Statystyka matematyczna - wykład dziesiąty[1]
Klasyczny model regresji liniowej II.
Jednoczynnikowa analiza wariancji.
kierunek: matematyka I°
specjalność: matematyka finansowa

dr Jarosław Kotowicz

Instytut Informatyki Uniwersytet w Białymstoku

[1] ©J.Kotowicz

Spis treści

Dokładność dopasowania prostej MNK. I

Odchylenie obserwowane wartości Y_i od średniej $\overline{Y}$ może być przedstawione, jako suma dwóch składników, z których pierwszy jest wyjaśniany regresją liniową Y względem X i reszt modelu (e_i) tzw. losowej części odchylenia nie wyjaśnianej regresją.
Zapisujemy to

$$Y_i - \overline{Y} = (\widehat{Y_i} - \overline{Y}) + (Y_i - \widehat{Y_i}). \tag{1}$$

Podnosząc obie strony równości do kwadratu, a następnie sumując po i otrzymujemy równanie

$$\sum_{i=1}^{n}(Y_i-\overline{Y})^2 = \sum_{i=1}^{n}(\widehat{Y_i}-\overline{Y})^2+2\sum_{i=1}^{n}(\widehat{Y_i}-\overline{Y})(Y_i-\widehat{Y_i})+\sum_{i=1}^{n}(Y_i-\widehat{Y_i})^2. \tag{2}$$

Udowodnimy, że środkowy składnik sumy równa się zero.

Dokładność dopasowania prostej MNK. II

Skorzystamy w tym celu z warunków

$$\begin{cases} \widehat{\beta_0} = \overline{Y} - \widehat{\beta_1}\overline{x} \\ \widehat{Y_i} = \widehat{\beta_1} x_i + \widehat{\beta_0} \end{cases}.$$

Stąd

$$\widehat{Y_i} - \overline{Y} = \widehat{\beta_1}(x_i - \overline{x}) \quad \text{oraz} \quad \widehat{Y_i} = \overline{Y} + \widehat{\beta_1}(x_i - \overline{x}).$$

Mamy wtedy

$$\sum_{i=1}^{n}(\widehat{Y_i} - \overline{Y})(Y_i - \widehat{Y_i}) = \widehat{\beta_1}\left[\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y}) - \widehat{\beta_1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right].$$

Wstawiając wartość estymatora $\widehat{\beta_1}$ otrzymujemy żądaną tezę.

Dokładność dopasowania prostej MNK. III

Stąd ostatecznie otrzymujemy równanie

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\widehat{Y_i} - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2. \qquad (3)$$

Współczynnik deterministyczny

Miarą dokładności dopasowania prostej jest współczynnik deterministyczny, który definiujemy jedną z równości

$$r^2 := \frac{\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \equiv 1 - \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}. \tag{4}$$

Współczynnik ten ma następujące własności

- $r^2 \in [0, 1]$,
- $r^2 = 1$ wtedy, gdy między zmiennymi X i Y zachodzi zależność liniowa (wszystkie punkty empiryczne leżą na prostej),
- $r^2 = 0$, gdy $\widehat{\beta_1} = 0$, czyli $\widehat{Y}_i = \widehat{\beta_0} = \overline{Y}$ (znajomość wartości zmiennej X nie dostarcza żadnych informacji na temat wartości zmiennej zależnej Y).

Wnioskowanie o klasycznym modelu normalnej regresji liniowej

Załóżmy, że warunkowe rozkłady zmiennej zależnej są normalne (składniki losowe modelu ε_i mają rozkład $\mathcal{N}(0, \sigma)$).
Parametry $\widehat{\beta_1}$ i $\widehat{\beta_0}$ mają rozkłady $\mathcal{N}(\beta_1, \mathbb{D}(\widehat{\beta_1}))$ i $\mathcal{N}(\beta_0, \mathbb{D}(\widehat{\beta_0}))$.
Konstruujemy statystyki dla nich

$$\begin{cases} t = \frac{\widehat{\beta_1} - \beta_1}{s^{\widehat{\beta_1}}} \\ t = \frac{\widehat{\beta_0} - \beta_0}{s^{\widehat{\beta_0}}} \end{cases} . \tag{5}$$

Są one rozkładami t-Studenta o $n-2$ stopniach swobody.
Dla współczynnika ufności $1-\alpha$ odpowiadające im przedział ufności wynoszą

$$]\widehat{\beta_1} - t_{\alpha,n-2} S_{\widehat{\beta_1}}, \widehat{\beta_1} + t_{\alpha,n-2} S_{\widehat{\beta_1}}[,$$
$$]\widehat{\beta_0} - t_{\alpha,n-2} S_{\widehat{\beta_0}}, \widehat{\beta_0} + t_{\alpha,n-2} S_{\widehat{\beta_0}}[.$$

