1 Regresja liniowa cz. II

1.1 Miary dopasowania modelu do danych

Dekompozycja wariancji

Idea:

$$\underbrace{y-\overline{y}}_{\text{Odchylenie całkowite}} = \underbrace{y-\widehat{y}}_{\substack{\text{Odchylenie nie wyjaśnione}\\ \text{(błąd)}}} + \underbrace{\widehat{y}-\overline{y}}_{\substack{\text{Odchylenie wyjaśnione}\\ \text{(regresyjne)}}}$$

$$\underbrace{\sum_{i=1}^{n}(y_i-\overline{y})^2}_{\substack{\text{Całkowita suma kwadratów}\\ SST}} = \underbrace{\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}_{\substack{\text{Suma kwadratów błędów}\\ SSE}} + \underbrace{\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2}_{\substack{\text{Suma kwadratów odchyleń regresyjnych}\\ SSR}}$$

SSR jest tzw. zmienność wyjaśniona
SSE jest tzw.zmienność niewyjaśniona (wynika z niedoskonałości regresji)

Przykład 1 *RYSUNEK!!!*

Definicja 1 *Współczynnik determinacji r^2[w programach czasem R^2]*

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

Wniosek 1

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Wartość współczynnika determinacji wyraża zmienność Y, która została wyjaśniona przez zachodzenie liniowego związku między X i Y. Wartość współczynnika korelacji liniowej to miara liniowego związku.

Lemat 1

$$r^2 = \widehat{\rho}^2_{X,Y}$$

Własności 1

1. *Współczynnik determinacji r^2 jest miarą siły liniowego związku między zmiennymi.*
2. $r^2 \in [0,1]$
3. *Wartość współczynnika determinacji nic nam nie mówi o kierunku zależności między zmiennymi.*
4. *Większe wartości r^2 mogą sugerować dokładniejsze dopasowanie modelu regresji do zbioru obserwacji.*

5. $r^2 = 1$ *oznacza, że zmienność* Y *jest w* 100% *wyjaśniona przez model regresji*

6. $r^2 \leq 0.3$ *oznacza, że regresja wyjaśnia mniej niż 30% zmienności* Y*.*

Uwaga 1 *Sama wartość współczynnika determinacji nie pozwala ocenić jakości dopasowania linii regresji do zbioru obserwacji. Istnieją przykłady, dla których* r^2 *jest duże, a model jest nieodpowiedni.*

Lemat 2 1. $SST = SS_Y$

2. $SSR = \widehat{\beta}_1 SS_{XY}$

3. $SSE = SS_Y - \widehat{\beta}_1 SS_{XY}$

Ćwiczenie 1 *Wskaż przykład danych gdzie model regresji jest nieodpowiedni, a* r^2 *jest duże. Wsk. parabola.*

Definicja 2

$$\begin{aligned} MSR &= \frac{SSR}{1} \\ MSE &= \frac{SSE}{n-2} \end{aligned}$$

MSR to średnie kwadratowe odchylenie regresyjne (mean square regression).
MSE to średni błąd kwadratowy (mean square error)

Uwaga 2 MSE *mierzy/wyraża rozproszenie danych od linii regresji. Jest to miara bezwzględna - jej wielkość w zależności od sytuacji może mieć inne znaczenie.*

Uwaga 3 MSE *wcześniej oznaczano* $\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2 = \frac{SSE}{n-2}$ *- bo to również estymator wariancji reszt.*

Lemat 3 *Jeśli* $\beta_1 = 0$ *to*

$$F = \frac{MSR}{MSE} \sim F_{1,n-2}$$

Obserwacja 1 *Stosunek zmienności wyjaśnionej do zmienności niewyjaśnionej* $\frac{MSR}{MSE}$ *może być potraktowany jako kolejna miara dopasowania modelu do danych. Przy braku liniowego związku między zmiennymi wielkość* MSR *powinna być bliska zeru. Fakt ten można wykorzystać do przeprowadzenia testu statystycznego badającego zachodzenie liniowego związku.*

$$\begin{aligned} H0 &: \beta_1 = 0 \textit{(nie zachodzi związek między } X \textit{ i } Y\textit{)} \\ H1 &: \beta_1 \neq 0 \end{aligned}$$

$$\frac{MSR}{MSE} \sim F_{1,n-2}$$

Jest to test jednostronny (prawostronny).

