

СТАТИСТИКА БОЛЬШИХ ДАННЫХ

Лекция 19. Задача множественной линейной регрессии. Часть 1



В этой лекции рассматриваются особенности решения задачи множественной линейной регрессии, связанные с увеличением размерности модели.

- Постановка задачи множественной линейной регрессии
- Оценивание коэффициентов регрессии
- Особенности решения и интерпретации задачи множественной линейной регрессии
- Заключение

Постановка задачи

Имеется p различных факторов X_1, X_2, \dots, X_p . **Модель множественной линейной регрессии** имеет следующий вид

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + e$$

где X_j представляет j -ый фактор, а β_j измеряет связь между этой переменной и откликом. Мы интерпретируем β_j как *средний эффект изменения Y на единице изменения X_j при постоянных значениях всех других факторов*.


Оценивание коэффициентов регрессии. Если оценки $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ заданы, то прогноз \hat{y} отклика Y имеет вид

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p.$$

Коэффициенты регрессии $\beta_0, \beta_1, \dots, \beta_p$ оцениваются методом наименьших квадратов (МНК). Минимизируют сумму квадратов остатков

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_p x_{ip})^2$$





Задача множественной линейной регрессии

Теория МНК и вывод аналитических выражений для оценок коэффициентов множественной линейной регрессии $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ вместе с сопутствующими им показателями точности (дисперсиями и стандартными ошибками) можно найти в продвинутых курсах по математической статистике (см., например, (Кендалл и Стьюарт, 1973)); соответственно процедуры вычисления МНК-оценок представлены в каждом статистическом пакете программ, в том числе пакете *R*, используемым в нашем курсе — поэтому мы приводим здесь их без вывода.

Напомним, что главная цель этой лекции — это трудности и особенности как решения задачи множественной линейной регрессии, так и интерпретации этого решения.

Задача множественной линейной регрессии

Рассмотрим модель множественной линейной регрессии

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + e_i, \quad i = 1, \dots, n,$$

где $x_{11}, x_{12}, \dots, x_{1n}, x_{21}, x_{22}, \dots, x_{2n}, \dots, x_{p1}, x_{p2}, \dots, x_{pn}$ — заданные значения факторов (X_1, \dots, X_p); y_1, \dots, y_n — наблюдаемые значения отклика; e_1, \dots, e_n — независимые, нормально распределенные с нулевым математическим ожиданием и одинаковой (неизвестной) дисперсией случайные величины; $\beta_0, \beta_1, \dots, \beta_p$ — неизвестные параметры, подлежащие оцениванию.


Матричные обозначения: матрица факторов $X = (1 \ x_1 \ \dots \ x_p)$,

$1 = (1, 1, \dots, 1)^T$ — вектор-столбец размерности n ,

$x_j = (x_{1j}, \dots, x_{nj})^T$ — вектор-столбец значений фактора X_j , $j = 1, 2, \dots, p$;

$y = (y_1, \dots, y_n)^T$ — вектор значений отклика Y .





Задача множественной линейной регрессии

Тогда, в предположении невырожденности матрицы факторов X , **оптимальные МНК-оценки** вектора параметров $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)^T$ множественной линейной регрессии даются в следующем виде (Кендалл и Стьюарт, 1973):

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

Несмещенная оценка дисперсии $\hat{\sigma}^2$ погрешностей измерений отклика имеет вид

$$\hat{\sigma}^2 = s_e^2 = \frac{1}{n - p - 1} (y - X\hat{\beta})^T (y - X\hat{\beta}),.$$

Задача множественной линейной регрессии

Этот результат позволяет построить несмещенную оценку для ковариационной матрицы оценок параметров множественной линейной регрессии, используя в них вместо σ^2 статистику s^2 , определенную формулой (5).

Ковариационная матрица оценок дается следующим выражением

$$V(\hat{\beta}) = \sigma^2 (X^T X)^{-1}, \quad \text{ее оценка имеет вид} \quad \hat{V}(\hat{\beta}) = s_e^2 (X^T X)^{-1}.$$

Границы **доверительного интервала** для прогноза значений регрессии при заданных значениях факторов $x = (1, x_1, \dots, x_p)^T$ определяются по формуле

$$\hat{y} \pm t_{1-\frac{\alpha}{2}}(n-p-1) s_e \sqrt{x^T (X^T X)^{-1} x},$$

где $t_{1-\alpha/2}(n-p-1)$ — t -статистика Стьюдента, а α — уровень значимости.



Особенности решения и интерпретации задачи множественной линейной регрессии

Важный блок вопросов связан с оценкой влияния факторов на прогноз отклика.

1. Имеется ли хотя бы один полезный фактор из X_1, X_2, \dots, X_p для прогноза отклика Y ?
2. Все факторы полезны для прогноза, или только какое-либо их подмножество?
3. Насколько хорошо регрессионная модель согласуется с данными?
4. При заданном множестве значений факторов какое значение отклика мы должны получить и насколько точен полученный прогноз?

Особенности решения и интерпретации

Имеется ли связь между откликом и факторами?

В задаче множественной линейной регрессии с p факторами нужно выяснить равны ли все коэффициенты регрессии нулю. Т.е. мы проверяем нулевую гипотезу

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$$

$$H_1: \text{хотя бы один } \beta_j \neq 0.$$

Тест для проверки этой гипотезы основан на вычислении F -статистики

$$F = \frac{(TSS - RSS)/p}{RSS/(n - p - 1)},$$

$$TSS = \sum (y_i - \bar{y})^2,$$

$$RSS = \sum (y_i - \hat{y})^2.$$



Особенности решения и интерпретации

Если принятые допущения об ошибках линейной модели (некоррелированность, нормальность и одинаковость дисперсии) верны, то можно показать, что

$$E[RSS/(n - p - 1)] = \sigma^2,$$

$$E[(TSS - RSS)/p] = \sigma^2.$$

Если верна нулевая гипотеза, то следует ожидать, что значения F -статистики будут принимать значения близкие к 1. Напротив, если верна альтернатива H_1 $E[(TSS - RSS)/p] > \sigma^2$, то следует ожидать значения F -статистики больше 1.

Ответ зависит от величин n и p . Если n велико, то даже небольшое превышение 1 значением F -статистики может свидетельствовать против нулевой гипотезы. Напротив, при малых n требуются большие значения F -статистики, чтобы отвергнуть H_0 . Окончательный ответ дают статистики p -значений (см. простую линейную регрессию), связанные с F -статистикой и ее распределением — во всех статистических пакетах вычисляются эти величины.