Test do weryfikacji hipotezy o parametrze β_1

$$H_0 : \beta_1 = \beta_1^0$$
$$H_1 : \beta_1 \neq \beta_1^0.$$

Przy założeniu prawdziwości hipotezy zerowej statystka ma postać

$$t = \frac{\widehat{\beta_1} - \beta_1^0}{s^{\widehat{\beta_1}}},$$

zaś obszar krytyczny dla poziomu istotności α opisany jest równaniem

$$P(\{|t| \geqslant t_{\alpha,n-2}\}) = \alpha.$$

Test do weryfikacji hipotezy o parametrze β_0

$$H_0 : \beta_0 = \beta_0^0$$
$$H_1 : \beta_0 \neq \beta_0^0.$$

Przy założeniu prawdziwości hipotezy zerowej statystka ma postać

$$t = \frac{\widehat{\beta_0} - \beta_0^0}{s^{\widehat{\beta_0}}},$$

zaś obszar krytyczny dla poziomu istotności α opisany jest równaniem

$$P(\{|t| \geqslant t_{\alpha,n-2}\}) = \alpha.$$

Uwagi

1. Najczęściej stosowaną wersją testu istotności dla β_1 jest $\beta_1^0 = 0$.
2. Najczęściej hipotezę dotyczącą wyrazu wolnego (β_0) pomijamy.

Analiza wariancji w modelu regresji

Podstawą analizy wariancji jest równanie

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\widehat{Y_i} - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2. \qquad (6)$$

Otrzymujemy z niego tzw. **tablicę analizy wariancji**.

Tablica analizy wariancji

Źródło zmienności	Suma kwadratów	Stopnie swobody	Średni kwadrat	Statystyka F
Regresja	$\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$	1	$\frac{\sum_{i=1}^{n}(\widehat{Y}_i-\overline{Y})^2}{1}$	$\frac{\sum_{i=1}^{n}(\widehat{Y}_i-\overline{Y})^2}{S_e^2}$
Reszta	$\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$	$n-2$	$\frac{\sum_{i=1}^{n}(Y_i-\widehat{Y}_i)^2}{n-2}$	
Całkowita	$\sum_{i=1}^{n}(Y_i - \overline{Y})^2$	$n-1$		

Hipoteza testowana to:

$$H_0 : \beta_1 = 0,$$
$$H_1 : \beta_1 \neq 0.$$

Statystyka z jaką mamy do czynienia, to statystyka F-Snedecora

$$\frac{\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2}{\frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{n-2}}$$

z liczbą stopni swobody licznika 1 i mianownika $n-2$.
Obszar krytyczny przy poziomie istotności α zadaje równość

$$P(\{F_{1,n-2} \geqslant F_{\alpha;1,n-2}\}) = \alpha.$$

Można udowodnić, że $F_{1,n-2} = t^2_{n-2}$.

Przypomnienie

Będziemy rozpatrywać klasyczny model regresji liniowej zadany warunkami zapisany w postaci alternatywnej

$$Y_i = \beta_1 x_i + \beta_2 + \varepsilon_i, \tag{7}$$

$$\mathbb{E}(\varepsilon_i) = 0 \tag{8}$$

$$\mathbb{D}^2(\varepsilon_i) = \mathbb{E}(\varepsilon_i^2) = \sigma^2, \tag{9}$$

$$\operatorname{cov}(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0 \text{ dla dowolnych } i \neq j, \tag{10}$$

gdzie $i, j \in \overline{1, n}$.

Sformułowanie modelu

Klasyczne model regresji liniowej może być zapisany w następującej postaci macierzowej

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_0 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}. \tag{11}$$

W skróconym zapisie macierzowym mamy

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \tag{12}$$

gdzie

- $\mathbf{Y}$ jest wektorem obserwacji zmiennej losowej Y o wymiarach $n \times 1$,
- $\mathbf{X}$ jest macierzą obserwacji dla zmiennej niezależnej X o wymiarach $n \times 2$,
- β jest wektorem współczynników o wymiarach 2×1,
- ε jest wektorem składników losowych o wymiarach $n \times 1$.

