mgr inż. Paweł Tarasiuk

1 Podstawowe pojęcia

1.1 Średnia

Niech $X = (X_1, X_2, \ldots, X_n)$ będzie krotką n obserwacji (lub próbą statystyczną). Wówczas

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

oznacza *średnią* z X. Zapis z pionową linią na górze jest wygodny również do oznaczania średnich z innych wyrażeń, na przykład

$$\overline{X^2} = \frac{1}{n} \sum_{i=1}^{n} X_i^2.$$

Łatwo pokazać, że dla równolicznych krotek X, Y i liczb rzeczywistych a, b:

$$\overline{aX + bY} = a\overline{X} + b\overline{Y}.$$

1.2 Wariancja

Wariancję (bez uwzględniania, że X może być próbą – wtedy warto byłoby zastosować „estymator nieobciążony wariancji" lub „wariancję z próby") obliczamy jako:

$$\operatorname{Var} X = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2.$$

Dodatkowo można zauważyć, że:

$$\operatorname{Var} X = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 = \overline{(X - \overline{X})^2} = \overline{X^2 - 2\overline{X}X + \overline{X}^2} =$$
$$= \overline{X^2} - 2\overline{\overline{X}X} + \overline{\overline{X}^2} = \overline{X^2} - 2\overline{X}^2 + \overline{X}^2 = \overline{X^2} - \overline{X}^2.$$

Stąd szczególnie łatwo pokazać, że dla dowolnej liczby rzeczywistej a:

$$\operatorname{Var}(aX) = a^2 \operatorname{Var} X$$

oraz

$$\operatorname{Var}(X + a) = \operatorname{Var} X.$$

Z ostatniego spostrzeżenia wynika także, że wariancja ze stałej zawsze wynosi 0. Wariancję można traktować jako miarę zróżnicowania elementów w próbie – czyli miarę odchylenia od sytuacji, w której wszystkie elementy są jednakowe.

1.3 Odchylenie standardowe

Odchylenie standardowe definiujemy jako:

$$\operatorname{sd} X = \sqrt{\operatorname{Var} X}.$$

Podobnie jak w przypadku wariancji, dodanie stałej do wszystkich elementów krotki nie zmienia wyniku, a wynik dla krotki jednakowych liczb wynosi 0.

Odchylenie standardowe jest przydatną, dodatnio liniową miarą rozrzutu obserwacji, gdyż dla dowolnej liczby rzeczywistej a:

$$\operatorname{sd}(aX) = |a| \operatorname{sd} X.$$

2 Kowariancja, korelacja liniowa

2.1 Kowariancja

Jeśli $X = (X_1, X_2, \ldots, X_n)$ oraz $Y = (Y_1, Y_2, \ldots, Y_n)$, to *kowariancję* definiujemy jako:

$$\operatorname{Cov}(X, Y) = \overline{(X - \overline{X})(Y - \overline{Y})}.$$

Jest to miara „wspólnego" zróżnicowania X i Y od średniej. Jeśli X_i są powyżej średniej dla tych samych $i = 1, \ldots, n$, co Y_i, to kowariancja jest dodatnia. Jeśli jest dokładnie przeciwnie (gdy X_i są powyżej średniej, to Y_i są poniżej), otrzymywane są wartości ujemne. Wartości bliskie zeru sugerują brak występowania takiej prostej, liniowej zależności.

Ważne własności (gdzie X, Y, Z – równoliczne krotki obserwacji, a, b – dowolne stałe, rzeczywiste współczynniki):

$$\operatorname{Cov}(X, X) = \operatorname{Var} X,$$

$$\operatorname{Cov}(X, Y) = \operatorname{Cov}(Y, X),$$

$$\operatorname{Cov}(X + a, Y) = \operatorname{Cov}(Y, X),$$

$$\operatorname{Cov}(X, a) = 0,$$

$$\operatorname{Cov}(aX, bY) = ab \operatorname{Cov}(X, Y),$$

$$\operatorname{Cov}(X, Y + Z) = \operatorname{Cov}(X, Y) + \operatorname{Cov}(X, Z).$$

2.2 Współczynnik korelacji liniowej Pearsona

To, ile wynosi kowariancja, zależy od zależności między próbami, ale również od charakterystyki samych prób – pomnożenie wszystkich elementów jednej z prób przez stały współczynnik powoduje również, że pomnożony przez niego zostaje wynik kowariancji. Prostszym w interpretacji wynikiem badania zależności między dwiema próbami jest *współczynnik korelacji liniowej Pearsona*, czyli:

$$\mathrm{r}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\mathrm{sd}\, X \,\mathrm{sd}\, Y}.$$

Współczynnik ten można obliczać dla krotek X, Y, które mają odchylenia standardowe (lub – równoważnie – wariancje) różne od zera. Potrzeba i wystarcza zatem, aby nie były to stałe.

