Множественная регрессия

Лекция

Цели лекции

- Обобщение парной регрессии на случай нескольких объясняющих переменных
- Интерпретация множественной регрессии
- Качество множественной регрессии
- Новые возможности регрессии

Виды множественной регрессии

1. Классическая линейная регрессия
2. Нелинейная регрессия
3. Специальные виды переменных

Модель множественной регрессии

Множественная регрессия имеет вид:

$$M[Y / x_1, x_2, \ldots, x_m] = f(x_1, x_2, \ldots, x_m)$$

Уравнение множественной регрессии:

$$Y = f(\beta, \mathbf{X}) + \varepsilon$$

где $\mathbf{X} = (X_1, X_2, \ldots, X_m)$ – вектор объясняющих переменных,

β – вектор параметров (подлежащих определению),

ε – вектор случайных ошибок (отклонений),

Y – зависимая переменная.

Линейная модель множественной регрессии

Теоретическое уравнение линейной множественной регрессии:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m + \varepsilon$$

или для индивидуальных наблюдений:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_m x_{im} + \varepsilon_i$$

$i = 1, 2, \ldots, n$, $n \geq m+1$, $k = n-m-1$ – число степеней свободы

Для обеспечения статистической надежности должно выполняться условие: $n > 3(m+1)$

Оценки параметров линейной множественной регрессии

Эмпирическое уравнение регрессии:

$$\hat{Y} = b_0 + b_1 X_1 + \ldots + b_m X_m \qquad \hat{y}_i = b_0 + b_1 x_{i1} + \ldots + b_m x_{im}$$

Самый распространенный метод оценки параметров – МНК

$$b_j,\ j = \overline{0,m}: \sum_{i=1}^{n} \left(y_i - (b_0 + \sum_{j=1}^{m} b_j x_{ij}) \right)^2 = \sum_{i=1}^{n} e_i^2 \to \min$$

$\hat{y}_i$

Предпосылки МНК

$1^0.\quad M(\varepsilon_i)=0,\ \ i=1,2,\ldots,n$

$2^0.\quad D(\varepsilon_i)=D(\varepsilon_j)=\sigma^2,\forall i,j$ Гомоскедастичность

$3^0.\quad Cov(\varepsilon_i,\varepsilon_j)=\begin{cases}0, & i\neq j\\ \sigma^2, & i=j\end{cases}$ Отсутствие автокорреляции

$4^0.\quad Cov(\varepsilon_i,x_i)=0$

5^0. Модель является линейной относительно параметров

Дополнительные предпосылки МНК

6^0. Отсутствие мультиколлинеарности: между объясняющими переменными отсутствует строгая (сильная) линейная зависимость

7^0. Ошибки ε_i имеют нормальное распределение:

$$\varepsilon_i - N(0, \sigma^2)$$

При выполнении этих предпосылок МНК-оценки коэффициентов множественной регрессии будут несмещенными, состоятельными и эффективными в классе линейных оценок

Оценка параметров классической регрессионной модели МНК

Матричная форма СЛАУ:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} b_0 \\ b_1 \\ \dots \\ b_m \end{pmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1m} \\ 1 & x_{21} & \dots & x_{2m} \\ . & . & . & . \\ 1 & x_{n1} & \dots & x_{nm} \end{bmatrix}$$

$$\mathbf{E} = \left(e_1\ e_2 \dots\ e_n\right)^T$$

Оценка параметров классической регрессионной модели МНК

$$\mathbf{X}^T\mathbf{X}\mathbf{B} = \mathbf{X}^T\mathbf{Y} \Rightarrow \mathbf{B} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y})$$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} n & \sum x_{i1} & \dots & \sum x_{im} \\ \sum x_{i1} & \sum x_{i1}^2 & \dots & \sum x_{i1}x_{im} \\ . & . & . & . \\ \sum x_{im} & \sum x_{i1}x_{im} & \dots & \sum x_{im}^2 \end{bmatrix} \quad \mathbf{X}^T\mathbf{Y} = \begin{bmatrix} \sum y_i \\ \sum y_i x_{i1} \\ \dots \\ \sum y_i x_{im} \end{bmatrix}$$

Интерпретация множественной линейной регрессии

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

Интерпретация: коэффициент регрессии при переменной X_1 выражает предельный прирост зависимой переменной при изменении переменной X_1 , при условии постоянства других переменных:

$$\beta_1 = \frac{dY}{dX_1} \approx \frac{\Delta Y}{\Delta X_1}, \quad X_2 = const$$

