Эконометрика

Модель линейной регрессии

Шишкин Владимир Андреевич

Пермский государственный национальный исследовательский университет

Модель множественной линейной регрессии

Модель парной регрессии:

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i, \quad i = 1, \dots, n$$

Модель множественной регрессии

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_m x_{im} + \varepsilon_i, \quad i = 1, \dots, n$$

Модель множественной линейной регрессии

Основные гипотезы

1. $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} + \varepsilon_i$, $i = 1, \ldots, n$ — модель правильно специфицирована
2.
   - $x_{i1}, \ldots, x_{im}$ — детерминированные величины
   - векторы $x_j = (x_{1j}, \ldots, x_{nj})^\top$, $j = 1, \ldots, m$, линейно независимы в $\mathbb{R}^n$ и не коллинеарны вектору $\imath = (1, \ldots, 1)^\top$
3. $\mathsf{E}\,\varepsilon_i = 0$, $\mathsf{E}\left(\varepsilon_i^2\right) = \mathsf{D}\left(\varepsilon_i\right) = \sigma^2$
4. $\mathsf{E}\left(\varepsilon_i \varepsilon_j\right) = 0$ при $i \neq j$
5. $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

Классическая нормальная линейная регрессионная модель

Модель множественной линейной регрессии

Основные гипотезы в матричной форме

1. $y = X\beta + \varepsilon$ — правильная спецификация модели
2. X — детерминированная матрица ранга $m+1$
3. $\mathsf{E}\,\varepsilon = 0$, $\mathsf{D}\,(\varepsilon) = \mathsf{E}\,(\varepsilon\varepsilon^\top) = \sigma^2 I_n$
4. $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \; X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nm} \end{pmatrix}, \; \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}, \; \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Модель множественной линейной регрессии

Теорема Гаусса–Маркова

Предположим, что

1. $y = X\beta + \varepsilon$;
2. X — детерминированная матрица ранга $(m+1)$;
3. $\mathsf{E}\,\varepsilon = 0$, $\mathsf{D}\,(\varepsilon) = \mathsf{E}\,(\varepsilon\varepsilon^\top) = \sigma^2 I_n$.

Тогда оценка, полученная методом наименьших квадратов, является наиболее эффективной (в смысле наименьшей дисперсии) оценкой в классе линейных несмещённых оценок.

Модель множественной линейной регрессии

Метод наименьших квадратов

$$e = y - \hat{y} = y - Xb, \qquad \mathsf{ESS} = \sum_{i=1}^{n} e_i^2 = e^\top e \to \min$$

$$\begin{aligned} e^\top e = (y - Xb)^\top (y - Xb) = \\ = y^\top y - y^\top Xb - b^\top X^\top y + b^\top X^\top Xb = \\ = y^\top y - 2b^\top X^\top y + b^\top X^\top Xb \end{aligned}$$

$$\frac{\partial \mathsf{ESS}}{\partial b} = 2X^\top Xb - 2X^\top y = 0$$

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

Модель множественной линейной регрессии
Коэффициент детерминации R^2

Центрированные величины: $y_* = y - \bar{y}\imath$, $\hat{y}_* = \hat{y} - \bar{y}\imath$

$$\underbrace{y_*^\top y_*}_{\text{TSS}} = \underbrace{e^\top e}_{\text{ESS}} + \underbrace{\hat{y}_*^\top \hat{y}_*}_{\text{RSS}}$$

Коэффициент детерминации:

$$R^2 = 1 - \frac{\text{ESS}}{\text{TSS}} = \frac{\text{RSS}}{\text{TSS}}, \qquad 0 \leqslant R^2 \leqslant 1$$

Скорректированный коэффициент детерминации:

$$R^2_{\text{adj}} = 1 - \frac{\text{ESS}/(n-m-1)}{\text{TSS}/(n-1)} = 1-(1-R^2)\frac{n-1}{n-m-1} \leqslant R^2$$

Модель множественной линейной регрессии

Проверка гипотезы $H_0\colon \beta = \beta^*$ (t-критерий Стьюдента)