Założenia klasycznego modelu regresji liniowej mają postać

$$\mathbb{E}(\varepsilon) = \Theta \tag{13}$$

$$\mathbb{E}(\varepsilon\varepsilon^T) = \sigma^2 I, \tag{14}$$

gdzie zero w pierwszym równaniu jest wektorem zerowym o wymiarze $n \times 1$, zaś I jest macierzą jednostkową stopnia n, a $\cdot^T$ jest transponowaniem macierzy.

Macierz kowariancji składników losowych

Uwaga 1

Macierz $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T)$ nazywamy macierzą kowariancji składników losowych.

Zauważmy, że dla dowolnych $i, j \in \overline{1, n}$ mamy

$$\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T)_{ij} = \mathbb{E}(\varepsilon_i \varepsilon_j) = \operatorname{cov}(\varepsilon_i, \varepsilon_j).$$

Warunek nielosowości zmiennej niezależnej. I

Ponieważ mamy założone, że wartości zmiennej niezależnej są nielosowe (deterministyczne), więc należy ten warunek ująć w ujęciu macierzowy modelu regresji liniowej.

$$\mathbf{X} \text{ jest macierzą o wymiarach } n \times 2 \text{ o ustalonych elementach.} \tag{15}$$

Aby ustalić wartość współczynników występujących w regresji liniowej musimy założyć, że rząd macierzy $\mathbf{X}$ jest równy 2, co odpowiada założeniu, że w próbie są co najmniej dwie obserwacje dokonane dla różnych wartości x.
W ujęciu macierzowym wyrażenie podlegające minimalizacji metodą najmniejszy kwadratów jest postaci

$$SSE = \varepsilon^T \varepsilon = (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta). \tag{16}$$

Warunek nielosowości zmiennej niezależnej. II

Różniczkując względem wektora $\boldsymbol{\beta}$ otrzymujemy

$$\frac{\partial}{\partial \boldsymbol{\beta}} SSE = -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}. \tag{17}$$

Korzystając z warunku koniecznego istnienia ekstremum otrzymujemy równanie

$$\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}, \tag{18}$$

które można zapisać w jawnej postaci macierzowej

$$\begin{bmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{bmatrix} \begin{bmatrix} \widehat{\beta_1} \\ \widehat{\beta_0} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} x_i Y_i \\ \sum_{i=1}^{n} Y_i \end{bmatrix}. \tag{19}$$

Warunek nielosowości zmiennej niezależnej. III

Wyznaczając z równania (18) wektor $\widehat{\beta}$ otrzymujemy

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{20}$$

gdzie macierz $(\mathbf{X}^T\mathbf{X})^{-1}$ jest postaci

$$\begin{bmatrix} \frac{1}{\sum_{i=1}^{n}(x_i-\overline{x})^2} & \frac{-\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n}(x_i-\overline{x})^2} \\ \frac{-\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n}(x_i-\overline{x})^2} & \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i-\overline{x})^2} \end{bmatrix}. \tag{21}$$

Warunek nielosowości zmiennej niezależnej. IV

Na podstawie wyznaczonej z próby wektora $\widehat{\boldsymbol{\beta}}$ wyznaczamy wektor $\widehat{\mathbf{Y}}$ teoretycznych wartości zmiennej losowej Y i wektor reszt $\mathbf{e}$

$$\begin{aligned} \widehat{\mathbf{Y}} &= \mathbf{X}\widehat{\boldsymbol{\beta}} \\ \mathbf{e} &= \mathbf{Y} - \widehat{\mathbf{Y}}. \end{aligned}$$

Ponieważ sumę kwadratów reszt można przedstawić wzorem

$$\sum_{i=1}^{n} e_i^2 = \mathbf{e}^T\mathbf{e},$$

więc nieobciążony estymator wariacji jest postaci

$$S_e^2 = \frac{\mathbf{e}^T\mathbf{e}}{n-2}.$$

Warunek nielosowości zmiennej niezależnej. V

Macierz kowariancji wektora losowego $\widehat{\boldsymbol{\beta}}$ definiujemy

$$V(\widehat{\boldsymbol{\beta}}) = \mathbb{E}((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})) \equiv \begin{bmatrix} \mathbb{D}^2(\widehat{\beta_1}) & \operatorname{cov}(\widehat{\beta_0}, \widehat{\beta_1}) \\ \operatorname{cov}(\widehat{\beta_0}, \widehat{\beta_1}) & \mathbb{D}^2(\widehat{\beta_0}) \end{bmatrix}.$$

Stwierdzenie 1

W klasycznym modelu regresji liniowej macierz $V(\widehat{\boldsymbol{\beta}})$ jest postaci $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

