Econometrics

Лекция 1. Общая информация и введение

Линейная регрессия и метод наименьших квадратов, МНК(OLS)

Из Марно Вербника

- $X = x_1, x_2, ..., x_n$ - вектор параметров с размерностью (m, n).
- Y - вектор значений с размерностью (n, 1)

Вопрос на который отвечает метод наименьших квадратов звучит следующим образом: **Какая линейная комбинация X с константой дает хорошую аппроксимацию для Y?**

- Для того, что бы ответить на этот вопрос запишем произвольную линейную комбинацию: $\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_n x_n$, где $\hat{\beta}_n$ - константы, которые мы в будующем подберем.
- Разность между наблюдаемым значением и его линейной аппроксимацией: $y_i - [\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_n x_{in}]$
- Следовательно мнимизировать будем функцию квадратов разностей

$$S(\hat{\beta}) = \sum_{i=1}^{n} (y_i - x_i^{'}\hat{\beta})^2$$

- Отсюда следует, что решениями проблемы минимизации является одно из двух следующих уравнений:

$$b = (\sum_{i=1}^{n} x_i x_i^{'})^1 \sum_{i=1}^{n} x_i y_i$$

$$\beta = (X'X)^{-1}(X'Y)$$

Распределения
$Z \sim N(0,1)$

$$\sum Z_i^2 \sim \chi_n^2$$

$E(\chi_n^2) = n, Var(\chi_n^2) = 2n$

$$\frac{Z}{\sqrt{\frac{\chi_n^2}{n}}} \sim t_n$$

$E(t_n) = 0, Var(t_n) = 2n$

$$\frac{\chi_n^2/n}{\chi_m^2/m} \sim F_{n,m}$$

$E(F_{n,m}) = \frac{m}{m-2}$

Лекция 1.2. Повторение ТВ и МС

Функия распределения $F_X(x)$СВ X называется: $F_X(x) = P(X \leq x)$, СВ X называется непрерывной, если существует функция $f(x)$, такая что $F_X^{'}(x) = f(x)$, где $f(x)$ обладает следующими свойствами:

$$f(x) \geq 0$$

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

$$P(a \leq x \leq b) = \int_a^b f(x)dx$$

МО для дискретной СВ: $E(X) = \sum_{i=1}^{N} X_i P_i$, МО для непреревной СВ: $E(X) = \int_{-\infty}^{+\infty} xf(x)dx$
Дисперсия СВ X: $\sigma^2 = Var(X) = E(X - E(X)) = E^2(X) - E(X^2)$, стандартное отклонение - $\sqrt{\sigma^2}$
Ковариация двух СВ X Y $Cov(X, Y) = E[X - E(X)(Y - E(Y)]$
Коэффициент корреляции $r_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$, со следующими свойствами:
1. $|r_{XY}| \leq 1$,
2. При $r_{XY} = 0$ линейная связь СВ X и Y отсутствует полностью,
3. При $r_{XY} = 1$ то между случайными величинами X и Y существует точная линейная связь: Y = aX + b

Лекция 2. Линейная регрессия для одного параметра

Линейный регрессионный анализ объединяет широкий круг задач, связанных с построением зависимостей между двумя переменными: X и Y. X – независимая, объясняющая, экзогенная переменная, регрессор, regressor, Y- зависимая, объясняемая, эндогенная переменная, regressand.

$E(Y|X = x) = f(x)$ – уравнение парной регрессии, из которого следующует следующее: $Y_i = E(Y|X = x_i) + \epsilon_i = f(X_i) + \epsilon_i$, где ϵ стохаистическое возмущение.

$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, i = 1, ..., n$– линейная регрессионная модель, β_0, β_1 параметры, которые необходимо оценить по выборке. $Y = \beta_0 + \beta_1 X$ - линия теоритической регрессии, а $\hat{Y} = \hat{\beta_0} + \hat{\beta_1} 1X$ - линия выборочной регрессии.

RSS - Residual Sum of Squares. OLS((МНК) основан на минимизации RSS. Таким образом задача OLS сводится к следующему:

$$RSS(\hat{\beta_0}, \hat{\beta_1}) \to min$$

Формулы для оценок OLS.

$$\hat{\beta_1} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X^2}} = \frac{\hat{Cov}(X, Y)}{\hat{Var}(X)}$$

$$\hat{\beta_0} = \bar{Y} - \hat{\beta_1}\bar{X}$$

Лекция 3. Линейная регрессия для одного параметра - дисперсионный анализ

Условия для линейной регрессии с константой

1. $\bar{Y} = \hat{\beta_0} + \hat{\beta_1}\bar{X}$ - Линия регрессии проходит через $\bar{X}, \bar{Y}$
2. $\sum e_i = 0$ - Отсутствует систематическая ошибка
3. $\sum Y_i = \sum \bar{Y_i}$ - Сумма всех значений Y совпадает с суммой всех выровненных Y
4. $\bar{Y} = \hat{Y}$ - Среднее арифметическое по всем значениям Y совпадает со средним арифметическим по всех выровненным Y
5. $\sum X_i e_i = 0$ - Векторы X и e ортогональны
6. $\sum \hat{Y_i} e_i = 0$ - Векторы $\hat{Y}$ и e ортогональны

Sums of squares
$\sum(Y_i - \bar{Y})^2$ = TSS (Total sum of squares)
$\sum(\hat{Y}_i - \bar{Y})^2$ = ESS (Explained sum of squares)
$\sum e_i^2$ = RSS (Residual sum of squares)
TSS = RSS + ESS

$$R^2 = \frac{ESS}{TSS} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{Var(\hat{Y})}{Var(Y)} = \frac{TSS - RSS}{TSS}$$

R^2 является отношением ESS к TSS,(или долей дисперсии Y,объясненной с помощью регрессии). Очевидно, это неотрицательная величина.

