1.6 Zmienne jakościowe i dyskretne w modelu regresji

1.6.1 Zmienne dyskretne i zero-jedynkowe (Dummy Variables)

W badaniach ekonometrycznych bardzo często występują zjawiska, które opisujemy zmiennymi nie posiadającymi charakteru ilościowego, a jakościowy (np. wykształcenie). Podobny charakter mają odpowiedzi na różne pytania ankietowe. Tego typu zmienne charakteryzuje się tym, że przyjmuje pewną skończoną liczbę dyskretnych wartości. Z powodu ograniczenia liczby ich wartości nie moża traktować tych zmiennych w sposób przyjęty dla zmiennych ciągłych w regresji. Dzieje się tak ponieważ pośrednie wartości tej zmiennej nie mają sensu ekonomicznego, a obliczone współczynniki modelu nie posiadają interpretacji ekonomicznej.

Przykład 1.

Budujemy model tłumaczący zarobki poziomem wykształcenia i liczbą godzin pracy w tygodniu.

$$zarobki = \beta_0 + \beta_1 liczba\ godzin\ pracy + \beta_2 wykształcenie + \varepsilon$$

*Zmienna liczba godzin pracy może być traktowana jako ciągła, ponieważ może przyjmować wszystkie wartości z przedziału [0,168]. Zmienna wykształcenie przyjmuje tylko trzy wartości: podstawowe, średnie i wyższe. Współczynnik przy zmiennej liczba godzin pracy informuje nas o tym o ile więcej zarobimy pracując godzinę dłużej a pozostałe zmienne pozostaną na swoim poziomie (*ceteris paribus*). Natomiast jak zinterpretować współczynnik przy zmiennej wykształcenie?*

Jest to zmiana dochodu spowodowana zmianą poziomu wykształcenia. Ale jeżeli w ten sposób zinterpretujemy tą zmienną to założymy, że różnica w zarobkach między osobami ze średnim wykształceniem a wykształceniem podstawowym jest taka sama jak różnica w zarobkach między osobą posiadającą wykształcenie wyższe a osobą ze średnim wykształceniem. Co więcej, założymy również, że różnica w zarobkach osoby z wyższym wykształceniem a osoby z wykształceniem podstawowym będzie równa dwukrotności różnicy w zarobkach między wykształceniem wyższym a średnim. Jednak takie założenia są nieuzasadnione teorią ekonomiczną!

Przykład 2.

Budujemy model tłumaczący zarobki liczbą godzin pracy w tygodniu i miejscem zamieszkania. Liczba godzin pracy jest zdefiniowana tak jak w przykładzie 1. Miejsce zamieszkania jest zmienną przyjmującą inną wartość dla

każdego województwa (16 wartości).

$$zarobki = \beta_0 + \beta_1 liczba\ godzin\ pracy + \beta_2 klm + \varepsilon$$

W takim przypadku, w odróżnieniu od poprzedniego przykładu współczynnik przy zmiennej miejsce zamieszkania nie będzie miał wogóle interpretacji!

W ekonometrii rozróżniane są dwa typy nieciągłych zmiennych. Zmienne jakościowe, to zmienne których wartości posiadają charakter opisowy, np kolor oczu (niebieski, zielony, brązowy). Podczas analizy badacz arbitralnie wybiera sposób kodowania wartości zmiennej. Drugim typem są zmienne dyskretne. Takie zmienne przyjmują z góry określoną liczbę wartości, ale sumowanie tych wartości jest pozbawione sensu.

W tego typu sytuacjach zmienną jakościową lub dyskretną o kilku czy kilkunastu kategoriach należy rozkodować na odpowiednią liczbę zmiennych zero-jedynkowych i używać tych zmiennych w równaniu regresji. Zmienne zero-jedynkowe są bardzo przydatnym narzędziem w analizie regresji. Taka zmienna przyjmuje wartość jeden, gdy jakieś zjawisko występuje i zero w przeciwnym przypadku.

Najprostszym typem zmiennej zero-jedynkowej jest zmienna wyróżniająca pewien okres czasu.

Przykład 3.