Интерпретация множественной логарифмической регрессии

$$\ln Y_i = \beta_0 + \beta_1 \ln X_{1i} + \beta_2 \ln X_{2i} + \varepsilon_i$$

Интерпретация: коэффициент регрессии при переменной $\ln X_1$ выражает эластичность зависимой переменной при изменении переменной X_1 , при условии постоянства других переменных:

$$\beta_1 = \frac{dY}{dX_1} \cdot \frac{X_1}{Y} \approx \frac{\Delta Y}{\Delta X_1} \cdot \frac{X_1}{Y}, \quad X_2 = const$$

Интерпретация множественной линейной регрессии

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

Оценка коэффициента регрессии:

$$b_1 = \frac{Cov(x_1, y)Var(x_2) - Cov(x_2, y)Cov(x_1, x_2)}{Var(x_1)Var(x_2) - [Cov(x_1, x_2)]^2}$$

Величина оценки коэффициента регрессии формируется под влиянием не только связи изучаемого фактора с зависимой переменной, но и структуры связей между объясняемыми переменными

Интерпретация множественной линейной регрессии

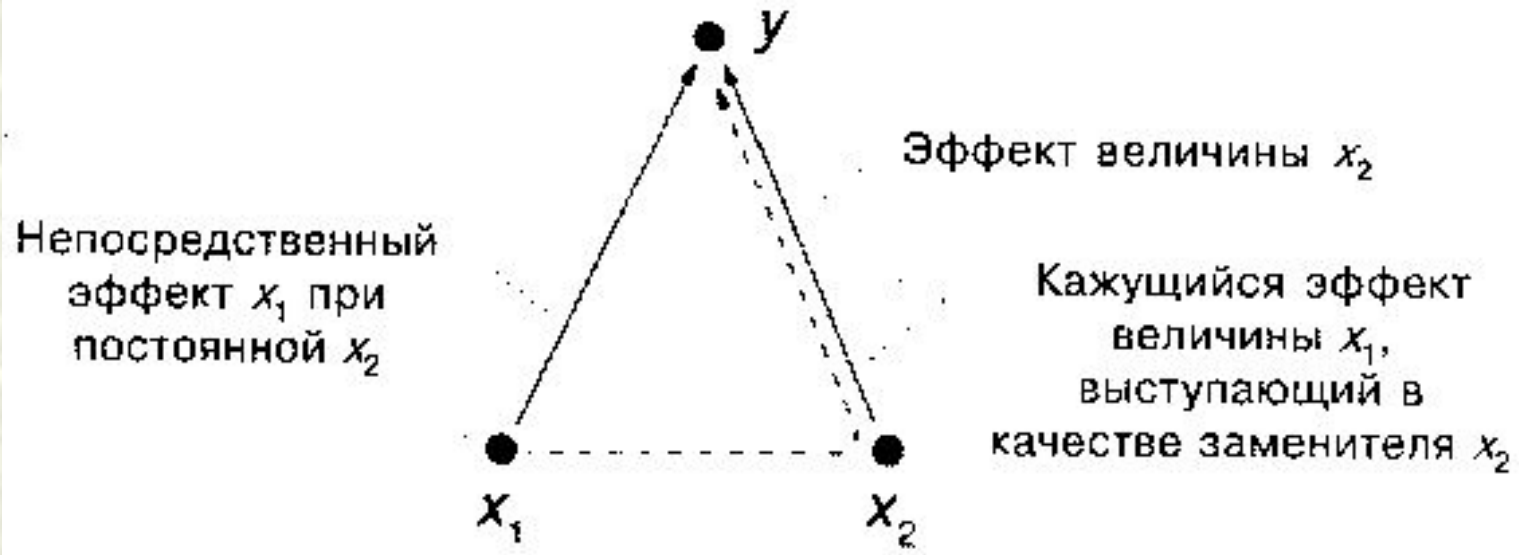
Рассмотрим проявление множественных связей в парной регрессии (в случае исключения значимой переменной X_2):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

$$\Downarrow$$

$$Y_i = \beta_0 + \beta_1' X_{1i} + \varepsilon_i'$$

Интерпретация множественной линейной регрессии



В случае исключения значимой переменной X_2 часть изменений Y за счет X_2 будет приписана X_1 , если переменная X_1 может замещать X_2. В результате оценка значения β_1 будет смещена.

