Efekty krańcowe w modelu regresji, gdy zmienna objaśniająca jest mierzona na skali nominalnej i wyraża więcej niż dwa stany (tzw. zmienna kategorialna). Konstrukcja pomocniczych zmiennych zero-jedynkowych odpowiadających ww. zmiennej kategorialnej.

Przykładowe zagadnienia, gdy zmienne objaśniające wyrażają

1) podokresy (dni) w tygodniu,
2) stan cywilny,
3) miesiące, gdy zbiór danych zawiera informacje pochodzące z okresów miesięcznych.

Ad 1. Przedmiotem zainteresowania jest przeciętna dzienna wielkości sprzedaży w pewnej sieci sklepów. Obserwuje się silne zróżnicowanie sprzedaży w zależności m.in. od dni tygodnia. Rozważamy trzy okresy w tygodniu, więc liczba kategorii wynosi trzy, tj.:

a) wtorek-piątek (wt.-pt.),
b) sobota-niedziela (sb.-nd.)[1],
c) poniedziałek (pon.).

Na potrzeby powyższego zagadnienia przyjmiemy model regresji o następującej strukturze (z wyrazem wolnym β_0, a dla uproszczenia pominięto inne zmienne objaśniające[2]):

$$y_t = \beta_0 + \beta_1 \cdot x_{t1} + \beta_2 \cdot x_{t2} + \varepsilon_t, \quad \text{gdzie } E(\varepsilon_t) = 0 \text{ dla } t=1,\dots,T \tag{1}$$

Kodowanie zmiennych wyrażających dzień tygodnia opiera się na zmiennych zero-jedynkowych (binarnych) i w tym przypadku może być następujące:

a) $x_{t1}=x_{t2}=0$, gdy sprzedaż miała miejsce w okresie wtorek-piątek, jest to tzw. kategoria referencyjna,
b) $x_{t1}=1$ i $x_{t2}=0$, gdy sprzedaż dotyczy dni sobota i niedziela,
c) $x_{t1}=0$ i $x_{t2}=1$, gdy sprzedaż miała miejsce w poniedziałek.

Oczywiście sposób kodowania może być inny, ale ten jest wygodny z punktu widzenia interpretacji. Obowiązuje reguła, że dla wyrażenia J kategorii stosuje się J–1 pomocniczych zmiennych binarnych.

[1] Przykład ten został skonstruowany przed obowiązującym w Polsce, od 2018 r., ograniczeniem handlu w niedziele (zob. Ustawa z dnia 10 stycznia 2018 r. o ograniczeniu handlu w niedziele i święta oraz w niektóre inne dni, Dz.U. z 2021 r. poz. 936).

[2] Z formalnego punktu wyraz wolny (odpowiadający zmiennej równej jeden dla każdego t=1,…,T) też można „ukryć „poprzez odjęcie od y_t i zmiennych objaśniających ich średnich wartości w próbie dla każdej z nich, np. y_t–średnia(y) itd. „Ukrycie” wyrazu wolnego w równaniu (1) nie zmienia postaci wzorów (3) i (4) i interpretacje efektów, a jedynie zmienia się interpretacja zmiennej y_t w kontekście wzoru (2), w którym wówczas „znika” β_0.

Efekty krańcowe liczymy jako różnice między wartościami oczekiwanymi zmiennej y_t dla różnych wartości zmiennych x_{t1} i x_{t2}. **Wyrażają one wpływ określonego okresu w tygodniu na zmianę wielkości sprzedaży względem okresu referencyjnego.**

Wartości oczekiwane zmiennej y_t w zależności od okresu sprzedaży są następujące (na podstawie wzoru 1):

$$\begin{aligned} &E(y_t|x_{t1}=x_{t2}=0)=\beta_0+\beta_1\cdot 0+\beta_2\cdot 0=\beta_0,\\ &E(y_t|x_{t1}=1, x_{t2}=0)=\beta_0+\beta_1,\\ &E(y_t|x_{t1}=0, x_{t2}=1)=\beta_0+\beta_2. \end{aligned} \tag{2}$$

Efekty krańcowe (η) są dane wzorami:

$$\begin{aligned} &\eta(\text{"sb} - \text{nd" wzgl."wt} - \text{pt"})=E(y_t|x_{t1}=1, x_{t2}=0)-E(y_t|x_{t1}=x_{t2}=0)=\beta_1,\\ &\eta(\text{"pon" wzgl."wt} - \text{pt"})=E(y_t|x_{t1}=0, x_{t2}=1)-E(y_t|x_{t1}=x_{t2}=0)=\beta_2. \end{aligned} \tag{3}$$

Zatem tutaj oba parametry (β_1, β_2) mają bezpośrednią interpretację jako efekty krańcowe.

