GeekBrains

Теория вероятностей и математическая статистика

Вебинары

Урок 7

Теория вероятностей и математическая статистика

Многомерный статистический анализ. Линейная регрессия

На этом уроке мы изучим:

1. Для чего применяют многомерный анализ.
2. Что такое линейная регрессия.
3. Коэффициент детерминации.
4. F-критерий Фишера.
5. t-статистику Стьюдента.

Многомерный статистический анализ — раздел статистики, который посвящен исследованиям экспериментов с многомерными наблюдениями.

Многомерный статистический анализ

Раздел статистики, который посвящен исследованиям экспериментов с многомерными наблюдениями

1. Зависимость между признаками и их влияние на некоторую переменную
2. Классификация объектов
3. Понижение размерности пространства

Модель регрессии

Модель зависимости количественной переменной y (объясняемой) от другой или нескольких других переменных x_i (факторов, предикторов)

$$y = f_b(x_1, \ldots, x_m) + \varepsilon$$

$f_b(x)$ — некоторая функция, имеющая набор параметров b, а ε — случайная ошибка. На ошибку накладывается условие, что её математическое ожидание равно 0

$$M(\varepsilon) = 0$$

Линейная регрессия

1. **Спецификация модели**

$$y = f_b(x_1, \ldots, x_m) + \varepsilon$$

функция fb(x) является линейной, модель имеет вид:

$$y = b_0 + b_1 x_1 + \cdots + b_m x_m + \varepsilon$$

Парная регрессия (Частный случай)

$$y = b_0 + b_1 x + \varepsilon$$

Линейная регрессия

Модель в матричном виде:

$$Y = X \cdot b + E,$$

где Y – вектор зависимой переменной, X – матрица, E – вектор ошибок, b – вектор оцениваемых коэффициентов

$$X = \begin{pmatrix} x_{10} & \cdots & x_{1k} \\ x_{20} & \cdots & x_{2k} \\ \vdots & \ddots & \vdots \\ x_{m0} & \cdots & x_{mk} \end{pmatrix}$$

Линейная регрессия

2. <u>Идентификация модели (оценка параметров)</u>

Y - реальные данные

$\hat{Y} = X\hat{b}$ - оцененные данные

Y - $\hat{Y}$ = e

Метод наименьших квадратов (МНК)

$$\hat{b} = \min(e^T e)$$

$$b = (X^\top X)^{-1} X^\top Y$$

Для парной регрессии: $$b_1 = \frac{\overline{yx} - \overline{y} \cdot \overline{x}}{\overline{x^2} - (\overline{x})^2}, \; b_0 = \overline{y} - b_1 \cdot \overline{x}.$$

Линейная регрессия

3. **Оценка качества модели**

Коэффициент детерминации (R^2)

$$R^2 = 1 - \frac{D(\varepsilon)}{D(y)}$$

Коэффициент детерминации принимает значения из интервала [0, 1]. Близкие к 1 значения коэффициента детерминации свидетельствуют о высоком качестве модели

$$R^2 = 1 - \frac{SS_{res}}{SS_y}$$

$$SS_Y = \sum_{i=1}^{n}(y_i - \overline{Y})^2$$ - сумма квадратов отклонений значений массива Y от среднего

SSres — остаточная сумма квадратов отклонений от их среднего

Корреляция и детерминация

Значение коэффициента детерминации ниже 1 не означает, что модель построена плохо (и могла бы быть лучше)

Для линейной модели, построенной с помощью метода наименьших квадратов верно равенство:

$$R^2 = r_{YZ}^2$$

где r_{YZ}^2 - коэффициент корреляции Пирсона между массивами

Коэффициент детерминации прямо зависит от уровня корреляции в данных и не может достигнуть 1, если в данных нет линейной зависимости

Значимость уравнения регрессии

Используем F-тест Фишера, который проверяет нулевую гипотезу о незначимости коэффициента детерминации (в данных нет зависимости):