Критерий лучшей оценки

Best – это оценки с наименьшей дисперсией в классе всех линейных несмещенных оценок.

$$\sigma^2_{\hat{\beta}_0} = \sigma^2_\epsilon \frac{\sum X_i^2}{n\sum x_i^2}, \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2_\epsilon}{\sum x_i^2}$$

Оценка дисперсии возмущений

$$\hat{\sigma_\epsilon}^2 = \frac{RSS}{n-2}$$

Является несмещенной оценкой дисперсии возмущений, а также стандартной оценкой регрессии в квадрате.

Лекция 4. Теорема Гаусса-Маркова, Классическая линейная регрессия

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Теорема Гаусса - Маркова

Т.к X - детерминированный, а ϵ - случайный вектор, то $\hat{\beta}_1$ - случайная величина.

1. 1) Если модель $Y = \beta_0 + \beta_1 X + \epsilon$ правильно специфицирована,
2. 2) $X_i, i = 1, ..., n$ детерминированы и не все равны между собой,
3. $E(\epsilon_i) = 0$
4. $Var(\epsilon_i) = 0$
5. $Cov(\epsilon_i, \epsilon_j) = 0$

То оценки МНК β_0, β_1 являются BLUE (Best Linear Unbiased Estimator)

ϵ - сумма влияния многих факторов, каждый из которых незначительно влияет на Y. По Центральной предельной теореме такая случайная величина имеет нормальное распределение.

Если ϵ_i, $i = 1, ..., n$ распределены нормально, то есть $\epsilon_i \sim N(0, \sigma^2_\epsilon)$, то оценки параметров β_0, β_1 тоже распределены нормально, причем $\beta_0 \sim N(\beta_0, \frac{\sum X_i^2}{n\sum x_i^2}\sigma^2_\epsilon)$, а $\beta_1 \sim N(\beta_1, \frac{\sigma^2_\epsilon}{\sum x_i^2})$, где $x_i = X_i - \bar{X}, i = 1, ..., n$

Проверка гипотез теория

1. Выбора основной и альтернативной гипотезы,
2. Вычисления некоторой тестовой статистики,
3. Выбора уровня значимости α (числа между 0 и 1),
4. Самые распространенные уровни значимости 0.05 и 0.01,
5. Разбиения множества значений тестовой статистики на две области: там, где основная гипотеза отвергается и там, где основная гипотеза не отвергается

Пример

Модель: $Y = \beta_0 + \beta_1 X + \epsilon$
Нулевая гипотеза: $H_0 : \beta_1 = \beta_1^0$
Двустороння альтернативная гипотеза: $H_1 : \beta_1 \neq \beta_1^0$

1. Сначала необходимо оценить по n наблюдениям модель: $\hat{Y} = \beta_0 + \beta_1 X + \epsilon$

2. Если нулевая гипотеза не отвергается, то тестовая статистика $t = \frac{\hat{\beta_1} - \hat{\beta_1^0}}{s.e.(\hat{\beta_1})} \sim t(n-2)$ имеет t - распределение с n-2 степенями свободы.

3. Гипотеза отвергается если: $|t| \geq t^c_{\alpha/2} r$

4. Если нулевая гипотеза отвергается, то говорят, что коэффициент значим. Если нулевая гипотеза не отвергается, то коэффициент называется незначимым

5. P – value – минимальный уровень значимости, при котором нулевая гипотеза отвергается. На рисунке это площадь всей заштрихованной области.

6. В таблице выделены P-value для проверки гипотез о значимости коэффициентов регрессии. Если P-value коэффициента регрессии меньше, чем выбранный уровень значимости , то нулевая гипотеза отвергается и соответствующий коэффициент является значимым. В приведенном примере при любом разумном уровне значимости константа незначима, а коэффициент наклона значим.

ДИ для оценок коэффициентов регрессии.

Найдем множество всех значений параметра β_1 , гипотеза о равенстве которым при заданном уровне значимости и двусторонней альтернативной гипотезе не отвергается.

Гипотеза $H_0 : \beta_1 = \beta_1^0$

$t_c r = \frac{\hat{\beta_1} - \beta_1^0}{s.e.(\hat{\beta_1})} \sim t(n-2)$

Следовательно формула ДИ для оценки коэффициента:

$$\hat{\beta_1} - t^{cr}_{\alpha/2} s.e.(\hat{\beta_1}) \leq \beta_1^0 \leq \hat{\beta_1} + t^{cr}_{\alpha/2} s.e.(\hat{\beta_1})$$