Интерпретация множественной регрессии: замещающие переменные

Замещающая переменная – это переменная, коррелирующая с отсутствующей переменной уравнения множественной регрессии, и выполняющая за счет этого функции отсутствующей переменной

Включение замещающей переменной позволяет правильно оценить роль других факторов, освободив их от функции замещения отсутствующих переменных

Анализ предельного вклада факторов

Множественная регрессия позволяет разложить суммарное влияние факторов на составные части, точнее выявив предельный вклад каждого фактора

Система показателей качества множественной регрессии

1. Показатели качества коэффициентов регрессии
2. Показатели качества уравнения в целом

Показатели качества коэффициентов регрессии

1. Стандартные ошибки оценок.
2. Значения *t*-статистик.
3. Интервальные оценки коэффициентов линейного уравнения регрессии.
4. Доверительные области для зависимой переменной.

Ковариационная матрица вектора оценок коэффициентов регрессии

$$\Sigma_\beta = M\left[(\beta - \mathbf{B})(\beta - \mathbf{B})^T\right] = \begin{bmatrix} \sigma_{00} & \sigma_{01} & \cdots & \sigma_{0m} \\ \sigma_{10} & \sigma_{11} & \cdots & \sigma_{1m} \\ \cdot & \cdot & \cdot & \cdot \\ \sigma_{m0} & \sigma_{m10} & \cdots & \sigma_{mm} \end{bmatrix}$$

$$\sigma_{ij} = Cov(b_i b_j) = \left[M[b_i] = \beta_i\right] = M\left[(b_i - \beta_i)(b_j - \beta_j)\right]$$

На главной диагонали матрицы Σ_β находятся дисперсии оценок коэффициентов регрессии:

$$\sigma_{jj} = \sigma^2_{b_j}$$

Ковариационная матрица вектора возмущений

$$\Sigma_\varepsilon = M[\varepsilon\varepsilon^T] = \begin{bmatrix} M[\varepsilon_1^2] & M[\varepsilon_1\varepsilon_2] & \dots & M[\varepsilon_1\varepsilon_n] \\ M[\varepsilon_2\varepsilon_1] & M[\varepsilon_2^2] & \dots & M[\varepsilon_2\varepsilon_n] \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ M[\varepsilon_n\varepsilon_1] & M[\varepsilon_n\varepsilon_2] & \dots & M[\varepsilon_n^2] \end{bmatrix}$$

Матрица Σ_ε обладает следующими свойствами:

1. Все элементы, не лежащие на главной диагонали, равны нулю (3^0).

2. Все элементы, лежащие на главной диагонали равны (1^0 и 2^0):

$$M[\varepsilon_i^2] = M[\varepsilon_i - 0]^2 = D[\varepsilon_i] = \sigma_\varepsilon^2 = \sigma^2 \quad \Rightarrow \quad \Sigma_\varepsilon = \sigma^2 E_n$$

Стандартные ошибки коэффициентов

Можно показать, что $$\Sigma_{\varepsilon} = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \quad (1)$$

Поскольку истинное значение дисперсии σ^2 по выборке определить нельзя, заменяем его несмещенной оценкой:

$$S_e^2 = S^2 = \frac{\mathbf{E}^T \mathbf{E}}{n-m-1} = \frac{\sum_{i=1}^{n} e_i^2}{n-m-1} \quad (2)$$

Стандартные ошибки коэффициентов

Из (1) и (2) следует формула для расчета выборочных дисперсий эмпирических коэффициентов регрессии:

$$S_{b_j}^2 = S^2 z'_{jj}, \quad j = \overline{0,m}$$

Здесь $z'_{jj}, \quad j = \overline{0,m}$ – диагональные элементы матрицы

$$\mathbf{Z}^{-1} = (\mathbf{X}^T \mathbf{X})^{-1}$$

Стандартные ошибки коэффициентов

Как и в случае парной регрессии:

$$S_{b_j} = \sqrt{S_{b_j}^2}, \quad j = \overline{0, m}$$

– стандартные ошибки коэффициентов

$$S = \sqrt{S^2}$$

– стандартная ошибка регрессии

Стандартные ошибки коэффициентов модели с двумя переменными

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

Расчет стандартных ошибок коэффициентов регрессии для случая двух факторов:

$$S_{b_j} = \sqrt{\frac{S_e^2}{nVar(X_1)} \cdot \frac{1}{1 - r_{x_1 x_2}^2}}, \quad j = 1,2$$