Warunek nielosowości zmiennej niezależnej. VI

Na podstawie tego mamy

$$V(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix} \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2} & \frac{-\sigma^2 \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n}(x_i-\overline{x})^2} \\ \frac{-\sigma^2 \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n}(x_i-\overline{x})^2} & \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i-\overline{x})^2} \end{bmatrix}.$$

Nieobciążonym estymatorem macierzy $V(\widehat{\boldsymbol{\beta}})$ jest macierz

$$\widehat{V}(\widehat{\boldsymbol{\beta}}) = S_e^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

Spis treści

Sformułowanie zagadnienia

Rozważamy zmienną $(k+1)$-wymiarową $(Y, X_1, \ldots, X_k)$, gdzie $X_1, \ldots, X_k$ są zmiennymi niezależnymi, a Y zmienną zależną.
Do opisu stosujemy klasyczny model regresji liniowej, o ile dla każdego układu wartości $x_1, \ldots, x_k$ warunkowe rozkłady zmiennej Y mają średnie

$$\mathbb{E}(Y|x_1, \ldots, x_k) = \beta_1 x_1 + \ldots + \beta_k x_k + \beta_{k+1}$$

oraz wariancję

$$\mathbb{D}^2(Y|x_1, \ldots, x_k) = \sigma^2.$$

Jeżeli dodatkowo warunkowe rozkłady zmiennej Y miałyby rozkład normalny, to mówilibyśmy o normalnej regresji liniowej.
Próbę losową stanowiącą podstawę sformułowania i oszacowania modelu określa n łącznych obserwacji postaci

$$(Y_i, x_{i1}, \ldots, x_{ik}), \quad i \in \overline{1, n}.$$

Model

Będziemy więc rozpatrywać model zadany warunkami

$$Y_i = \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \beta_{k+1} + \varepsilon_i, \tag{22}$$

$$\mathbb{E}(\varepsilon_i) = 0 \tag{23}$$

$$\mathbb{D}^2(\varepsilon_i) = \mathbb{E}(\varepsilon_i^2) = \sigma^2, \tag{24}$$

$$\operatorname{cov}(\varepsilon_i, \varepsilon_j) = \mathbb{E}(\varepsilon_i \varepsilon_j) = 0 \text{ dla dowolnych } i \neq j, \tag{25}$$

gdzie $i, j \in \overline{1, n}$.

Założenie 1

Będziemy zakładać, że $k + 1 < n$ tzn. liczba obserwacji jest większa od liczby parametrów modelu.

Postać macierzowa

Klasyczne model regresji liniowej może być zapisany w następującej postaci macierzowej

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1k} & 1 \\ x_{21} & \dots & x_{2k} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{nk} & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{k+1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}. \tag{26}$$

W skróconym zapisie macierzowym mamy

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \tag{27}$$

gdzie

- $\mathbf{Y}$ jest wektorem obserwacji zmiennej losowej Y o wymiarach $n \times 1$,
- $\mathbf{X}$ jest macierzą obserwacji dla zmiennej niezależnej X o wymiarach $n \times (k+1)$,
- β jest wektorem współczynników o wymiarach $(k+1) \times 1$,
- ε jest wektorem składników losowych o wymiarach $n \times 1$.

Założenia klasycznego modelu regresji liniowej mają postać

$$\mathbb{E}(\varepsilon) = \Theta \tag{28}$$
$$\mathbb{E}(\varepsilon\varepsilon^T) = \sigma^2 I, \tag{29}$$

gdzie zero w pierwszym równaniu jest wektorem zerowym o wymiarze $n \times 1$, zaś I jest macierzą jednostkową stopnia n.

Założenie 2

Będziemy dodatkowo zakładać, że macierz $\mathbf{X}$ *jest macierzą pełnego rzędu. Oznacza to, że łącznie z założeniem 1 rząd macierzy* $\mathbf{X}$ *równy jest* $k+1$ *tzn.* $rz(\mathbf{X}) = k+1$.