1. Оценка дисперсии ошибки: $s_e^2 = \frac{1}{n-m-1}\sum_{i=1}^n e_i^2$
2. Оценка дисперсии коэффициента b_i: $s_{b_i}^2 = s_e^2 q_i$
   q_i — $(i+1)$-й диагональный элемент матрицы $(X^\top X)^{-1}$
3. Наблюдаемое значение t-статистики для коэффициента b_i:
$$t_{i,\text{набл}} = \frac{b_i - \beta_i^*}{s_{b_i}} \sim t_{n-m-1}$$
4. Нулевая гипотеза отклоняется на уровне значимости α, если $t_{i,\text{набл}} > t_{1-\frac{\alpha}{2},n-m-1}$

Доверительный интервал для β_i:
$$\left[b_i + t_{\frac{\alpha}{2},n-m-1}s_{b_i}, b_i + t_{1-\frac{\alpha}{2},n-m-1}s_{b_i}\right]$$

Модель множественной линейной регрессии

Проверка гипотезы $\mathrm{H}_0\colon \beta_1 = \dots = \beta_m = 0$ (проверка значимости модели с помощью F-критерия Фишера)

$$F_{\text{набл}} = \frac{R^2}{1-R^2}\frac{m}{n-m-1} = \frac{\text{RSS}}{\text{ESS}}\frac{m}{n-m-1} = $$
$$= \frac{\hat{y}_*^\top \hat{y}_*/m}{e^\top e/(n-m-1)} \sim F_{m,n-m-1}$$

Гипотеза H_0 отклоняется на уровне значимости α, если $F_{\text{набл}} > F_{1-\alpha,m,n-m-1}$

Модель множественной линейной регрессии
Прогнозирование

$$y = X\beta + \varepsilon$$
$$y' = x'\beta + \varepsilon'$$

Прогнозное значение и стандартная ошибка прогноза:

$$\boxed{\hat{y}' = x'b, \qquad s_{\hat{y}'} = s_e\sqrt{1 + x'(X^\top X)^{-1}(x')^\top}}$$

Доверительный интервал для прогнозного значения:

$$\hat{y}' + t_{\frac{\alpha}{2},n-m-1} \cdot s_{\hat{y}'} \leqslant y' \leqslant \hat{y}' + t_{1-\frac{\alpha}{2},n-m-1} \cdot s_{\hat{y}'}$$

Нарушения условий Гаусса–Маркова

1. $y \neq X\beta + \varepsilon$ — неправильная спецификация модели
2. $\operatorname{rank} X < m + 1$ — мультиколлинеарность
3. $\mathrm{D}\,\varepsilon_i = \sigma_i^2 \neq \mathrm{const}$ — гетероскедастичноть
4. $\mathrm{E}\,(\varepsilon_i \varepsilon_j) \neq 0$ при $i \neq j$ — автокорреляция ошибок

Спецификация модели

Нарушения спецификации:

1. Исключены существенные переменные:

$$\text{Процесс, порождающий данные:} \quad y = X\beta + Z\gamma + \varepsilon$$
$$\text{Модель:} \quad y = X\beta + \varepsilon$$

2. Включены несущественные переменные:

$$\text{Процесс, порождающий данные:} \quad y = X\beta + \varepsilon$$
$$\text{Модель:} \quad y = X\beta + Z\gamma + \varepsilon$$

$\dim(X) = n \times m$, $\dim Z = n \times l$, $\dim(y) = n \times 1$,
$\dim(\beta) = m \times 1$, $\dim(\gamma) = l \times 1$

Нарушение спецификации модели

Исключены существенные переменные

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

$$y = X\beta + Z\gamma + \varepsilon$$

Смещённая оценка β:

$$\mathsf{E}\, b = (X^\top X)^{-1} X^\top \mathsf{E}\, y = \beta + (X^\top X)^{-1} X^\top Z\gamma$$