Значимость коэффициентов регрессии

Значимость коэффициентов множественной регрессии проверяется по *t*-критерию Стьюдента:

$$|t| = \frac{|b_j|}{s_{b_j}} > t_{\frac{\alpha}{2};n-m-1}$$

$$t = \frac{b_j}{s_{b_j}}$$

– расчетное значение *t*-статистики коэффициента b_j

t-тесты обеспечивают проверку значимости предельного вклада каждой переменной при допущении, что все остальные переменные уже включены в модель

Незначимость коэффициента регрессии не всегда может служить основанием для исключения соответствующей переменной из модели

Доверительные интервалы для коэффициентов регрессии

$$b_j - t_{\frac{\alpha}{2};n-m-1} S_{b_j} < \beta_j < b_j + t_{\frac{\alpha}{2};n-m-1} S_{b_j}$$

Данный доверительный интервал накрывает с надежностью $(1-\alpha)$ истинное значение коэффициента регрессии

Доверительная область для условного математического ожидания зависимой переменной

$$\hat{Y}_p - t_{\frac{\alpha}{2};k} S_{\overline{y(\mathbf{X}_p)}} < M(Y_p / \mathbf{X}_p^T) < \hat{Y}_p + t_{\frac{\alpha}{2};k} S_{\overline{y(\mathbf{X}_p)}}$$

$$k = n - m - 1, \quad \hat{Y}_p = \hat{y}(\mathbf{X}_p) = b_0 + \sum_{j=1}^{m} b_j x_{pj}$$

$$S_{\overline{y(\mathbf{X}_p)}} = S\sqrt{\mathbf{X}_p^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_p}$$

Доверительная область для индивидуальных значений *Y*

$$\hat{Y}_p - t_{\frac{\alpha}{2};k} S_{\hat{y}(\mathbf{X}_p)} < Y_p^* < \hat{Y}_p + t_{\frac{\alpha}{2};k} S_{\hat{y}(\mathbf{X}_p)}$$

$$k = n - m - 1, \quad \hat{Y}_p = \hat{y}(\mathbf{X}_p) = b_0 + \sum_{j=1}^{m} b_j x_{pj}$$

$$S_{\hat{y}(\mathbf{X}_p)} = S\sqrt{1 + \mathbf{X}_p^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_p}$$

Показатели качества уравнения регрессии в целом

Основные показатели качества:

1. Коэффициент детерминации R^2
2. Скорректированный коэффициент детерминации $\overline{R}^2$
3. Значение F-статистики
4. Сумма квадратов остатков (RSS)
5. Стандартная ошибка регрессии S_e
6. Прочие показатели: средняя ошибка аппроксимации, индекс множественной корреляции и т.д.

Коэффициент детерминации R^2

Коэффициент R^2 показывает долю объясненной вариации зависимой переменной:

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}$$

R^2 всегда увеличивается с включением новой переменной

Низкое значение R^2 не свидетельствует о плохом качестве модели, и может объясняться наличием существенных факторов, не включенных в модель

Коэффициенты R^2 в разных моделях с разным числом наблюдений (и переменных) несравнимы

Скорректированный коэффициент детерминации $\overline{R}^2$

$\overline{R}^2$ показывает долю объясненной вариации зависимой переменной с учетом числа объясняющих переменных уравнения регрессии:

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

Добавление переменной приведет к увеличению $\overline{R}^2$, если ее *t*-статистика будет по модулю больше 1. Следовательно, увеличение $\overline{R}^2$ при добавлении новой переменной ***необязательно означает***, что ее коэффициент значимо отличается от нуля

Скорректированные коэффициенты $\overline{R}^2$ в разных моделях с разным числом наблюдений (и переменных) ограниченно сравнимы

F-статистика для проверки качества уравнения регрессии

F-статистика представляет собой отношение объясненной суммы квадратов (в расчете на одну независимую переменную) к остаточной сумме квадратов (в расчете на одну степень свободы)

$$F = \frac{\frac{ESS}{m}}{\frac{RSS}{n-m-1}}$$

n – число выборочных наблюдений, *m* – число объясняющих переменных

F-статистика для проверки значимости коэффициента R^2

F-статистика рассчитывается на основе коэффициента детерминации

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

Для проверки значимости *F*-статистики используются таблицы *F*-распределения с m и $(n–m–1)$ степеней свободы