Można dodatkowo policzyć efekt krańcowy wynikający ze zróżnicowania sprzedaży w dniach sobota i niedziela względem poniedziałku:

$$\eta(\text{"sb} - \text{nd" wzgl."pon"})=E(y_t|x_{t1}=1, x_{t2}=0)-E(y_t|x_{t1}=0, x_{t2}=1)=\beta_1-\beta_2. \tag{4}$$

Interpretacja. Niech zgodnie z przewidywaniami $\beta_1>0$ i $\beta_2<0$. Wówczas $\beta_1-\beta_2>0$. Średnia dzienna sprzedaż w okresie sobota-niedziela jest średnio wyższa o β_1 jednostek w stosunku do okresu wtorek-piątek. W poniedziałek sprzedaż ta jest średnio o β_2 jednostek niższa niż średnia w okresie od wtorku do piątku. Natomiast przeciętna sprzedaż w sobotę i niedzielę jest o ($\beta_1-\beta_2$) jednostek wyższa niż poniedziałek.

Zauważmy także ile wynosi pomiar zależność przeciwnej:

$$\eta(\text{"wt} - \text{pt" wzgl."pon"})=-\eta(\text{"pon" wzgl."wt - pt"})=-\beta_2. \tag{5}$$

W badaniach empirycznych można zastosować także taką definicję zmiennych zero-jedynkowych (x_{t1}, x_{t2}), aby oceny parametrów β_1 i β_2 zawsze były dodatnie. W omawianym przykładzie ma to miejsce, gdy za kategorię referencyjną, $x_{t1}=0$ i $x_{t2}=0$, przyjmie się sprzedaż w poniedziałek, czyli kategorię, w przypadku której obserwuje się najniższą dzienną sprzedaż.

Ad 2. Przykład dotyczy wprowadzenia do modelu zmiennej wyrażającej stan cywilny.

Posiadamy surowe dane, np. niech: x=1 – niezamężna (kawaler/panna), x=2 – osoba w związku cywilnym (żonaty/zamężna), x=3 – osoba, której prawne małżeństwo przestało istnieć z powodu śmierci współmałżonka (wdowiec/wdowa), x=4 – osoba, której małżeństwo zostało rozwiązane orzeczeniem sądu (rozwiedziony/rozwiedziona). Zauważmy, że do modelu nie wprowadza się x-sa, który przyjmuje wartości 1, 2, 3 i 4. Natomiast przyjmuje się, że jeden ze stanów jest referencyjny i następnie konstruuje się trzy zmienne zero-jedynkowe.

Przykładowe kodowanie zmiennych objaśniających, gdy kategorią referencyjną jest stan cywilny *rozwiedziony* lub *rozwiedziona* (pominięto indeks obserwacji *t*):

Stan cywilny/zmienna	d_1	d_2	d_3	$\cancel{d_4}$
kawaler/panna	**1**	0	0	~~0~~
żonaty/zamężna	0	**1**	0	~~0~~
wdowiec/wdowa	0	0	**1**	~~0~~
rozwiedziony lub *rozwiedziona*	**0**	**0**	**0**	~~1~~

Wówczas w modelu pojawią się trzy (a nie cztery) sztuczne zmienne zero-jedynkowych ($d_{t,j}$):

$$y_t = \beta_0 + \beta_1 \cdot d_{t,1} + \beta_2 \cdot d_{t,2} + \beta_3 \cdot d_{t,3} + \varepsilon_t, \qquad \text{dla } t=1,\dots,T \tag{6}$$

Efekty krańcowe są określone jak wcześniej. Jakie są konsekwencje wprowadzenia w modelu (6) dodatkowej zmiennej sztucznej d_4, która określona jak w powyższej tabeli?