Estymujemy funkcję konsumpcji typu keynesowskiego:

$$C_t = \beta_0 + \beta_1 Y_t + \varepsilon_t \tag{1}$$

dysponujemy danymi kwartalnymi, i chcemy sprawdzić czy funkcja konsumpcji jest taka sama w każdym kwartale. W tym celu wprowadzamy zmienne zero-jedynkowe (dummy variables) po jednej dla każego kwartału.

$$C_t = \beta_0 + \beta_1 Y_t + \delta_1 D_{t1} + \delta_2 D_{t2} + \delta_3 D_{t3} + \delta_4 D_{t4} + \epsilon_t \tag{2}$$

ale wprowadzenie czterech zmiennych zero-jedynkowych do modelu spowoduje, że pojawi się współliniowość bowiem:

$$\begin{pmatrix} C_1 \\ C_2 \\ C_3 \\ C_4 \\ .. \\ C_n \end{pmatrix} = \begin{bmatrix} 1 & y_{t1} & 1 & 0 & 0 & 0 \\ 1 & y_{t2} & 0 & 1 & 0 & 0 \\ 1 & y_{t3} & 0 & 0 & 1 & 0 \\ 1 & y_{t4} & 0 & 0 & 0 & 1 \\ ... & ... & ... & ... & ... & ... \\ 1 & y_{tn} & 0 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

zmienne zero-jedynkowe sumują się dla każdej obserwacji w próbie do wektora jednostkowego, powodując że macierz X staje się osobliwa.

To zjawisko w literaturze ekonometrycznej nazywane jest pułapką związaną ze zmiennymi zero-jedynkowymi (*dummy variable trap*). Takiego modelu nie można oszacować, ponieważ wystąpi liniowa zależność między regresorami, a macierz $X'X$ będzie osobliwa. Dzieje się tak, ponieważ zmienne zero-jedynkowe sumują się do jedności $\sum \delta_i = l$.

Aby ominąć pułapkę w modelu ekonometrycznym należy pominąć zmienną zero-jedynkową dla jednej z kategorii. Zazwyczaj w praktyce odrzuca się tę kategorię dla której jest najwięcej obserwacji. Wtedy model jest prawidłowy dla większości obserwacji, a zmienne zero-jedynkowe mierzą odchylenia od stanu średniego powodowane przez inne kategorie rozpatrywanej zmiennej.

1.6.2 Interakcje

W modelu ekonometrycznym zakłada się, że poszczególne zmienne wpływają na zmienną zależną w sposób od siebie niezależny. Niekiedy to założenie jest mało realistyczne. Na przykład przy tłumaczeniu wysokości zarobków za pomocą między innymi płci i wykształcenia respondentów. Teoretycznie możemy uznać, że wykształcenie i płeć wpływają na wysokość uzyskiwanych zarobków w sposób od siebie niezależny. Chociaż z drugiej strony, z teorii rynku pracy wiemy o istnieniu zjawiska dyskryminacji płacowej kobiet, i że jest ona silniejsza wraz ze wzrostem poziomu wykształcenia. W takim przypadku warto taką informację wykorzystać w modelu zjawiska wprowadzając interakcje.

Badanie interakcji między zmiennymi ciągłymi sprowadza się do wprowadzenia do modelu odpowiednich iloczynów zmiennych. Zostanie ono szczegółowo omówione przy okazji doboru formy funkcyjnej modelu.

Jeżeli w modelu zawartych jest kilka cech jakościowych, np. wykształcenie o m_1 kategoriach, klasa miejscowości o m_2 kategoriach to w równaniu regresji mamy $(m_1 - 1) + (m_2 - 1)$ dodatkowych regresorów. Ale postępując w ten sposób zakładamy, że rozpatrywane cechy są niezależne i nie zachodzi żadna interakcja między nimi. Gdy chcemy zbadać efekty interakcyjne to powinniśmy wprowadzić $m_1 * m_2 - 1$ dodatkowych regresorów w równaniu regresji. Przy takim kodowaniu wybór zmiennej pozostającej poza zbiorem regresorów jest dowolny, tyle że od tego wyboru może zależeć interpretacja wyników.

Niekiedy może wystąpić sytuacja, że istnieje współzależność między zmiennymi objaśniającymi i jedna z tych zmiennych jest ciągła a druga dyskretna. Interakcję między tymi zmiennymi uwzględniamy wstawiając do modelu ilo-

czyny zmiennych zero-jedynkowych związanych z poziomami zmiennej dyskretnej i interesującej nas zmiennej ciągłej.

1.6.3 Zmienne o wielu kategoriach i efekty progowe (Threshold Effects)

W wielu zastosowaniach zmienne zero-jedynkowe używane są w celu modelowania czynników jakościowych takich jak np. przynależność do danej grupy, czy występowanie zjawiskaw danym okresie czasu. Jednak na tym nie kończą się możliwości stosowania tych zmiennych. Zmienne jakościowe mogą być również stosowane do pomiaru pewnych zjawisk, które mogą być mierzone metodą bezpośrednią. Wracając do przykładu wykształcenia lepszą jego miarą jest wzięcie pod uwagę osiągniętego poziomu wykształcenia, niż rozpatrywanie ilości lat nauki.

