STATYSTYKA OPISOWA

Dr Alina Gleska

Instytut Matematyki WE PP

28 września 2018

Dane dwuwymiarowe

Analiza współzależności – badanie związków między cechami X i Y, między którymi istnieje więź przyczynowo-skutkowa.

Jednostka statystyczna – para (X, Y).

Rodzaje analiz:

- korelacji – pozwala ustalić kształt, kierunek i siłę związku między badanymi cechami;
- regresji – pozwala na ilościowy opis powiązań między cechami mierzalnymi, dla których analiza korelacji wskazała istnienie współzależności.

Analiza regresji

Analiza regresji – ilościowy opis powiązań między cechami

Funkcja regresji – funkcja matematyczna charakteryzująca zależność między cechami (zmiennymi)

Rodzaje współzależności:

- funkcyjna (deterministyczna) – danej wartości jednej cechy odpowiada dokładnie jedna wartość drugiej cechy; zależność taką można zapisać jako $Y = f(X)$;
- statystyczna (stochastyczna) – danej wartości jednej cechy odpowiada kilka różnych wartości drugiej cechy; oznacza to, że danym wartościom X nie są jednoznacznie przypisane wartości zmiennej Y.

Zależność statystyczna

Na kształtowanie się zmiennej będącej efektem ma wpływ nie tylko zmienna przyjęta jako przyczyna, ale również czynniki zakłócające, nazywane składnikiem losowym.

Zależność statystyczną można zapisać wówczas jako

$$Y = f(X) + \varepsilon,$$

gdzie ε – składnik losowy.

Równanie regresji (model regresji) – równanie opisujące związek między cechami po uwzględnieniu obecności składnika losowego.

Regresja

Rodzaje regresji:

$Y = f(X) + \varepsilon$ – regresja prosta;

$Y = f(X_1, \ldots, X_n) + \varepsilon$ – regresja wieloraka;

X – przyczyna – zmienna (cecha) objaśniająca (niezależna);

Y – skutek – zmienna (cecha) objaśniana (zależna).

Rodzaje regresji prostej:

- liniowa – gdy najlepiej dopasowaną do danych empirycznych jest linia prosta;
- krzywoliniowa – gdy najlepiej dopasowaną do danych empirycznych jest pewna linia krzywa.

Wybór modelu regresji

Wybór modelu – na podstawie wizualnej oceny rozrzutu punktów empirycznych na diagramie korelacyjnym

UWAGA! Zaletą wyboru liniowej funkcji regresji jest jasna interpretacja jej parametrów. Na podstawie funkcji liniowej można nie tylko przewidzieć wartość zmiennej zależnej Y dla danej wartości zmiennej niezależnej X, ale również w łatwy sposób ocenić efekty decyzji związanych ze zmianą wartości X (zmiany cechy objaśnianej są proporcjonalne do zmian cechy objaśniającej).

Liniowy model regresji

$$Y = a + bX + \varepsilon,$$

X – zmienna objaśniająca (niezależna),
Y – zmienna objaśniana (zależna),
a – wyraz wolny regresji,
b – współczynnik regresji,
ε – składnik losowy.

Współczynnik regresji – informuje, o ile jednostek przeciętnie zmieni się (wzrośnie lub zmaleje) wartość zmiennej Y, jeśli wartość zmiennej X wzrośnie o jednostkę.

UWAGA! $b = 0$ – brak zależności między cechami

Liniowy model regresji

LP	Wartości X	Wartości Y
1	x_1	y_1
2	x_2	y_2
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

$$y_i = a + bx_i + \varepsilon_i,$$

x_i, y_i - zaobserwowane (rzeczywiste) wartości cechy X i Y
ε_i - tzw. reszta modelu

Reszty modelu regresji

Reszty modelu – odchylenia wartości zaobserwowanych y_i od wartości teoretycznych $\hat{y}_i$, czyli tych, które wynikają z równania regresji (wartości, które wystąpiłyby, gdyby nie istniał wpływ losowości)

$$e_i = y_i - \hat{y}_i$$

$$\sum_{i=1}^{n} e_i = 0$$

Równanie prostej regresji:

$$\hat{y}_i = a + bx_i$$

Równanie prostej regresji

$$\hat{y}_i = a + bx_i$$

$$b = \frac{cov(X, Y)}{S_x^2}, \qquad a = \overline{y} - b \cdot \overline{x}$$

Ocena dopasowania

Miary dopasowania:

- wariancja reszt: $s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- odchylenie standardowe reszt: $s_e = \sqrt{s_e^2}$
- współczynnik zmienności losowej: $V_e = \frac{s_e}{\overline{y}} \cdot 100\%$
- współczynnik determinacji: $R^2 = r^2 \cdot 100\%$
  (gdzie r - współczynnik regresji liniowej Pearsona)
- współczynnik indeterminacji: $\varphi^2 = 100\% - R^2$

Współczynnik zmienności losowej V_e

Współczynnik zmienności losowej V_e informuje, jaką część średniego poziomu cechy Y stanowi przeciętna reszta (o ile procent przeciętnie mylimy się określając wartość cechy Y za pomocą wyznaczonego równania regresji).

Własności:

- przyjmuje wartości z przedziału $[0\%, 100\%]$;
- im mniejsza jest jego wartość, tym lepsze jest dopasowanie funkcji regresji do danych empirycznych.

Współczynnik determinacji R^2

Współczynnik determinacji R^2 służy do oceny funkcji regresji pod względem stopnia wyjaśniania przez model zróżnicowania cechy Y.

Całkowite zróżnicowanie – rozrzut obserwowanych wartości cechy Y względem średniego poziomu tej cechy, tzn. rozrzut $y_i - \overline{y}$:

- zróżnicowanie wyjaśnione przez model – wynika ze zmian w wartościach cechy X; rozrzut $\hat{y}_i - \overline{y}$
- zróżnicowanie niewyjaśnione przez model – wynika z obecności składnika losowego; rozrzut reszt $y_i - \hat{y}_i$

Współczynnik determinacji R^2

Odchylenia:

$$y_i - \overline{y} = (\hat{y}_i - \overline{y}) + (y_i - \hat{y}_i)$$

Sumy kwadratów:

$$\begin{array}{ccccc} \sum_{i=1}^{n}(y_i - \overline{y})^2 & = & \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 & + & \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ SST & = & SSR & + & SSE \end{array}$$

Współczynnik determinacji R^2

SST – całkowita suma kwadratów – mierzy zróżnicowanie zaobserwowanych wartości cechy objaśnianej Y względem ich średniej;

SSR – suma kwadratów dla regresji – mierzy zróżnicowanie zaobserwowanych wartości cechy objaśnianej Y ze względu na zmienność cechy X;

SSE – suma kwadratów dla błędu – mierzy zróżnicowanie wartości cechy Y wynikające z wpływu innych czynników niż cecha X.

Współczynnik determinacji:

$$R^2 = \frac{SSR}{SST} = r^2.$$

Współczynnik determinacji R^2

Własności:

- przyjmuje wartości z przedziału $[0\%, 100\%]$;
- im jego wartość jest bliższa 100%, tym lepsze jest dopasowanie funkcji regresji do danych empirycznych, przy czym:
  - $R^2 = 100\%$ – cała zmienność cechy Y została wyjaśniona zmiennością cechy X;
  - $R^2 = 0\%$ – cała zmienność cechy Y jest spowodowana tylko składnikiem losowym.

Współczynnik indeterminacji φ^2

Współczynnik indeterminacji (zbieżności) – określa, jaka część całkowitej zmienności cechy Y nie została wyjaśniona wpływem zmienności cechy X, czyli wynika z innych przyczyn:

$$\varphi^2 = \frac{SSE}{SST} = 100\% - R^2.$$

Własności:

- przyjmuje wartości z przedziału $[0\%, 100\%]$;
- im jego wartość jest bliższa 0%, tym lepsze jest dopasowanie funkcji regresji do danych empirycznych, przy czym:
  - $\varphi^2 = 0\%$ – dopasowanie jest "doskonałe";
  - $\varphi^2 = 100\%$ – cała zmienność cechy Y jest spowodowana składnikiem losowym.