- Gdy w (6) są cztery zmienne d_1, d_2, d_3 i d_4, to zachodzi dokładna współliniowość (liniowa zależność) między (sumą) $d_1 + d_2 + d_3 + d_4$ a kolumną jedynek (przy β_0).
- Wówczas estymator MNK nie istnieje dla tych pięciu parametrów (w tym β_4 przy d_4). W przypadku modeli nieliniowych (np. logitowy lub probitowy) uzyskuje się „nietypowe wyniki” dla parametrów przy tych czterech zmiennych i wyrazu wolnego oraz bardzo duże (i prawie identyczne co do wartości) średnie błędy estymacji.

Zauważmy, że w obu przykładach modeli regresji zaprezentowanych powyżej występuje wyraz wolny β_0 i towarzysząca mu zmienna przyjmująca wartość 1 dla każdej obserwacji. Istnieje równoważna postać modelu bez wyrazu wolnego, ale ze wszystkimi pomocniczymi zmiennymi zero-jedynkowych:

$$y_t = \alpha_1 \cdot d_{t,1} + \alpha_2 \cdot d_{t,2} + \alpha_3 \cdot d_{t,3} + \alpha_4 \cdot d_{t,4} + \varepsilon_t, \tag{7}$$

W powyższym przypadku pojedyncze parametry α nie są bezpośrednio interpretowalne, a efekty krańcowe mają inną postać, gdyż są wyrażone przez różnice tych parametrów. Przykładowo efekty

$$\begin{aligned}
&\eta(\text{"kawaler" wzgl."rozwiedziony"}) = \\
&\quad E(y_t | d_{t1} = 1, d_{t2} = d_{t3} = d_{t4} = 0) - E(y_t | d_{t1} = d_{t2} = d_{t3} = 0, d_{t4} = 1) = \alpha_1 - \alpha_4 \\
&\eta(\text{"zonaty" wzgl."rozwiedziony"}) = \alpha_2 - \alpha_4 \\
&\eta(\text{"rozwiedziony" wzgl."wdowiec"}) = \alpha_4 - \alpha_3 \quad \text{itd.}
\end{aligned} \tag{8}$$

Ad 3. Trzeci przykład dotyczy modelu sformułowanego dla danych w formie szeregu czasowego dla okresów miesięcznych (np. sprzedaż detaliczna napojów, lodów, czekolady, opon samochodowych), gdy warto uwzględnić wpływ sezonowości.

Jeśli decyzje zakupowe konsumentów mogą zależeć m.in. od okresu, to w modelu pojawia się 11 zmiennych zero-jedynkowych ($d_{t,j}$):

$$y_t = \beta_0 + \beta_1 d_{t,1} + \beta_2 d_{t,2} + .. + \beta_{11} d_{t,11} + \varepsilon_t, \qquad \text{dla } t=1,\dots,T \tag{9}$$

Pozostaje ustalić znaczenie tych sztucznych zmiennych w celu identyfikacji miesięcy. W jednym z wariantów, gdy kategorią referencyjną jest 12-sty miesiąc, czyli grudzień, dla poszczególnych zmiennych $d_{t,j}$ można przyjąć wartości 0 lub 1 w poniższy sposób (pominięto indeks t):

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}	d_{11}
styczeń	**1**	0	0	0	0	0	0	0	0	0	0
luty	0	**1**	0	0	0	0	0	0	0	0	0
marzec	0	0	**1**	0	0	0	0	0	0	0	0
kwiecień	0	0	0	**1**	0	0	0	0	0	0	0
maj	0	0	0	0	**1**	0	0	0	0	0	0
czerwiec	0	0	0	0	0	**1**	0	0	0	0	0
lipiec	0	0	0	0	0	0	**1**	0	0	0	0
sierpień	0	0	0	0	0	0	0	**1**	0	0	0
wrzesień	0	0	0	0	0	0	0	0	**1**	0	0
październik	0	0	0	0	0	0	0	0	0	**1**	0
listopad	0	0	0	0	0	0	0	0	0	0	**1**
grudzień	0	0	0	0	0	0	0	0	0	0	0

Parametry ($\beta_1, \beta_2, \dots$) mają bezpośrednią interpretację jako efekty krańcowe wyrażające zmianę sprzedaży w danym miesiącu względem miesiąca referencyjnego (grudnia).