Смещённая оценка дисперсии остатков:

$$\mathsf{E}\,(e^\top e) = \mathsf{E}\left(y^\top \underbrace{(I - X(X^\top X)^{-1} X^\top)}_{M} y\right) =$$

$$= \mathsf{E}\,(\varepsilon^\top M\varepsilon + 2\gamma^\top Z^\top M\varepsilon + \gamma^\top Z^\top MZ\gamma) =$$

$$= \sigma^2(n - m) + \gamma^\top Z^\top MZ\gamma$$

Нарушение спецификации модели

Включение несущественных переменных

$$b = (X^\top M_Z X)^{-1} X^\top M_Z y, \quad M_Z = I - Z(Z^\top Z)^{-1} Z^\top$$

Оценка β:

$$\mathsf{E}\, b = \beta$$

$$\mathsf{D}\, b = \sigma^2 (X^\top M_Z X)^{-1} =$$
$$= \sigma^2 \Big((X^\top X)^{-1} + (X^\top X)^{-1} X^\top Z (Z^\top M Z)^{-1} Z^\top X (X^\top X)^{-1} \Big)$$

Спецификация модели

Сравнение моделей

$$y = X\beta + z\gamma + \varepsilon \qquad \text{модель без ограничений}$$

$$y = X\beta + \varepsilon \qquad \text{модель с ограничениями}$$

$$F = \frac{(e_r^\top e_r - e_u^\top e_u)/1}{e_u^\top e_u/(n-m-1)} \sim F_{1,n-m-1}$$

Мультиколлинеарность

$$\mathrm{Cor}\,(x_i, x_j) \neq 0 \text{ при } i \neq j$$

Система нормальных уравнений

$$X^\top X b = X^\top y$$

плохо определена: решение неустойчиво.

1. Использование ридж-оценок:

$$(X^\top X - \alpha I_{m+1}) b = X^\top y$$

2. Использование сингулярного разложения (метод главных компонент):

$$X^\top X = USV^\top$$

Обобщённая линейная модель

Обобщённая линейная модель множественной регрессии

$$y = X\beta + \varepsilon$$

1. $y = X\beta + \varepsilon$ — спецификация модели
2. X — детерминированная матрица ранга $m + 1$
3. $\mathsf{E}\,\varepsilon = 0$, $\mathsf{E}\left(\varepsilon\varepsilon^\top\right) = \Omega$

Теорема Айткена

В классе линейных несмещённых оценок вектора β для обобщённой регрессионной модели оценка

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

имеет наименьшую ковариационную матрицу.

Гетероскедастичность

Тестирование

Тест ранговой корреляции Спирмена

Наблюдения ранжируются по x_i и e_i и вычисляется коэффициент ранговой корреляции

$$\rho_{x_i,e} = 1 - \frac{6}{n^3 - n} \sum_{i=1}^{n} d_i^2,$$

где d_i — разность между рангами x_i и e_i.

$$|\rho_{x_i,e}| \sqrt{\frac{n-2}{1-\rho_{x_i,e}^2}} \sim t_{n-2}$$

Гетероскедастичность

Тестирование

Тест Голдфелдта–Квандта

$$F = \frac{e_1^\top e_1}{e_2^\top e_2} \sim F_{n/2-d/2, n/2-d/2}$$

$d \sim n/3 \div n/4$

Тест Уайта

Проверяется значимость уравнения

$$e_i^2 = f(x_i) + \nu_i$$

Тест Глейзера

Проверяется значимость уравнения

$$|e_i| = f(x_i) + \nu_i$$

Гетероскедастичность

Устранение гетероскедастичности

Метод взвешенных наименьших квадратов

$$\frac{y_i}{\sigma_i} = \sum_{j=1}^{m} \beta_j \frac{x_{i,j}}{\sigma_i} + u_i, \quad i = 1, \ldots, n$$

- Если $\sigma_i^2 = \sigma^2 x_{ik}^2$, то $x_{ij}^* = x_{ij}/x_{ik}$ и $y_i^* = y_i/x_{ik}$, $i = 1, \ldots, m$, $j = 1, \ldots, m$.
- $\sigma_i^2 = \omega_1^2$, $i = 1, \ldots, n_1$ и $\sigma_i^2 = \omega_2^2$, $i = n_1 + 1, \ldots, n$:
    - Провести обычную регрессию, разбить вектор остатков на два подвектора e_1 и e_2.
    - Построить оценки $\hat{\omega}_1^2 = e_1^\top e_1/n_1$ и $\hat{\omega}_2^2 = e_2^\top e_2/n_2$.
    - Преобразовать переменные и провести обычную регрессию для преобразованной модели.
- Состоятельное оценивание дисперсий.

