛМПР

M_n=$\{(X_i, y_i),\ i = 1, .., n\}$ - массив из n наблюдений (объём выборки = n)

$\hat{\beta}(M_n)$ - оценка для параметра β, как функция от массива. $\hat{\beta}(n)$, т.е. зависит от n.

- Если x_i-детерминированы, то y_i случайны (т.к. в модели есть случайные ошибки и поэтому переменная y является случайной величиной).

  Каждый из $y_1, y_2, ..., y_n$ - случайная величина

  Считаем, что все наборы объясняющих переменных - детерминированы. Т.е. интерпретируются как числа, а не как случайные величины.

  $\hat{\beta}_{1,MHK} = \overline{y} - \hat{\beta}_2 \cdot \overline{x}$

  $\hat{\beta}_{2,MHK} = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\overline{x^2} - (\overline{x})^2} = c_1 y_1 + c_2 y_2 + ... + c_n y_n$

  $c_i = \frac{\frac{1}{n}(x_i - \overline{x})}{\overline{x^2} - (\overline{x})^2} = const$ по $y_1, ..., y_n$

- Обе оценки являются **линейными функциями относительно** y

- **Свойства хороших оценок:**

- **Несмещённая**: $\boxed{E\hat{\beta}_{2,MHK} = \hat{\beta}_2 \quad \forall n}$ (для МНК существуют условия, при которых это свойство будет верным)

- **Состоятельная**: $\boxed{\hat{\beta}(M_n) \xrightarrow[n\to\infty]{P} \beta}$ (сходимость по вероятности)

  сходимость по вероятности=состоятельность : $\boxed{\forall\delta \quad \mathrm{P}\{|\hat{\beta(n)} - \beta| > \delta\} \xrightarrow[n\to\infty]{} 0}$

  Состоятельность - ассимптотическое свойство. Поэтому когда наблюдений немного, оценка может очень сильно отклоняться от оцениваемого параметра. При большом объёме выборки с большой уверенностью можно сказать, что состоятельная оценка окажется очень хорошим аналогом для значения параметра β

- **Эффективная (оптимальная)**:

  1. Относительная эффективность: Из двух несмещённых оценок одна относительно эффективнее другой, если её дисперсия не больше дисперсии другой.
     $\hat{\beta}$ относительно эффективнее $\widetilde{\beta}$, если $D\hat{\beta} \le D\widetilde{\beta} \quad (E(\hat{\beta} - \beta)^2 = D\hat{\beta})$
  2. Абсолютная эффективность: Эффективность в классе

     Рассматривается класс несмещённых оценок $K = \{\widetilde{\beta} : E\widetilde{\beta} = \beta\}$. В этом классе одна конкретная оценка $\hat{\beta}$ называется эффективной, если $\forall\widetilde{\beta} \in K \ \ D\hat{\beta} \le D\widetilde{\beta}$

Несмещённость и эффективность рассматриваются при фиксированном n. Состоятельность - это асимптотическое свойство (т.е. при $n \to \infty$).

4.1.2 Условия, при которых МНК даёт хорошие оценки. Теорема Гаусса-Маркова.

Если

- $E\varepsilon_i = 0 \qquad i = 1, ..., n$
- $D\varepsilon_i = \sigma^2 = const \qquad i = 1, ..., n$ (одинаковая, не зависящая от номера наблюдения)
- $cov(\varepsilon_i, \varepsilon_j) = 0 \qquad i \neq j;\ i = 1, ..., n;\ j = 1, ..., n,$

то МНК-оценки коэффициентов β_1, β_2 в ЛМПР будут:

- несмещёнными;
- эффективными в классе всех несмещённых, линейных по $y_1, ..., y_n$ оценок.

Т.е. для β_2 рассматриваем класс $K_2 = \{\widetilde{\beta_2} = c_1 y_1 + ... + c_n y_n$, где $c_1, .., c_n$ не зависят от y, при этом $E\widetilde{\beta_2} = \beta\}$.

Теорема Гаусса-Маркова утверждает, что $\hat{\beta}_{2,MHK}, \in K_2$ и $D\hat{\beta}_{2,MHK} \leq D\widetilde{\beta_2} \quad \forall \widetilde{\beta_2} \in K_2$.

То же самое верно для $\hat{\beta}_{1,MHK}$

Эти МНК-оценки называются BLUE - Best Linear Unbiased Estimator.