Warunek nielosowości zmiennej niezależnej. I

Ponieważ mamy założone, że wartości zmiennych niezależnych są nielosowe (deterministyczne), więc należy ten warunek ująć w ujęciu macierzowy modelu regresji liniowej.

$$\mathbf{X} \text{ jest macierzą o wymiarach } n \times (k+1) \text{ o ustalonych elementach.} \quad (30)$$

Aby ustalić wartość współczynników występujących w regresji liniowej musimy założyć, że rząd macierzy $\mathbf{X}$ jest równy $k+1$, co odpowiada założeniu, że w próbie są co najmniej $k+1$ obserwacje dokonane dla różnych wartości x.
Podobnie jak w przypadku dwóch zmiennych wyrażenie podlegające minimalizacji metodą najmniejszy kwadratów jest postaci

$$SSE = \varepsilon^T \varepsilon = (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta). \quad (31)$$

Warunek nielosowości zmiennej niezależnej. II

Otrzymujemy

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{Y}, \tag{32}$$

które można zapisać w jawnej postaci macierzowej

$$\begin{bmatrix} \sum x_{i1}^2 & \sum x_{i1}x_{i2} & \dots & \sum x_{i1}x_{ik} & \sum x_{i1} \\ \sum x_{i2}x_{i1} & \sum x_{i2}^2 & \dots & \sum x_{i2}x_{ik} & \sum x_{i2} \\ \vdots & \vdots & & \vdots & \vdots \\ \sum x_{i1} & \sum x_{i2} & \dots & \sum x_{ik} & n \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{k+1} \end{bmatrix} = \begin{bmatrix} \sum x_{i1}Y_i \\ \sum x_{i2}Y_i \\ \vdots \\ \sum Y_i \end{bmatrix}.$$

Z założeń 1 i 2 wynika, że macierz $\mathbf{X}^T\mathbf{X}$ jest odwracalna, więc możemy wyznaczyć z ostatniego równania wektor $\widehat{\boldsymbol{\beta}}$. Otrzymujemy

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{33}$$

Warunek nielosowości zmiennej niezależnej. III

Na podstawie wyznaczonej z próby wektora $\widehat{\boldsymbol{\beta}}$ wyznaczamy wektor $\widehat{\mathbf{Y}}$ teoretycznych wartości zmiennej losowej Y i wektor reszt $\mathbf{e}$

$$\begin{aligned} \widehat{\mathbf{Y}} &= \mathbf{X}\widehat{\boldsymbol{\beta}} \\ \mathbf{e} &= \mathbf{Y} - \widehat{\mathbf{Y}}. \end{aligned}$$

Nieobciążony estymator wariacji jest postaci

$$S_e^2 = \frac{\mathbf{e}^T\mathbf{e}}{n-k-1}.$$

Macierz kowariancji wektora losowego $\widehat{\boldsymbol{\beta}}$ definiujemy

$$V(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1},$$

a jej estymator to

$$V(\widehat{\boldsymbol{\beta}}) = S_e^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

Współczynnik korelacji wielorakiej

Podobnie jak w przypadku dwóch zmiennych mamy współczynnik determinacji

$$r^2 := \frac{\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \equiv 1 - \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}. \tag{34}$$

Natomiast dodatni pierwiastek z współczynnika determinacji nazywany jest współczynnikiem korelacji wielorakiej.
Współczynnik determinacji ma następujące własności

- $r^2 \in [0, 1]$,
- $r^2 = 1$ wtedy, gdy wszystkie punkty leżą w hiperpłaszczyźnie,
- $r^2 = 0$ – znajomość wartości zmiennych $X_1, \ldots, X_k$ nie dostarczają żadnych informacji na temat wartości zmiennej zależnej Y.

Uwagi. I

Założenia i ich testowanie (zobacz [1]):

1. zmienne niezależne (predyktory) nie są ze sobą silnie skorelowane (sposób weryfikacji: analiza współczynnika korelacji),
2. mamy do czynienia z zależnością liniową (sposób weryfikacji: analizę wykresu rozrzutu (rozrzutów) pomiędzy predyktorami a zmienną zależną),
3. brak znaczących obserwacji odstających (sposób weryfikacji: inspekcja wykresów punktowych, IQR, z-score, odległość Cooka, test Grubbs'a, test Dixona),
4. liczba obserwacji musi być większa bądź równa liczbie parametrów wyprowadzonych z analizy regresji (współczynniki dla predyktorów, wyraz wolny),

Uwagi. II

5. wariancja reszt, składnika losowego jest taka sama dla wszystkich obserwacji – homoskedastyczność (sposób weryfikacji: test Goldfelda-Quandta, test Breuscha-Pagana; dla dwóch prób: test Fishera F[2]; dla wielu prób: testy Barletta[3], Flingera-Killeena, Levene'a[4], Browna-Forsythe'a, Hartley'a[5]),
6. nie występuje autokorelacja reszt, składnika losowego (sposób weryfikacji: test Durina-Watsona),
7. reszty mają rozkład zbliżony do rozkładu normalnego (sposób weryfikacji: test Shapiro-Wilka, test Kołmogorowa-Smirnowa, test Jarque'a-Berry),
8. brak współliniowości predyktorów - regresja wieloraka (sposób weryfikacji: współczynnik VIF).