Obserwacja 2

$$F = t^2$$

gdzie $t = \frac{\widehat{\beta}_1}{SE(\widehat{\beta}_1)}$.

Wniosek 2 *Wnioskowanie za pomocą testu F i testu t jest równoważne. F jest testem jednostronnym, a t dwustronnym. W obu testach otrzymamy to samo p-value, gdy dla t ustalimy symetryczne przedziały - zbiory krytyczne.*

Prognoza

Jednym z zastosowań modelu regresji jest prognozowanie dalszego poziomu zjawiska.

Prognoza punktowa:

Niech x_0 jest ustaloną wartością X. Odpowiadającą jej wartość (teoretyczną) zmiennej Y obliczymy ze wzoru

$$\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$$

Prawdziwa wartość wynosi

$$y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$$

Błąd prognozy

$$\widehat{y}_0 - y_0 = \widehat{\beta}_0 - \beta_0 + \left(\widehat{\beta}_1 - \beta_1\right) x_0 - \varepsilon_0$$

$$E\left(\widehat{y}_0 - y_0\right) = 0$$

$$Var\left(\widehat{y}_0 - y_0\right) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]$$

Wniosek 3 *Niepewność mierzona odchyleniem standardowym błędu prognozy rośnie wraz ze zwiększaniem się odległości* x_0 *od* $\overline{x}$*, oraz maleje wraz ze zwiększaniem się liczebności próbki.*

Obserwacja 3 *Jeśli* $x_0 = \overline{x}$ *to* $\widehat{y}_0 = \overline{y}$ *oraz*

$$Var\left(\widehat{y}_0 - y_0\right) = Var\left(\overline{y} - y_0\right) = \sigma^2 \left[1 + \frac{1}{n}\right] \underset{n\to\infty}{\to} \sigma^2$$

Uwaga 4 *W praktyce nie znamy* σ^2 *zastępujemy go estymatorem* $\widehat{\sigma}^2$. $\sqrt{\widehat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]}$ *nazywamy standardowym błędem predykcji.*

Prognoza przedziałowa:

Twierdzenie 1 $(1-\alpha)\,100\%$ *przedział predykcji zmiennej* Y *dla ustalonej wartości* x_0 *zmiennej objaśniającej jest określony wzorem*

$$\widehat{y}_0 \pm t_{n-2,\alpha/2}\sqrt{\widehat{\sigma}^2\left[1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{SS_X}\right]}$$

gdzie $P\left(t > t_{n-2,\alpha/2}\right) = \alpha/2$, $t \sim t_{n-2}$.

Prognoza przeciętnego poziomu zmiennej Y

Motywacja:

y_0 - prognoza przychodów ze sprzedaży

$E(y_0)$ - prognoza przeciętnych przychodów ze sprzedaży

Mając ustaloną wartość x_0 jesteśmy zainteresowani prognozą $E(y_0)$ a nie y_0. Zauważmy, że $E(y_0) = \beta_0 + \beta_1 x_0$, wielkość tę przybliżamy wartością $\widehat{E}(y_0) = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$. **Wartość prognozy dla** $E(y_0)$ **i** y_0 **są takie same** ($\widehat{E(y_0)} = \widehat{y_0}$) **!** Zobaczmy czy zmieni się standardowy bład predykcji:

$$E(y_0) - \widehat{E}(y_0) = \left(\beta_0 - \widehat{\beta}_0\right) + \left(\beta_1 - \widehat{\beta}_1\right) x_0$$

$$Var\left(E(y_0) - \widehat{E}(y_0)\right) = \sigma^2\left[\frac{1}{n}+\frac{(x_0-\overline{x})^2}{S_{xx}}\right]$$

Twierdzenie 2 $(1-\alpha)\,100\%$ *przedział predykcji średniego poziomu zmiennej* Y *dla ustalonej wartości* x_0 *zmiennej objaśniającej jest określony wzorem*

$$\widehat{E}(y_0) \pm t_{n-2,\alpha/2}\sqrt{\widehat{\sigma}^2\left[\frac{1}{n}+\frac{(x_0-\overline{x})^2}{SS_X}\right]}$$

gdzie $P\left(t > t_{n-2,\alpha/2}\right) = \alpha/2$, $t \sim t_{n-2}$.

Uwaga 5 *Zgodnie z oczekiwaniami przedział ufności dla przeciętnego poziomu jest węższy niż dla samego poziomu.*