Сумма квадратов остатков *RSS*

Является оценкой необъясненной части вариации зависимой переменной

$$RSS = \sum_{i=1}^{n} e_i^2$$

Используется как основная минимизируемая величина в МНК, а также для расчета других показателей

Значения *RSS* в разных моделях с разным числом наблюдений и (или) переменных несравнимы

Стандартная ошибка регрессии S_e

Является оценкой величины квадрата ошибки, приходящейся на одну степень свободы модели

$$S_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n - m - 1}}$$

Используется как основная величина для измерения качества модели (чем она меньше, тем лучше)

Значения S_e в однотипных моделях с разным числом наблюдений и (или) переменных сравнимы

Расчет эластичности для линейной регрессии

Средние коэффициенты эластичности:

$$\overline{L}_{YX_j} = b_j \frac{\overline{X}_j}{\overline{Y}}$$

Частные коэффициенты эластичности:

$$L_{Y_{X_j}} = b_j \frac{X_j}{\hat{Y}_{X_j / X_1, X_2, \ldots, X_{j-1}, X_{j+1}, \ldots, X_M}}$$

Индекс множественной корреляции

Тесноту совместного влияния факторов на результат характеризует индекс (показатель) множественной корреляции:

$$R = R_{yx_1 \square \ x_m} = \sqrt{1 - \frac{S_e^2}{S_y^2}} = \sqrt{R^2}$$

Диапазон значений лежит от 0 до 1. Чем ближе его значение к 1, тем теснее связь результативного признака Y со всем набором объясняющих факторов X_i

Индекс множественной корреляции

Справедливо неравенство:

$$R_{yx_1x_2\boxtimes\ x_m} \geq \max_i r_{yx_i}$$

При правильном включении факторов в модель индекс множественной корреляции будет существенно превосходить наибольшее из значений коэффициента парной корреляции

Новые возможности множественной регрессии

1. Многочлены от объясняющих переменных
2. Исследование структуры связи во времени: запаздывающие переменные – лаги
3. Анализ структурных сдвигов

Многочлены от объясняющих переменных

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i + \varepsilon_i$$

Появляются возможности:

- исследования зависимостей, для которых существенно наличие максимумов и минимумов,
- прямой анализ нелинейных эффектов

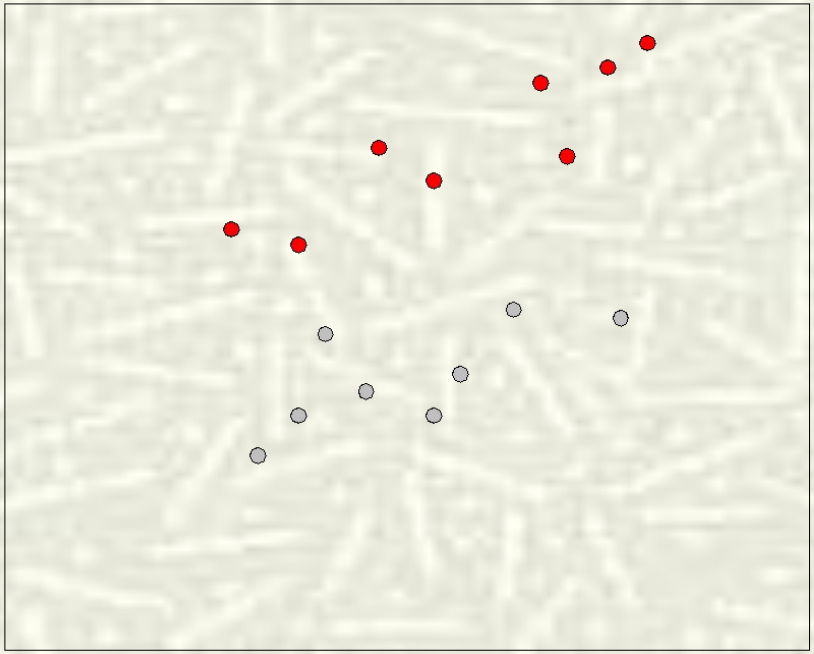
Лаговые переменные

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_{i-1} + \varepsilon_i$$

Учет структуры взаимосвязей во времени зависимой и объясняющих переменных

Анализ структурных сдвигов

- Тест Чоу на наличие структурного сдвига
- Фиктивные переменные сдвига и наклона



Конец лекции