Jeśli wiele z założeń jest niespełniony nie korzystamy z przedstawionych metod weryfikacji

Uwagi. III

- bardziej adekwatny skorygowany współczynnik determinacji (także stosowalny gdy nie ma wyrazu wolnego).

Metody doboru zmiennych do modelu

- zmienne wybiera się na podstawie wiedzy dziedzinowej,
- wymagania dotyczące własności zmiennych niezależnych:
  1. są silnie skorelowanych ze zmienną, którą objaśniają,
  2. są nieskorelowane lub co najwyżej słabo skorelowane ze sobą,
  3. charakteryzują się dużą zmiennością.

W literaturze przyjmuje się, że budując model regresji powinno być co najmniej 15 obserwacji na każdą zmienną. Wtedy można uzyskać dobry model.

[2] Założenia: normalność.
[3] Założenia: normalność, równa liczebność grup.
[4] Założenia: niezależność prób.
[5] Założenia: normalność, równa liczebność grup.

Spis treści

Wybrane typy regresji

1. Regresja nieliniowa (np. wielomianowa).
2. Regresja logistyczna.
3. Regresja porządkowa.

Regresja nieliniowa

Regresja nieliniowa i transformacje do modelu liniowego.

- Między zmienną objaśnianą a zmiennymi objaśniającymi mogą zachodzić związki nieliniowe.
- W wielu przypadkach można dokonać transformacji do modelu liniowego poprzez odpowiednie przekształcenia zmiennych.
- Model $Y = f(X, b)$ jest liniowy względem parametrów, jeśli można go przedstawić jako liniową funkcję jednoznacznych przekształceń X, przy czym współczynniki tych przekształceń muszą być znane.

Typowe modele nieliniowe i ich transformacje do modelu liniowego. I

Poza modelem regresji liniowej występują także modele regresji nieliniowej. Są nimi między innymi

- model wielomianowy (wielomian stopnia k)

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_k X^k + \varepsilon.$$

Podstawiając $V_j = X^j$ dla $j = 1, 2, \ldots, k$, model sprowadza się do modelu liniowego.

Typowe modele nieliniowe i ich transformacje do modelu liniowego. II

- model potęgowy

$$Y = \beta_0 X_1^{\beta_1} \cdot X_2^{\beta_2} \cdot \ldots \cdot X_k^{\beta_k} e^{\varepsilon}.$$

Logarytmując otrzymujemy

$$\ln Y = \ln \beta_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 + \ldots + \beta_k \ln X_k + \varepsilon.$$

Podstawiając $V_j = \ln X_j$ dla $j = 1, 2, \ldots, k$ i $Z = \ln Y$, model sprowadza się do modelu liniowego.

Typowe modele nieliniowe i ich transformacje do modelu liniowego. III

- model wykładniczy

$$Y = \beta_0 \cdot \beta_1^{X_1} \cdot \beta_2^{X_2} \cdot \ldots \cdot \beta_k^{X_k} \cdot e^{\varepsilon}.$$

Logarytmując otrzymujemy

$$\ln Y = \ln \beta_0 + X_1 \ln \beta_1 + X_2 \ln \beta_2 + \ldots + X_k \ln \beta_k + \varepsilon.$$

Podstawiając $Z = \ln Y$ i $\tilde{\beta}_j = \ln \beta_j$ dla $j = 1, 2, \ldots, k$, model sprowadza się do modelu liniowego.

Krokowa konstrukcja modelu regresji.

Definicja 1

Krokowa konstrukcja modelu regresji polega na wprowadzaniu do modelu jedynie istotnych statystycznie predyktory, które „poprawiają" zbudowany model.

1. Postępująca (forward).
   - Zakłada kolejne dołączanie do listy zmiennych objaśniających tych zmiennych, które mają najistotniejszy wpływ na zmienną zależną.
2. Wsteczna (backward).
   - Usuwamy ze zbioru zmiennych, ta które mają najmniejszy wpływ na zmienną zależną.
   - Stosując r^2 lub testy istotności współczynników modelu (F).