Для сформулированной теоремы: Рассматриваются ЛМПР, считая, что эта модель правильная и что нет других существенных переменных, кроме , которые влияют на y.
$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ (т.е. безошибочно задали спецификацию модели).

В случае использования ЛМПР и неучёта существенных переменных, эти МНК оценки не будут обладать хорошими свойствами.

4.1.3 ЛММР

Три формы записи:

1. $y_i = \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + ... + \beta_k x_i^{(k)} + \varepsilon_i \qquad i = 1, ..., n$
2. Векторная форма: $y_i = \beta^T x_i + \varepsilon_i \qquad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \qquad x_i = \begin{pmatrix} x_i^{(1)} \\ \vdots \\ x_i^{(k)} \end{pmatrix}$
3. Векторно-матричная форма: $y = X\beta + \varepsilon \qquad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \qquad X = \begin{pmatrix} x_1^{(1)} & x_1^{(2)} ... x_1^{(k)} \\ x_2^{(1)} & x_2^{(2)} ... x_2^{(k)} \\ & \vdots \\ x_n^{(1)} & x_n^{(2)} ... x_n^{(k)} \end{pmatrix}$

   $y(n \times 1)$-набор всех наблюдений за переменной y
   $\beta(k \times 1)$
   $X(n \times k)$ -матрица плана (i-ая строчка - наблюдения, снятые с i-го объекта)
   X называется **матрицей плана**, потому что если рассматривается управляемый планируемый эксперимент, когда мы сами назначаем значения x, то всю матрицу мы должны зафиксировать.
   (X, y) - массив M

4.1.4 Вектор МНК-оценок

$\hat{\beta}_{MHK}=\begin{pmatrix}\hat{\beta}_{1,MHK}\\ \hat{\beta}_{2,MHK}\\ \vdots\\ \hat{\beta}_{k,MHK}\end{pmatrix}$ - вектор МНК-оценок

Зафиксируем матрицу X, получим случайным образом реализации на всех n объектах, тем самым мы получили массив данных. Затем мы можем повторить процедуру получения массива, не меняя матрицу X, но получая новую реализацию вектора y. Таких реализаций можно сделать много и получить огромное количество массивов с одним и тем же X и разными y.
$\hat{\beta}_{i,MHK}$ является функцией от массива М.
$M=(X,y)$, где X-фиксирован, y-случайный (меняется)
Возьмём большое количество реализаций массива (например, 10^6), будут получаться разные векторы МНК-оценок параметров.
Для 10^6 реализаций мы получим 10^6 реализаций случайной величины $E\hat{\beta}_{i,MHK}$.
В соотвествии с теоремой Гаусса-Маркова: $E\hat{\beta}_{i,MHK}=\beta_i$ (говорит о том, что оценка хорошая).

Как посчитать вектор МНК-оценок?
$\boxed{\hat{\beta}_{MHK}=(X^TX)^{-1}X^Ty}$
По умолчанию требуется условие $rank(X)=k$

- Всегда ли существует $(X^TX)^-1$?
$(X^TX)^-1$ не существует, когда $det(X^TX)=0\Leftrightarrow rank(X)<k$
Это означает, что одна из объясняющих переменных $x^{(1)}...x^{(k)}$является линейной комбинацией остальных. А это означает, что нет смысла включать в модель все k переменных: ту, которая линейно выражается через остальные можно выкинуть, тем самым сократив количество регрессоров и сделав полный ранг.

Свойства вектора МНК-оценок.

1. МНК оценка является **линейной оценкой по** y:
$A=(X^TX)^{-1}X^T$
$\hat{\beta}_{MHK}=Ay$
2. **Несмещённость**: $E\hat{\beta}_{MHK}=\beta$
3. **Теорема Гаусса-Маркова**: В классе линейных по y и несмещённых оценок МНК оценки имеют наименьшую дисперсию.
В классе K_j=$\{\widetilde{\beta}_j=c_1y_1+...+c_ny_n$ - несмещённые$\}$ $\forall\widetilde{\beta}_j\in K_j$ $D\hat{\beta}_{j,MHK}\le D\widetilde{\beta}_j$

4.1.5 Коэффициент детерминации R^2

Меры разброса:

- $\boxed{TSS=\sum_{i=1}^{n}(y_i-\overline{y})^2}$ -общая сумма квадратов

- $\boxed{\frac{TSS}{n} = \hat{Var}(y)}$ - выборочная дисперсия

Рассматриваем модель $y = X\beta + \varepsilon$, применяем МНК $\Rightarrow$ получаем $\hat{\beta}$
$y - X\hat{\beta} = \hat{\varepsilon} = e$ - вектор оценки случайной ошибки Модель $y_i = \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + ... + \beta_k x_i^{(k)} + \varepsilon_i$