Dla tak obliczanego współczynnika otrzymujemy, że:

$$\mathrm{r}(X, X) = 1,$$

$$\mathrm{r}(X, -X) = -1,$$

$$\mathrm{r}(X, aX) = 1 \quad \text{gdy } a > 0,$$

$$\mathrm{r}(X, aX) = -1 \quad \text{gdy } a < 0,$$

$$-1 \leqslant \mathrm{r}(X, Y) \leqslant 1,$$

$$\mathrm{r}(X, Y) = \mathrm{r}(Y, X),$$

$$\mathrm{r}(X + a, Y) = \mathrm{r}(X, Y).$$

Taki współczynnik łatwo jest interpretować – wartość 1 oznacza niezaburzoną, idealną, rosnącą zależność liniową. Podobnie wartość -1 oznacza idealną, malejącą zależność liniową. Z otrzymanej wartości 0 wynikałoby, że kowariancja wynosi 0, czyli brak jest tego rodzaju zależności.

Nawet w przypadku „idealnych zależności liniowych" nie wskazaliśmy jeszcze odpowiedzi, jakie to są zależności (na przykład – jakim wzorem można je opisać). Problem ten można rozwiązać stosując metodę ogólniejszą, która służy do szukania „najlepszych zależności liniowych" – najprostsza wersja takiej metody zostanie opisana w kolejnej sekcji.

3 Prosta regresja liniowa jednej zmiennej

Niech X i Y będą równolicznymi krotkami obserwacji (po n obserwacji). Przedstawiona metoda będzie służyła wskazywaniu współczynników „najlepszej" funkcji afinicznej, która pozwala obliczać Y na podstawie X. Zatem jeśli oznaczymy:

$$Z = f(X) = aX + b,$$

to chcemy, aby Z było „najbardziej zbliżone" do Y. Wzajemne podobieństwo dwóch sekwencji liczb można definiować w różny sposób, jednak w proponowanej metodzie zostanie zastosowana *metoda najmniejszych kwadratów*, lub minimalizacja średniego błędu kwadratowego. Najprościej to zapisać jako:

$$\overline{(Z-Y)^2} \to \min,$$

lub

$$\operatorname{cost}(a,b) = \overline{(aX+b-Y)^2} \to \min.$$

Jeśli X jest stałą, to trudno rozwiązać problem w jakikolwiek sprytny lub przydatny sposób. Można wtedy co najwyżej przyjąć $a = 0$ oraz $b = \overline{Y}$ (łatwo pokazać, że takie rozwiązanie jest wówczas optymalne). Lub dla dowolnego a przyjąć $b = \overline{Y} - a\overline{X}$. Wtedy rozwiązanie nie jest nawet jednoznaczne.

Skupmy zatem uwagę na przypadku, gdy X stałą nie jest. Wtedy dążenie współczynnikiem a lub b do wartości ∞ lub $-\infty$ powoduje, że również koszt jest rozbieżny do nieskończoności. Funkcja $\operatorname{cost}(a,b)$ jest zatem koercytywną funkcją o dziedzinie $\mathbb{R}^2$. Intuicyjnie lub z wykorzystaniem analizy matematycznej można stwierdzić, że taka funkcja na pewno osiąga gdzieś swoje minimum globalne.

Z twierdzenia Fermata o zerowaniu się pochodnej wynika, że kandydatami na takie minimum będą wyłącznie te punkty, w których pochodne cząstkowe $\operatorname{cost}(a,b)$ po a oraz b zerują się, czyli:

$$\begin{cases} \frac{\partial \operatorname{cost}(a,b)}{\partial a} = 0, \\ \frac{\partial \operatorname{cost}(a,b)}{\partial b} = 0. \end{cases}$$

Zacznijmy od prostszego z przekształceń, które szybko da nam przydatne wnioski:

$$\begin{gathered} \frac{\partial \operatorname{cost}(a,b)}{\partial b} = \frac{\partial \overline{(aX+b-Y)^2}}{\partial b} = \frac{1}{n}\frac{\partial\left(\sum_{i=1}^{n}(aX_i+b-Y_i)^2\right)}{\partial b} = \\ = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\left((aX_i+b-Y_i)^2\right)}{\partial b} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\left(b^2+2(aX_i-Y_i)b+(aX_i-Y_i)^2\right)}{\partial b} = \\ = \frac{1}{n}\sum_{i=1}^{n}\left(2b+2(aX_i-Y_i)\right) = 2\left(b+a\overline{X}-\overline{Y}\right). \end{gathered}$$