Spis treści

Wprowadzenie. I

Rozważmy zagadnienie porównywania kilku próbek. Chodzi o sprawdzenie, czy wszystkie pochodzą z tej samej populacji, czy też z populacji o różnych średnich. Najprostszy model zakłada, że mamy kilka niezależnych próbek z rozkładów normalnych.
Analiza wariancji, ANOVA (**AN**alysis **O**f **VA**riance) — metoda statystyczna służąca do badania obserwacji, które zależą od jednego lub wielu działających równocześnie czynników. Metoda ta wyjaśnia, z jakim prawdopodobieństwem wyodrębnione czynniki mogą być powodem różnic między obserwowanymi średnimi grupowymi.
Modele analizy wariancji można podzielić na:

1. jednoczynnikowe — wpływ każdego czynnika jest rozpatrywany oddzielnie,
2. wieloczynnikowe — wpływ różnych czynników jest rozpatrywany łącznie.

Wprowadzenie. II

Według kryterium podział modeli przebiega następująco:

1. model efektów stałych — obserwacje są z góry podzielone na kategorie,
2. model efektów losowych — kategorie mają charakter losowy,
3. model mieszany — część kategorii jest ustalona, a część losowa.

Założenia analizy wariancji (jednoczynnikowej):

1. każda populacja musi mieć rozkład normalny,
2. pobrane do analizy próby są niezależne,
3. próby pobrane z każdej populacji muszą być losowymi próbami prostymi,
4. wariancje w populacjach są równe,
5. zmienna zależna mierzona jest na skali co najmniej przedziałowej,

Wprowadzenie. III

Uwaga 2

1. *Często zakłada się, że analizowane grupy są równoliczne (niektóre źródła podają, że ich liczność nie powinna różnić się o więcej niż 10%).*
2. *Wyniki uzyskane metodą analizy wariancji mogą być uznane za prawdziwe, gdy spełnione powyższe założenia.*
3. *W przypadku, gdy założenia analizy wariancji nie są spełnione należy posługiwać się testem Kruskala-Wallisa*

Jednoczynnikowa analiza wariancji. I

Rozważmy r populacji (próbek) o rozkładzie normalnym, jednakowej wariancji σ^2 i wartości oczekiwanej μ_i, gdzie $i = 1, \ldots, r$. Z populacji tych losujemy niezależne próby o liczebnościach n_i tj. $Y_{i1}, \ldots, Y_{in_i}$, na których przeprowadzamy pomiary, otrzymując wartości y_{ij} dla $i \in \overline{1, r}$, $j \in \overline{1, n_i}$. Całkowita wielkość próby wynosi $n = n_1 + n_2 + \cdots + n_r$.

Uwaga 3

Jeżeli $n_1 = n_2 = \cdots = n_r$, mówimy o modelu zrównoważonym.

Mamy następujący układ hipotez

$$H_0: \quad \mu_1 = \mu_2 = \cdots = \mu_r \tag{35}$$

$$H_1: \quad \text{nie wszystkie } \mu_i \text{ są sobie równe } i \in \overline{1, r} \tag{36}$$

Jednoczynnikowa analiza wariancji. II

Niech $\overline{Y}$ oznacza średnią arytmetyczną ze wszystkich obserwacji ze wszystkich r prób tzn.

$$\overline{Y} := \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{n_i} Y_{ij},$$

a $\overline{Y_i}$ średnią arytmetyczną z i-tej próby $(i \in \overline{1, r})$

$$\overline{Y_i} := \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}.$$

Jednoczynnikowa analiza wariancji. III

Definicja 2

Sumą kwadratów odchyleń od wartości średnich (ang. Total Sum of Squares lub Sum of Squares Total)) lub zmiennością całkowitą nazywamy statystykę

$$TSS := \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y})^2.$$

Definicja 3

Zmiennością międzygrupową (ang. Sum of Squares due to Treatment) nazywamy statystykę

$$SST := \sum_{i=1}^{r} n_i (\overline{Y_i} - \overline{Y})^2.$$

Jednoczynnikowa analiza wariancji. IV

Definicja 4

Sumą kwadratów błędów (ang. Sum of Squares of Errors) lub zmiennością wewnątrz grupową nazywamy statystykę

$$SSE := \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y_i})^2.$$

Fakt 1

Zachodzi równość

$$TSS = SST + SSE.$$

Zadanie 1

Pokazać powyższy fakt.