Применили МНК и получили: $\hat{\beta}_1, ..., \hat{\beta}_k$
$\hat{y}_i = \hat{\beta}_1 x_i^{(1)} + \hat{\beta}_2 x_i^{(2)} + ... + \hat{\beta}_k x_i^{(k)}$
$\hat{y}_i$ - оценка y_i, рассчитанное значение y на i-ом объекте.
y_i-наблюдаемая реализация случайной величины y на i-ом объекте.
$y_i - \hat{y}_i = \hat{\varepsilon}_i = e_i$
ε_i-регрессионная случайная ошибка
$\hat{\varepsilon}_i = e_i$ - регрессионный остаток (является случайной величиной, так как y случайны).

Разница между случайной ошибкой (ε_i) и остатком ($\hat{\varepsilon}_i$) состоит в том, что:
случайная ошибка (ε_i) - ненаблюдаемая случайная величина,
остаток ($\hat{\varepsilon}_i$) - вычисленная, то есть наблюдаемая случайная величина.

$TSS = \sum\limits_{i=1}^{n} (y_i - \overline{y})^2$

$\overline{\hat{y}} = \frac{1}{n} \sum\limits_{i=1}^{n} (\hat{y}_i) = \overline{y}$
$RSS = \sum\limits_{i=1}^{n} (\hat{y}_i - \overline{\hat{y}})^2 = \sum\limits_{i=1}^{n} (\hat{y}_i - \overline{y})^2$

- $\boxed{RSS = \sum\limits_{i=1}^{n} (\hat{y}_i - \overline{y})^2}$ - Regression sum of squares
- $\boxed{ESS = \sum\limits_{i=1}^{n} (y_i - \hat{y}_i)^2 = e_i^2}$ - Error sum of squares

$ESS = \frac{1}{n} \sum\limits_{i=1}^{n} (e_i - \overline{e})^2 = \frac{1}{n} \sum\limits_{i=1}^{n} e_i^2$
$e_i = y_i - \hat{y}_i$ - число
$e = y - \hat{y}$ - вектор
$\overline{e} = \overline{y} - \overline{\hat{y}} = 0$ - число

TSS-мера разброса y вокруг своего центра;
RSS-мера разброса $\hat{y}$;
ESS-мера малости остатков.

Теорема
Если модель содержит свободный коэффициент, т.е. записывается в виде:
$y_i = \beta_1 + \beta_2 x_i^{(2)} + ... + \beta_k x_i^{(k)}$, т.е. $x_i^{(1)} \equiv 1$,
то $TSS = RSS + ESS$

Таким образом, TSS разбивается на две части RSS и ESS, одна в другую перетекает.
Если остатки очень маленькие, то ESS маленькая величина (хорошее качество модели), значит доля RSS в общей сумме большая.
Поэтому вводится мера качества ЛММР: $R^2 = 1 - \frac{ESS}{RSS}$ - чем больше, тем лучше

$\boxed{\mathrm{R}^2 = 1 - \frac{ESS}{RSS}}$ - **коэффициент детерминации** - мера качества линейной модели множественной регрессии

Свойства R^2 (в рамках теоремы - т.е. при наличии свободного коэффициента):

- $0 \leqslant R^2 \leqslant 1$
  Для моделей без свободного коэффициента может быть неверным.

- $R^2 = 1 \Leftrightarrow ESS = 0 \Leftrightarrow$ все $e_i = 0 \Leftrightarrow$ все $y_i = \hat{y}_i$
  $R^2 = 1$ означает, что никаких случайных ошибок нет, т.е. абсолютно точная подгонка с помощью линейной функции.
  Если $R^2 \approx 1 \quad (0,8 - 0,99)$ - это говорит о хорошей подгонке модели, т.е. ошибки очень маленькие, **качество модели хорошее**.
  Если облако точек подгоняется с помощью прямой линии очень хорошо, то это значит, что размазанность вокруг этой прямой линии маленькая.

- $R^2 = 0$: $1 - \frac{ESS}{TSS} = 0$
  $TSS = ESS$
  $RSS = 0$
  Т.е. горизонтальное облако. Нет связи. Свидетельствует о **плохой модели**

Таким образом, R^2 является мерой качества модели: чем ближе он к 1, тем лучше.