Skoro $\frac{\partial \operatorname{cost}(a,b)}{\partial b} = 0$, to:

$$2\left(b + a\overline{X} - \overline{Y}\right) = 0$$
$$b = \overline{Y} - a\overline{X}.$$

Badając pochodną cząstkową po a otrzymamy:

$$\frac{\partial \operatorname{cost}(a,b)}{\partial a} = \frac{\partial \overline{(aX+b-Y)^2}}{\partial a} = \frac{1}{n}\frac{\partial\left(\sum_{i=1}^{n}(aX_i+b-Y_i)^2\right)}{\partial a} =$$
$$= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\left((aX_i+b-Y_i)^2\right)}{\partial a} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\left(a^2X_i^2+2aX_i(b-Y_i)+(b-Y_i)^2\right)}{\partial a} =$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left(2aX_i^2 + 2X_i(b-Y_i)\right)$$

Po podstawieniu $b = \overline{Y} - a\overline{X}$:

$$\frac{\partial \operatorname{cost}(a,b)}{\partial a} = \frac{1}{n}\sum_{i=1}^{n}\left(2aX_i^2 + 2X_i(b-Y_i)\right) =$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left(2aX_i^2 + 2X_i(\overline{Y} - a\overline{X} - Y_i)\right) = \overline{2aX^2 + 2X(\overline{Y} - a\overline{X} - Y)} =$$
$$= \overline{2aX^2 + 2X\overline{Y} - 2aX\overline{X} - 2XY} = 2a\left(\overline{X^2} - X\overline{X}\right) - 2\left(\overline{X(Y-\overline{Y})}\right) =$$
$$= 2a\left(\overline{X^2} - \overline{X}^2\right) - 2\left(\overline{X(Y-\overline{Y})}\right) = 2a\operatorname{Var}X - 2\operatorname{Cov}(X-\overline{X}, Y) =$$
$$= 2a\operatorname{Var}X - 2\operatorname{Cov}(X,Y).$$

Skoro $\frac{\partial \operatorname{cost}(a,b)}{\partial a} = 0$, to:

$$2a\operatorname{Var}X - 2\operatorname{Cov}(X,Y) = 0$$
$$a = \frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}X}.$$

Współczynnik b możemy natychmiast obliczyć ze wskazanej wcześniej zależności:

$$b = \overline{Y} - a\overline{X}.$$

Warto zauważyć, że jedyne założenie tutaj dotyczyło $\operatorname{Var}X \neq 0$. Gdy jest ono spełnione, to wskazany wynik stanowi jedyny punkt zerowania się pochodnych cząstkowych. Jest to zatem globalne minimum kwadratowej funkcji kosztu.

4 Ocena przydatności wyniku regresji

Jeśli $Z = f(X)$ jest sekwencją proponowanych wyników (przy oznaczeniach jak w poprzedniej sekcji), to krotkę błędów można oznaczyć jako:

$$Err = Y - Z.$$

Dla optymalnego modelu na pewno $\operatorname{Var} Err \leqslant \operatorname{Var} Y$ – równość występowałaby dla modelu stałego $Z = 0$. Iloraz

$$FUV = \frac{\operatorname{Var} Err}{\operatorname{Var} E}$$

określa się zatem jako „współczynnik niewyjaśnionej wariancji". Jest to liczba z przedziału $[0; 1]$ – im mniejsza, tym lepiej.

Jeszcze popularniejszym w literaturze jest „współczynnik determinacji" lub po prostu „R kwadrat", czyli

$$R^2 = 1 - FUV.$$

W przypadku R^2 interpretacja jest nieco podobna, co w przypadku współczynnika korelacji – 0 oznacza zupełny brak zależności, która pasowałaby do modelu, a 1 oznacza zależność bezbłędną (bez uwzględniania takich szczegółów jak monotoniczność zależności liniowej). Wartość tę często podaje się procentowo, jako „odsetek wyjaśnionej wariancji". W zależności od dziedziny zastosowań, wyboru modelu i wyboru źródła bibliograficznego, różne wartości mogą być interpretowane w różny sposób. Wartości poniżej 70% sugerują, że zależność występuje, ale model może nie być najlepszym z możliwych dla wybranych danych. Wartości powyżej 85% często są oceniane jako sukces w dopasowywaniu modelu do danych z pewnym szumem.

Satysfakcjonująca wartość R^2 może zostać wskazana jako warunek konieczny uznania regresji za udaną, ale nie powinien to być warunek wystarczający. W tym celu warto zbadać rozkład wartości Err. W przypadku regresji liniowej zawsze $\overline{Err} = 0$, ale to jeszcze nawet nie znaczy, że rozkład jest symetryczny. Jeśli błędy nie przystają do rozkładu normalnego, to trudno je nazwać przypadkowym szumem, a wskazane jest poszukiwanie modelu, który wyjaśniłby dane lepiej. Ocena, czy dane „przystają do rozkładu normalnego" może obejmować narysowanie histogramu i ocenę „na oko", ale powinna w sobie zawierać również formalne podejście takie jak test Shapiro-Wilka albo test zgodności χ^2 Pearsona.

