Регрессионная модель в матричном виде

В матричной форме регрессионная модель имеет вид:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \qquad (1)$$

- где Y - случайный вектор - столбец размерности (n x 1)
- X - матрица размерности [n x (k+1)] наблюдаемых значений аргументов.
- β - вектор - столбец размерности [(k+1) x 1] неизвестных, подлежащих оценке параметров (коэффициентов регрессии) модели;
- ε - случайный вектор - столбец размерности (n x 1) ошибок наблюдений (остатков).

Основы регрессионного анализа

Исходные статистические данные могут быть представлены в виде вектора значений результативной переменной $Y = (y_1, \ldots, y_i, \ldots, y_n)^T$ и матрицы **X** значений объясняющих переменных

$$X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1j} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2j} & \ldots & x_{2k} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{i1} & \ldots & x_{ij} & \ldots & x_{ik} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & \ldots & x_{nj} & \ldots & x_{nk} \end{pmatrix}$$

размерности $(n \times (k+1))$, где x_{ij} – значение j–й объясняющей переменной для i-го наблюдения.

Единицы в первом столбце матрицы необходимы для обеспечения свободного члена в регрессионной модели.

Основы регрессионного анализа

$$\mathbf{Y = X\beta + \varepsilon}$$

$$X = \begin{pmatrix} 1 & x_{11} & . & . & x_{1k} \\ 1 & x_{i1} & . & . & x_{ik} \\ 1 & x_{n1} & . & . & x_{nk} \end{pmatrix} \qquad \begin{pmatrix} y_1 \\ y_i \\ y_n \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_j \\ \beta_k \end{pmatrix}$$

Основная задача регрессионного анализа заключается в нахождении по выборке объемом n оценки неизвестных коэффициентов регрессии $\beta_0, \beta_1, \ldots, \beta_k$

Регрессионная модель в матричном виде

Так как x_j - неслучайные величины,

$M\varepsilon_i=0$,

оценка уравнения регрессии в матричной форме имеет вид:

$$\tilde{Y} = Xb$$

- где - $\tilde{Y}$ вектор-столбец с элементами

$$(\tilde{y}_1,\ldots,\tilde{y}_i,\ldots,\tilde{y}_n)^T$$

Регрессионная модель в матричном виде

- Для оценки вектора b наиболее часто используют метод наименьших квадратов (МНК)

$$Q = \sum_{i=1}^{n} \left(y_i - \tilde{y}_i \right)^2 = (Y - Xb)^T (Y - Xb) \Rightarrow \min_{b_0, b_1, \ldots, b_k}$$

Регрессионная модель в общем виде

- Дифференцируя квадратичную форму Q по $b_0, b_1, \ldots, b_к$ и приравнивая производные к нулю, получим систему нормальных уравнений:

$$\begin{cases} \dfrac{\partial Q}{\partial b_j} = 0 \\ j = 0,1,\ldots,к \end{cases}$$

- Решая которую, получим вектор оценок b, где $b=(b_0\ b_1 \ldots b_к)^T$

$$b = (X^T X)^{-1} X^T Y \qquad (2)$$

Свойства оценки

Из (2) с учетом (1) будем иметь:

$$b = \hat{\beta}_{МНК} = \left(X^T X\right)^{-1} X^T \left(X\beta + \varepsilon\right)$$

$$\hat{\beta}_{МНК} = \beta + \left(X^T X\right)^{-1} X^T \varepsilon \qquad (4)$$

$$M\hat{\beta} = \beta + (X^T X)^{-1} X^T M\varepsilon$$

$$M\varepsilon = 0$$

$$M\hat{\beta}_{МНК} = \beta$$

Таким образом, b - несмещенная оценка β

Пример

- Пусть $y_i = \beta x_i + \varepsilon_i$, i=1,2,…,n
- Определить МНК-оценку $\hat{\beta}_{МНК}$ параметра β

$$\tilde{y}_i = My_i = bx_i$$

$$Q = \sum_{i=1}^{n} (y_i - bx_i)^2 \qquad \frac{\partial Q}{\partial b} = 2\sum_{i=1}^{n} (y_i - bx_i)\cdot(-x_i) = 0$$

$$\sum_{i=1}^{n} x_i y_i = b \cdot \sum_{i=1}^{n} x_i^2$$

$$\hat{\beta}_{МНК} = b = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Оценка ковариационной матрицы

- Оценка ковариационной матрицы коэффициентов регрессии вектора b определяется из выражения:

$$S(b) = \hat{S}^2 (X^T X)^{-1}$$

- На главной диагонали ковариационной матрицы находятся дисперсии коэффициентов регрессии.

$$\hat{S}^2 = \frac{1}{n-\kappa-1}(Y-Xb)^T(Y-Xb) = \frac{1}{n-\kappa-1}\sum(y_i-\hat{y}_i)^2$$