Автокорреляция остатков

Тестирование

$$\varepsilon_i = \rho\varepsilon_{i-1} + \nu$$

Тест Дарбина–Уотсона

на наличие автокорреляции первого порядка:

$$\mathrm{DW} = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \approx 2(1-r)$$

Тест серий (Бреуша–Годфри)

проверяется значимость коэффицинта ρ в модели

$$e_i = \rho e_{i-1} + \nu_i, \quad i = 2, \ldots, n$$

Автокорреляция остатков
Тестирование

Если ряд стационарный, то выборочный частный коэффициент корреляции r_p совпадает с оценкой обычного метода наименьших квадратов коэффициента β_p в авторегрессионной модели $\mathrm{AR}(p)$:

$$y_i = \beta_0 + \beta_1 y_{i-1} + \cdots + \beta_p y_{i-p} + \varepsilon_i.$$

Q-тест Льюинга–Бокса

$$Q_p = n(n+1)\sum_{i=1}^{p} \frac{r_i}{n-i} \sim \chi^2_p$$

Автокорреляция остатков

Коррекция данных

$$\varepsilon_i = \rho\varepsilon_{i-1} + \nu_i \quad \Rightarrow \quad \varepsilon_i - \rho\varepsilon_{i-1} = \nu_i$$

$$y_i - \rho y_{i-1} = (x_i - \rho x_{i-1})\beta + \nu_i, \quad i = 2, 3, \ldots, n$$
$$\sqrt{1-\rho^2}y_1 = \sqrt{1-\rho^2}x_1\beta + \sqrt{1-\rho^2}\varepsilon_1$$

Автокорреляция остатков

Оценка параметра авторегрессии ρ

1. *Процедура Кохрейна–Оркатта*

$$e_i = \rho e_{i-1} + \nu_i$$

2. *Процедура Хилдрета–Лу*
   Сеточный поиск
3. *Процедура Дарбина*

$$y_i = \beta_1(1-\rho) + \rho y_{i-1} + \beta_2 x_{i,2} - \rho\beta_2 x_{i-1,2} + \dots + \\ + \beta_m x_{i,m} - \rho\beta_m x_{i-1,m} + \nu_i$$

y_{i-1} включается в число регрессоров, а ρ — в число оцениваемых параметров.

Регрессионные модели

1. Модели, линейные по параметрам

$$y = \sum_{i=1}^{m} \beta_i \phi_i(x) + \varepsilon$$

2. Линеаризуемые модели

$$y = f(x, \varepsilon) \quad \Rightarrow \quad \tilde{y} = \sum_{i=1}^{m} \beta_i \phi_i(x) + \varepsilon$$

3. Существенно нелинейные модели

$$y = f(x) + \varepsilon$$

Модели с фиктивными переменными

Модели, имеющие в своём составе качественные независимые переменные.

Качественная переменная, имеющая k значений, заменяется $(k-1)$-й бинарной переменной (принимающей значения 0 или 1).

	x_1	x_2	x_3
зима	0	0	0
весна	1	0	0
лето	0	1	0
осень	0	0	1

Модели с фиктивными переменными

- *Аддитивная модель*:

$$y = \beta x + \tilde{\beta}\tilde{x} + \varepsilon$$

  (сдвиг по вертикали на величину $\tilde{\beta}$)
- *Мультипликативная модель*:

$$y = \beta(1 + \tilde{\beta}\tilde{x})x + \varepsilon$$

  (изменение угла наклона)
- *Смешанная модель*