Statystyka - wykład II

Pomoce do samodzielnej pracy

REGRESJA LINIOWA

Prosta ($y = ax + b$) najlepiej charakteryzująca zależność między x i y.

y – zmienna objaśniana (czasem nazywana też zależną)

x – zmienna objaśniająca (czasem nazywana też niezależną)

a – współczynnik kierunkowy, współczynnik regresji y względem x

b – wyraz wolny

REGRESJA LINIOWA

Metoda najmniejszych kwadratów ma na celu dopasowanie do zebranych danych, pary wyników takiej linii prostej, która jest do nich najlepiej dopasowana (obliczeniowo).

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b = \bar{y} - a\bar{x}$$

gdzie: $\bar{x}$ oznacza wartość średnią zmiennej objaśniającej
$\bar{y}$ oznacza wartość średnią zmiennej objaśnianej

RESZTY

Kiedy mamy wyliczone parametry funkcji regresji liniowej, możemy wyliczyć o ile nasze doświadczalne y różnią się od teoretycznego y obliczonego na podstawie funkcji regresji dla konkretnych wartości x. Każda taka różnica jest nazywana resztą albo składnikiem resztowym.

Wykres reszt pokazuje, na ile reszty odchylają się od linii regresji. Na wykresie reszt otrzymanych z modelu względem zmiennej objaśniającej nie powinno być widać żadnej zależności. Jeśli taka zależność jest widoczna, to regresja liniowa nie jest optymalną metodą szacowania y na podstawie x.

OCENA OSZACOWANIA FUNKCJI REGRESJI

- Odchylenie standardowe składnika resztowego.
- Współczynnik determinacji.

ODCHYLENIE STANDARDOWE SKŁADNIKA RESZTOWEGO

To inaczej błąd standardowy szacunku. Informuje, o ile średnio wartości empiryczne zmiennej odchylają się od wartości teoretycznych obliczonych na podstawie oszacowanej funkcji regresji:

$$S_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

gdzie: $\hat{y}$ oznacza estymowaną wartość zmiennej objaśnianej

WSPÓŁCZYNNIK DETERMINACJI

- Pozwala stwierdzić czy oszacowane równanie regresji jest przydatne do przewidywania czy nie.
- Nazywa się go często współczynnikiem dopasowania regresji i oznacza jako R^2.
- Określa on stopień w jakim linia regresji najmniejszych kwadratów wyjaśnia zmienność obserwowanych danych.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

OBSERWACJE NIETYPOWE

Obserwacje nietypowe (odstające) to obserwacje, które nie pasują do modelu, tzn. są o wiele za duże lub o wiele za małe. Często (choć nie zawsze) zmienne odstające biorą się z błędów pomiarowych lub pomyłek przy wprowadzaniu informacji do baz danych.

Obserwacja odstająca to obserwacja, która nie należy do poniższego przedziału:

$$x_{ods} \notin [Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)]$$

gdzie najczęściej przyjmuje się $k = 1{,}5$.

EKSTRAPOLACJA

Prognozowanie wartości pewnej zmiennej lub funkcji poza zakresem, dla którego mamy dane, przez dopasowanie do istniejących danych pewnej funkcji, następnie wyliczenie jej wartości w szukanym punkcie.

Pokrewną metodą jest interpolacja, gdzie po dopasowaniu funkcji wyliczamy jej wartość pomiędzy znanymi jej punktami.