Например, найдено

$$\hat{s}^2 = 0{,}498$$

$$(X^T X)^{-1} = \begin{pmatrix} 0{,}31 & -0{,}027 \\ -0{,}027 & 0{,}0037 \end{pmatrix}$$

$$\hat{S}(b) = \hat{s}^2 (X^T X)^{-1} = 0{,}498 \begin{pmatrix} 0{,}31 & -0{,}027 \\ -0{,}027 & 0{,}0037 \end{pmatrix} = \begin{pmatrix} 0{,}15 & -0{,}014 \\ -0{,}014 & 0{,}0018 \end{pmatrix}$$

тогда оценка ковариационной матрицы

$$\hat{s}^2_{b_0} = 0{,}15 \qquad \hat{s}_{b_0} = 0{,}39$$

$$\hat{s}^2_{b_1} = 0{,}0018 \qquad \hat{s}_{b_1} = 0{,}042$$

Проверка значимости уравнения регрессии

- H_0: $\beta=0$ $(\beta_0=\beta_1=...=\beta_к=0)$, проверяется по F-критерию Фишера

$$F_{набл} = \frac{Q_R/(к+1)}{Q_{ост}/(n-к-1)}$$

- где

$$Q_R = (Xb)^T(Xb),$$

$$Q_{ост} = (Y-Xb)^T(Y-Xb) = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

Проверка значимости уравнения регрессии

2. По таблице F-распределения находят $F_{кр}$ для заданных α, $\nu_1=к+1$, $\nu_2=n-к-1$
3. Если $F_{набл}>F_{кр}$, то гипотеза H_0 отклоняется с вероятностью ошибки α.

Из этого следует, что уравнение регрессии является значимым, т. е. хотя бы один из коэффициентов регрессии отличен от нуля.

Проверка значимости отдельных коэффициентов регрессии

- В случае значимости уравнения регрессии представляет интерес проверка значимости отдельных коэффициентов регрессии и построение для них интервальных оценок.
- Значимость коэффициентов регрессии можно проверить с помощью t-критерия, основанного на статистике:

$$t_j(набл) = \frac{b_j}{\hat{s}[(X^T X)^{-1}]^{1/2}} = \frac{b_j}{s_{b_j}}$$

- которая при выполнении гипотезы $H_0: \beta_j = 0$
- имеет t-распределение (распределение Стьюдента)

Проверка значимости коэффициентов регрессии

2. $t_{кр}(\alpha,\ \nu = n-к-1)$
3. Гипотеза H_0 отвергается с вероятностью α,

$$\left|t_{набл}\right| > t_{кр}(\alpha, \nu = n - к - 1)$$

- В противном случае коэффициент регрессии незначим и соответствующая переменная в модель не включается.

Интервальное оценивание коэффициентов регрессии

Интервальная оценка с доверительной вероятностью γ для параметра β_j имеет вид:

$$b_j - t_\alpha S_{b_j} \le \beta_j \le b_j + t_\alpha S_{b_j}$$

где t_α находят по таблице t-распределения Стьюдента

при $\alpha = 1-\gamma$ и $\nu = n-к-1$.

Явление мультиколлинеарности

Мультиколлинеарность - это негативное явление, обусловленное тесной взаимосвязью объясняющих переменных $x_1, x_2, \ldots, x_k$

- 1. При наличии мультиколлинеарности матрица **(X^TX)** становится **слабо обусловленной**, т.е. ее определитель близок к нулю.
- 2. Нахождение обратной матрицы связано с делением на определитель (т.е. величину близкую к нулю). Следовательно, все решения становятся неустойчивыми.

Явление мультиколлинеарности

3. ⇒ вектор $b=(b_0\ b_1 ... b_к)^T$

содержит элементы, знаки которых не поддаются содержательной интерпретации.

4. Находящиеся на главной диагонали ковариационной матрицы $S(b) = \hat{S}^2 (X^T X)^{-1}$

дисперсии $\hat{s}^2_{bj}$ могут оказаться неоправданно завышенными

5. В этой связи значимые коэффициенты βj могут оказаться статистически незначимыми, т.к.

$$t_j = \frac{b_j}{\hat{s}_{bj}}$$

Явление мультиколлинеарности

6. Мультиколлинеарность ведет к неоправданно **завышенному множественному коэффициенту корреляции R_y**

$$r_y = \sqrt{1 - \frac{|R|}{R_{yy}}}$$

Наличие мультиколлинеарности можно проверить по матрице парных коэффициентов корреляции

$$R=(r_{jl})\ j,l=1,2,\ldots,p.$$

О мультиколлинеарности говорят, если $r_{jl}>0,8$ (0<85). В этом случае при построении регрессии в модель необходимо включить либо x_j , либо x_l. Избавиться от мультиколлинеарности позволяют пошаговые алгоритм регрессионного анализа (*метод пошагового включения переменных*).