5 Regresja liniowa wielokolumnowa

Przyjmijmy, że mamy wiele objaśniających kolumn X_1, X_2, ..., X_m, które dotyczą n opisanych w wierszach obserwacji. Wówczas model liniowy wyjaśniający zmienną Y polega na wskazaniu takiego

$$Z = f(X_1, X_2, \ldots) = A_0 + A_1 X_1 + A_2 X_2 + \ldots$$

gdzie A to wektor o odpowiedniej długości, dla którego minimalizowany jest średni błąd kwadratowy, czyli

$$cost(A) = \overline{(Z - Y)^2} \to \min.$$

Niech X będzie macierzą podsumowującą wszystkie „kolumny z danymi", które są brane pod uwagę we wzorze na Z. Jeśli chcemy, żeby w tym wzorze występował składnik A_0 (nie jest to wymagane), potrzebna będzie odpowiednia kolumna, z której brane są liczby mnożone przez A_0 – czyli kolumna samych jedynek (oznaczymy ją również jako X_0). Taka macierz X ma wymiary $n \times (m+1)$.

$$X = \begin{bmatrix} 1 & (X_1)_1 & (X_2)_1 & \ldots & (X_m)_1 \\ 1 & (X_1)_2 & (X_2)_2 & \ldots & (X_m)_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & (X_1)_n & (X_2)_n & \ldots & (X_m)_n \end{bmatrix}.$$

Przyjmując, że A, Z i Y to macierze kolumnowe, w szczególności:

$$A = \begin{bmatrix} A_0 \\ A_1 \\ A_2 \\ \vdots \\ A_m \end{bmatrix},$$

otrzymamy, że

$$Z = XA$$

jest macierzą $n \times 1$, czyli da się porównywać z kolumną Y.

Zauważmy, że:

$$cost(A) = \overline{(XA - Y)^2} = \tfrac{1}{n}(XA - Y)^T(XA - Y) =$$
$$= \tfrac{1}{n}(A^T X^T - Y^T)(XA - Y) = \tfrac{1}{n}(A^T X^T XA - Y^T XA - A^T X^T Y + Y^T Y) =$$
$$= \tfrac{1}{n}(Y^T Y - 2A^T X^T Y + A^T X^T XA).$$

Obliczając gradient po A (wektor pochodnych cząstkowych po wszystkich elementach) otrzymujemy

$$\begin{gathered}\nabla_A(cost(A)) = \\ = \tfrac{1}{n}(\nabla_A(Y^TY) - 2\nabla_A(A^TX^TY) + \nabla_A(A^TX^TXA)) = \\ = \tfrac{2}{n}\left(-X^TY + X^TXA\right).\end{gathered}$$

Gradient ten się wyzeruje, gdy:

$$\begin{gathered}X^TXA - X^TY = 0 \\ X^TXA = X^TY \\ (X^TX)^{-1}(X^TX)A = (X^TX)^{-1}X^TY \\ A = (X^TX)^{-1}X^TY.\end{gathered}$$

Pod warunkiem, że macierz X^TX jest odwracalna. Na pewno jest to macierz kwadratowa (taką macierzą nie jest raczej samo X ani równoważnie X^T, więc takiego skracania nie można było wykonać). Macierz X^TX na pewno nie będzie odwracalna, jeśli układ kolumn X jest liniowo zależny. Taka sytuacja wydarzy się między innymi wtedy, gdy pewne dwie kolumny X będą jednakowe (lub proporcjonalne) lub gdy X ma więcej kolumn, niż wierszy. Warto zatem sprawdzić odwracalność macierzy X^TX, a nie zakładać, że wynik zawsze uda się wskazać – istnienie jednoznacznego zestawu współczynników jest równoważne tej odwracalności.

Ważna uwaga: w przypadku modeli realizujących kombinacje liniowe, na przykład $f(X_1, X_2) = a + bX_1 + cX_2 + dsin(X_1X_2)$, problem można rozwiązać tę samą metodą, poprzez sztuczne stworzenie dodatkowych kolumn. Tutaj należy sztucznie wprowadzić $X_3 = sin(X_1X_2)$ i rozwiązać problem dla $g(X_1, X_2, X_3) = a + bX_1 + cX_2 + dX_3$. Możliwe problemy wynikające z braku niezależności tak skonstruowanych kolumn zostaną wykryte na etapie oceny przydatności regresji (współczynnik R^2, rozkład błędów).