

10. Модель парной линейной регрессии. Метод наименьших квадратов. Теорема Гаусса-Маркова.

Модель парной регрессии

Рассмотрим модели регрессии. В них присутствует объясняемая (зависимая) переменная y и набор объясняющих (независимых) переменных $x_1, x_2, \dots, x_n$. Связь между ними выражается следующим образом: $y = f(x_1, x_2, \dots, x_n, \beta_1, \beta_2, \dots, \beta_m)$, где f — некоторая функция, а $\beta_1, \beta_2, \dots, \beta_m$ — набор параметров. В зависимости от вида функции $f(x, \beta)$ модели делятся на линейные и нелинейные.

Рассмотрим набор двух переменных $X_t, Y_t, t = \overline{1, n}$. Можно отображать (X_t, Y_t) точками на плоскости.

Задача: подобрать $Y = f(X)$ из параметрического семейства $f(X, \beta)$ наилучшим образом описывающую зависимость Y от X .

Подобрать функцию означает выбрать β . Примером параметрического семейства является модель парной линейной регрессии: $f(X, \beta) = a + bX$.

Определим следующие меры отклонения полученных значений от реальных значений:

1. $F = \sum_{t=1}^n (Y_t - f(X_t, \beta))^2$ — сумма квадратов отклонений.
2. $F = \sum_{t=1}^n |Y_t - f(X_t, \beta)|$ — сумма модулей отклонений.
3. $F = \sum_{t=1}^n g(Y_t - f(X_t, \beta))$, где g — мера с которой отклонение входит в F .

Метод наименьших квадратов (МНК).

Пусть $f(x, \beta)$ — линейная функция с параметрами α и β : $Y = f(X, a, b) = a + bX$. Метод наименьших квадратов позволяет подобрать наилучшие параметры путем решения задачи оптимизации:

$$F = \sum_{t=1}^n (Y_t - f(X_t, \beta))^2 \rightarrow \min$$

Записав необходимые условия экстремума (производная равна 0), и решив систему уравнений, получаем оптимальные параметры:

$$\hat{a} = \frac{1}{n} \sum_{t=1}^n Y_t, \quad \hat{b} = \frac{n \sum X_t Y_t - \sum X_t \sum Y_t}{n \sum X_t^2 - (\sum X_t)^2} = \frac{Cov(X, Y)}{Var(X)}$$

Линейная регрессионная модель с двумя переменными.

Добавим к постановке задачи статистические свойства данных. Запишем уравнение зависимости Y_t от X_t , добавив в него случайную величину ε_t . Получим регрессионное уравнение:

$$Y_t = a + bX_t + \varepsilon_t, t = \overline{1, n}.$$

В этом уравнении X_t — неслучайная детерминированная величина, называется регрессор. Y_t, ε_t — случайные величины.

Случайная ошибка может возникать:

- Если модель является упрощением действительности, и в ней опускаются некоторые параметры.
- Если возможны неточности в измерении данных, используемых в модели.

Можно перечислить основные гипотезы линейной регрессионной модели с двумя переменными:

Основные гипотезы:

1. $Y_t = a + bX_t + \varepsilon_t, t = 1, \dots, n$, - спецификация модели.
2. X_t - детерминированная величина; вектор $(X_1, \dots, X_n)'$ не коллинеарен вектору $i = (1, \dots, 1)'$.
3. (a) $E\varepsilon_t = 0, E(\varepsilon_t)^2 = V(\varepsilon_t) = \sigma^2$ - не зависит от t ($E\varepsilon = 0, V(\varepsilon) = \sigma^2 I_n$)
(b) $E(\varepsilon_t \varepsilon_s) = 0$ при $t \neq s$, некоррелированность ошибок для разных наблюдений.
Часто добавляется условие:
(c) Ошибка $\varepsilon_t, t = 1, \dots, n$, имеют совместное нормальное распределение:
 $\varepsilon_t \sim N(0, \sigma^2)$

Условие **3(a)** – независимость дисперсии ошибок от номера наблюдения называется **гомоскедастичностью**, случай когда это условие не выполняется называют **гетероскедастичностью**.

Условие **3(b)** указывает на некоррелированность ошибок для разных наблюдений. Оно часто нарушается, когда имеем дело с временными рядами. В случае, когда это условие не выполняется, говорят об **автокорреляции ошибок**.

Теорема Гаусса-Маркова

В модели линейной регрессии с двумя параметрами оценки параметров $\hat{a}$ и $\hat{b}$, полученные методом наименьших квадратов, имеют наименьшую дисперсию в классе всех линейных несмещенных оценок.

Исходя из теоремы, полученные МНК оценки параметров наилучшие для модели (в смысле наименьшей дисперсии). Остается оценить оставшийся параметр σ^2 .

Оценка дисперсии ошибок.

Найдем значение регрессионной функции с посчитанными коэффициентами: $\hat{Y}_t = \hat{a} + \hat{b}X_t$. Тогда, остатком регрессии будет называться следующая случайная величина: $e_t = Y_t - \hat{Y}_t$.

Между оценкой σ^2 и суммой квадратов ошибки регрессии e_t существует связь, называемая несмещённый оценкой дисперсии ошибок:

$$s^2 = \widehat{\sigma^2} = \frac{1}{n-2} \sum_{t=1}^n e_t^2$$

Если дисперсия ошибок σ^2 известна, то значения коэффициентов $\hat{a}$ и $\hat{b}$ можно найти по формулам из теоремы Гаусса-Маркова.

На практике, как правило, σ^2 неизвестна, и, заменив σ^2 на s^2 , можем получить оценки дисперсий коэффициентов $\hat{a}$ и $\hat{b}$:

$$\widehat{Var}(\hat{b}) = s^2 \frac{1}{\sum (X_t - \bar{X})^2} \quad \widehat{Var}(\hat{a}) = s^2 \frac{\sum X_t^2}{\sum (X_t - \bar{X})^2}$$

Эти формулы используются для нахождения стандартных отклонений оценок коэффициентов регрессии:

$$s_{\hat{b}} = \sqrt{\widehat{Var}(\hat{b})}, s_{\hat{a}} = \sqrt{\widehat{Var}(\hat{a})}$$