Modele Liniowe
Powtórka do kolosa

3 grudnia 2018

1 Regresja liniowa

X_i - zmienna objaśniająca
Y_i - zmienna odpowiadająca

Model regresji:

$$Y_i = \beta_0 + \beta_1 X_i + \xi_i$$

β_0 - intercept, wyraz wolny β_1 - nachylenie, $\xi_i \sim N\left(0, \sigma^2\right)$ - błąd losowy
Dodatkowo:

$$E(Y_i|X_i) = \beta_0 + \beta_1 X_i$$
$$Var(Y_i|X_i) = \sigma^2$$

2 Estymacja

Dopasowany model regresji:

$$\hat{Y} = b_0 + b_1 X_i$$

z resztą:

$$e_i = Y_i - \hat{Y} = Y_i - (b_0 + b_1 X_i)$$

Minimalizując $\sum e_i^2$ metodą najmniejszych kwadratów otrzymujemy wzory na estymatory b_0 i b_1

$$b_1 = \frac{\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum\left(X_i - \bar{X}\right)^2}$$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

Estymator wariancji:

$$s^2 = \frac{\sum \left(Y_i - \hat{Y}_i\right)^2}{n-2} = \frac{\sum e_i^2}{n-2} = \frac{SSE}{dfE} = MSE$$

3 Przedziały ufności

3.1 Dla β_1

$$b_1 \sim N(\beta_1, \sigma^2(b_1))$$

, gdzie

$$\sigma^2(b_1) = \frac{\sigma^2}{\sum \left(X_i - \bar{X}\right)^2}$$

Statystyka testowa:

$$t = \frac{(b_1 - \beta_1)}{s(b_1)} \sim t(n-2)$$

, gdzie

$$s^2(b_1) = \frac{s^2}{\sum \left(X_i - \bar{X}\right)^2}$$

Przedział ufności:

$$b_1 \pm t_c s(b_1)$$

, gdzie

$$t_c = (1 - \alpha/2, n-2)$$

3.1.1 Testowanie

$$H_0 : \beta_1 = 0$$

,

$$H_a : \beta_1 \neq 0$$

odrzuć H_0, gdy $|t| \geqslant t_c$
$P(|z| \geqslant |t|)$ - p-wartość , $z \sim t(n-2)$
odrzuć H_0, gdy p-wartość przekracza poziom istotności

3.2 Dla β_0

$$b_1 \sim N(\beta_0, \sigma^2(b_0))$$

, gdzie

$$\sigma^2(b_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum \left(X_i - \bar{X}\right)^2}\right]$$

Statystyka testowa:

$$t = \frac{b_0 + \beta_0}{s(b_0)} \sim t(n-2)$$

$$s^2(b_0) = s^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum \left(X_i - \bar{X}\right)^2} \right]$$

Przedział ufności:

$$b_0 \pm t_c s(b_0)$$

$$t_c = (1 - \alpha/2, n-2)$$

3.2.1 Testowanie

$$H_0 : \beta_0 = \beta_{00}$$

,

$$H_a : \beta_1 \neq \beta_{00}$$

odrzuć H_0, gdy $|t| \geqslant t_c$

$$t = \frac{b_0 + \beta_{00}}{s(b_0)}$$

$P(|z| \geqslant |t|)$ - p-wartość ,

$$z \sim t(n-2)$$

odrzuć H_0, gdy p-wartość przekracza poziom istotności

3.3 Inne

Normalność b_0 i b_1 wynika z tego, że b_0 i b_1 są kombinacją liniową Y_i(iid). Jeśli ξ_i nie są z rozkładu normalnego, ale zbliżonego to można stosować z powodzeniem powyższe testy i przedziały ufności.

4 Moc

Moc testu (moc statystyczna) to prawdopodobieństwo niepopełnienia błędu drugiego rodzaju – przyjęcia hipotezy zerowej, gdy w rzeczywistości jest ona fałszywa. Im większe jest to prawdopodobieństwo, tym lepszy jest dany test jako narzędzie do różnicowania między hipotezą prawdziwą i fałszywą. Moc można wyrazić jako dopełnienie prawdopodobieństwa popełnienia błędu drugiego rodzaju (β), czyli $1 - \beta$.

$$H_0 : \beta_1 = 0$$

,

$$H_a : \beta_1 \neq 0$$

Statystyka testowa:

$$t = \frac{b_1}{s(b_1)} \sim t(n-2, \delta)$$

$$\delta = \frac{\beta_1}{\sigma(b_1)}$$

$b_1 \sim N(\beta_1, \sigma_2(b_1))$ - parametr niecentralności
Musimy znać σ, SSX i n, żeby obliczyć:

$$\sigma^2(b_1) = \frac{\sigma^2}{\sum \left(X_i - \bar{X}\right)^2}$$

5 Estymacja średniej

$$E(Y_h) = \mu_h = \beta_0 + \beta_1 X_h$$

$$\hat{\mu}_h = b_0 + b_1 X_h$$

$$\hat{\mu}_h \sim N(\mu_h, \sigma^2)$$

$$\sigma^2(\hat{\mu}_h) = \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2}\right]$$

$$s^2(\hat{\mu}_h) = s^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2}\right]$$

$$\frac{\hat{\mu}_h - E(Y_h)}{s(\hat{\mu}_h)} \sim t(n-2)$$

5.1 Przedział ufności

$$\hat{\mu}_h \pm t_c s(\hat{\mu}_h)$$

6 Prognoza

$$Y_h = \beta_0 + \beta_1 X_h + \xi_h$$

$$s^2(pred) = s^2 \left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2}\right]$$

$$\frac{Y_h - \hat{\mu}_h}{s(pred)} \sim t(n-2)$$

6.1 Przedział prognozy

$$\hat{\mu}_h \pm t_c s(pred)$$

7 ANOVA

7.1 Ogólnie - Total

$$SST = \sum (Y_i - \bar{Y})^2$$
$$dfT = n - 1$$
$$MST = \frac{SST}{dfT}$$

7.2 Model

$$SSM = \sum (\hat{Y}_i - \bar{Y})^2$$
$$dfM = 1$$
$$MSM = \frac{SSM}{dfM}$$

7.3 Błąd - Error

$$SSE = \sum (Y_i - \hat{Y}_i)^2$$
$$dfE = n - 2$$
$$MSE = \frac{SSE}{dfE}$$