Jednoczynnikowa analiza wariancji. V

Uwaga 4

1. *Statystyka TSS wykorzystuje n zmiennych i warunek* $\sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y})=0$, *a więc ma* $n-1$ *stopni swobody.*
2. *Statystyka SST wykorzystuje r zmiennych i warunek* $\sum_{i=1}^{r} n_i(\overline{Y_i}-\overline{Y})=0$, *a więc ma* $r-1$ *stopni swobody.*
3. *Statystyka SSE wykorzystuje n zmiennych i r warunków* $\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y_i})=0$ $(i\in\overline{1,r})$, *a więc ma* $n-r$ *stopni swobody.*

Jednoczynnikowa analiza wariancji. VI

Definicja 5

Średnią zmiennością międzygrupową (ang. Mean Sum of Squares due to Treatment) nazywamy statystykę

$$MST := \frac{SST}{r-1}.$$

Średnią sumą kwadratów błędów (ang. Mean Sum of Squares of Errors) lub średnią zmiennością wewnątrz grupową nazywamy statystykę

$$MSE := \frac{SSE}{n-r}.$$

Jednoczynnikowa analiza wariancji. VII

Statystyką testową służącą do weryfikacji hipotezy (35) przeciwko hipotezie (36) stosowana jest statystyka F postaci

$$F = \frac{MST}{MSE}.$$

Przy założeniu prawdziwości hipotezy zerowej statystyka F ma rozkład F-Snedecora z $r-1$ stopniami swobody w liczniku i $n-r$ stopniami swobody w mianowniku.

Testy post-hoc.

Uwaga 5

ANOVA pozwala jedynie odrzucić hipotezę zerową o równości średnich w grupach. Nie wskazuje jednak, które średnie znacząco różnią się między sobą.

Dla znalezienia takich grup stosuje się testy typu post-hoc.
Typy testów post-hoc:

1. test HSD Tukeya (HSD – Honestly Significant Difference),
2. test Studenta-Newmana-Keulsa,
3. test LSD Fishera (LSD – Least Significant Difference).

Jednoczynnikowa analiza wariancji, jako model liniowy (model 1). I

Jednoczynnikowa analiza wariancji jest to szczególny przypadek modelu liniowego. Zapiszmy w postaci

$$Y_{ij} = \mu_1 + \alpha_i + \epsilon_{ij},$$

gdzie $\alpha_i = \mu_i - \mu_1$ dla $i \in \overline{1, r}$ oraz $\epsilon_{ij} = Y_{ij} - \mu_i$.
Ponieważ spełnione są założenia analizy wariancji, więc $\epsilon_{ij} \sim \mathcal{N}(0, \sigma)$ są niezależnymi zmiennymi losowymi.
Wystarczy wprowadzić sztuczne (nieme ang. *dummy variables*) zmienne objaśniające $X_1, \ldots, X_r$. Przyjmiemy umownie, że dla obserwacji z i-tej próbki mamy $X_1 = 1$, $X_i = 1$, zaś wszystkie inne zmienne x-owe są zerami. Otrzymamy wtedy

$$Y_{ij} = \mu_1 + \alpha_i X_i + \epsilon_{ij},$$

Jednoczynnikowa analiza wariancji, jako model liniowy (model 1). II

Zauważmy, że w tym modelu μ_1 odgrywa rolę *wyrazu wolnego*. Można sobie wyobrazić, że średnią μ_1 traktujemy jako *poziom bazowy* zaś pozostałe parametry uznajemy za *odchylenia od poziomu bazowego*. Hipoteza

$$H_0 : \alpha_2 = \ldots = \alpha_r = 0$$

sprowadza się do stwierdzenia, że wszystkie próbki pochodzą z tego samego rozkładu. Alternatywa jest postaci

$$H_1 : \text{ nie jest prawdą, że } \alpha_2 = \ldots = \alpha_r = 0,$$

czyli nie wszystkie średnie μ_i są jednakowe.

Jednoczynnikowa analiza wariancji, jako model liniowy (model 2). I

Zapiszmy model w postaci

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

gdzie

$$\mu = \frac{\sum_{i=1}^{r} n_i \mu_i}{n}$$

oraz $\epsilon_{ij} \sim \mathcal{N}(0, \sigma)$ są niezależnymi zmiennymi losowymi.

Uwaga 6

W powyższym podejściu μ nazywane jest ogólnym efektem średnim, zaś μ_i $(i \in \overline{1, r})$ efektem i-tej grupy.

Jednoczynnikowa analiza wariancji, jako model liniowy (model 2). II

Hipoteza

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_r = 0$$

sprowadza się do stwierdzenia, że wszystkie próbki pochodzą z tego samego rozkładu. Alternatywa jest postaci

$$H_1 : \text{ nie jest prawdą, że } \alpha_1 = \alpha_2 = \ldots = \alpha_r = 0,$$

czyli nie wszystkie średnie μ_i są jednakowe.

Bibliografia

Założenia analizy regresji liniowej. URL: https://www.naukowiec.org/wiedza/statystyka/zalozenia-anal (term. wiz. 24. 04. 2020).