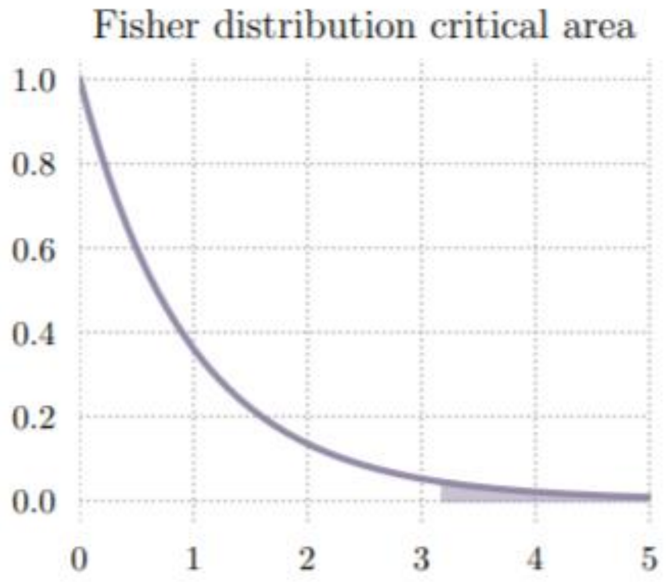
$$F = \frac{R^2/m}{(1 - R^2)/(n - m - 1)}$$

R2 — коэффициент детерминации, n — число наблюдений, m — число факторов

Эта статистика имеет распределение Фишера с параметрами k1 = m, k2 = n − m − 1

Значимость уравнения регрессии

Распределение Фишера имеет один хвост, поэтому рассматривается правосторонняя критическая область



Если статистика попадает в критическую область, то гипотеза о равенстве нулю коэффициента детерминации отвергается. Это означает, что построенная нами модель значимо соответствует данным

Доверительные интервалы для коэффициентов парной регрессии

Получили оценку коэффициента наклона b1, и пусть b1 — реальное значение этого коэффициента. Рассмотрим статистику:

$$t = \frac{\hat{b}_1 - b_1}{S_{slope}}$$

$$S_{slope} = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(x_i - \overline{X})^2}}$$

S slope — стандартная ошибка коэффициента наклона

Доверительные интервалы для коэффициентов парной регрессии

Статистика t имеет распределение Стьюдента с параметром $df = n - 2$. Отсюда можно, имея доверительную вероятность p, построить доверительный интервал для коэффициента наклона по формуле:

$$P\left(\hat{b}_1 + t_{\alpha/2,\, n-2} \cdot S_{slope} \leq b_1 \leq \hat{b}_1 + t_{1-\alpha/2,\, n-2} \cdot S_{slope}\right) = p$$

где $\alpha = 1 - p$, $t_{\beta, n-2}$ — квантиль порядка β для распределения Стьюдента

Доверительные интервалы для коэффициентов парной регрессии

Доверительный интервал для коэффициента сдвига b0

Стандартная ошибка коэффициента сдвига вычисляется по формуле:

$$S_{intercept} = S_{slope} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

$$t = \frac{\hat{b}_0 - b_0}{S_{intercept}}$$

Доверительный интервал для коэффициента наклона:

$$P\left(\hat{b}_0 + t_{\alpha/2,\, n-2} \cdot S_{intercept} \le b_0 \le \hat{b}_0 + t_{1-\alpha/2,\, n-2} \cdot S_{intercept}\right) = p$$

Резюме

1. Непосредственно факт наличия линейной взаимосвязи проверяется с помощью корреляционного анализа
2. Если линейная зависимость наблюдается, можно построить модель линейной регрессии. Она укажет на характер этой зависимости (т.е. на то, каким именно образом изменяется переменная под влиянием факторов)
3. С помощью F-критерия Фишера можно проверить, является ли уровень зависимости в данных статистически значимым
4. С помощью доверительных интервалов можно оценить реальный вклад каждого фактора в изменение переменной

Итоги

1. Для чего применяют многомерный анализ.
2. Что такое линейная регрессия.
3. Коэффициент детерминации.
4. F-критерий Фишера.
5. t-статистика Стьюдента.