Przykład 4.

Przypuśćmy, że analizujemy następujący model badający zależność zarobków od wykształcenia i wieku:

$$zarobki = \beta_0 + \beta_1 wiek + wykształcenie + \varepsilon \tag{3}$$

Zbiór danych zawiera informacje o zarobkach, wieku oraz najwyższym osiągniętym wykształceniu przez respondenta. Zmienna ta przyjmuje trzy poziomy: podstawowe (P), średnie (S) i wyższe (W). Najprostszym sposobem analizy, aczkolwiek nie najlepszym, jest użycie zmiennej E równej 0 dla pierwszej grupy obserwacji, 1 dla drugiej i 2 dla trzeciej. Powstanie wtedy model:

$$zarobki = \beta_0 + \beta_1 wiek + \beta_2 E + \varepsilon \tag{4}$$

Jednak sprawia on trudności w analizie i interpretacji wyników. Tak jak w przykładzie 1, zakładamy że każda zmiana poziomu wykształcenia, czyli przekroczenie pewnej wartości progowej zmiennej objaśniającej, powoduje taki sam przyrost zarobków. Jednak w rzeczywistości takie zjawisko jest mało prawdopodobne i to założenie ogranicza regresję powodują obciążenie estymatorów. Zamiast modelu (4), możemy użyć modelu z dwoma zmiennymi zero-jedynkowymi:

$$zarobki = \beta_0 + \beta_1 wiek + \delta_w W + \delta_s S + \varepsilon \tag{5}$$

Zależność pomiędzy wykształceniem a dochodami wtedy jest dana przez:

$$\text{wyższe: } E[zarobki \mid wiek, W] = \beta_0 + \beta_1 wiek + \delta_w$$

$$\text{średnie: } E[zarobki \mid wiek, S] = \beta_0 + \beta_1 wiek + \delta_s$$
$$\text{podstawowe: } E[zarobki \mid wiek, P] = \beta_0 + \beta_1 wiek$$

Tym co nas interesuję są współczynniki δ_w i δ_s, oraz różnica między nimi. Jest ona łatwa do policzenia i interpretacji. Każdy współczynnik δ w równaniu (5) interpretujemy jako wzrost dochodu osiągany dzieki wyższemu poziomowi wykształcenia niż podstawowe, natomiast różnica $\delta_w - \delta_s$ pokazuje nam o ile więcej zarabiają ludzie z wyższym wykształceniem niż ludzie ze średnim wykształceniem przyjmując inne czynniki na stałym poziomie.

Przyjęty sposób rozkodowania zmiennej nie jest jedynym możliwym. Istnieje również inny sposób rozbicia zmiennej *wykształcenie* na zmienne zero-jedynkowe. Wartość 1 zmiennej zero-jedynkowej może oznaczać, że dana jednostka posiada dany poziom wykształcenia. W takim przypadku dla osoby z wyższym wykształceniem wszystkie zmienne zero-jedynkowe oznaczające niższe poziomy wykształcenia, które osoba osiągnęła, przyjmą wartość 1. Definiując zmienne w ten sposób zmieniamy również zależność między wykształceniem a dochodami:

$$\text{wyższe: } E[zarobki \mid wiek, W] = \beta_0 + \beta_1 wiek + \delta_w + \delta_s$$
$$\text{średnie: } E[zarobki \mid wiek, S] = \beta_0 + \beta_1 wiek + \delta_s$$
$$\text{podstawowe: } E[zarobki \mid wiek, P] = \beta_0 + \beta_1 wiek$$

Zamiast różnicy między wykształceniem wyższym a podstawowym, w tym modelu δ_w jest krańcową wartością wyższego wykształcenia.

Sposób w jaki rozbijemy zmienna o kilku kategoriach jest wyborem badacza i powinien odpowiadać celowi modelu. Oba sposoby są matematycznie równoważne.

Przykład 5.

Na podstawie danych pochodzących z Badania Aktywności Ekonomicznej Ludności (BAEL) dwóch badaczy zbudowało modele tłumaczące wysokość płacy w zależności od poziomu wykształcenia i zmiennych kontrolnych (płeć - 1 mężczyzna, staż pracy oraz jego kwadrat, zamieszkiwanie w dużym mieście). W danych źródłowych zmienna *wykształcenie* przyjmowała 5 wartości (podstawowe, zawodowe, średnie, policealne, wyższe). Badacz A tworząc zmienne 0-1 dla poziomów wykształcenia przypisał wartość jeden dla najwyższego osiągniętego wykształcenia i 0 dla pozostałych. Z kolei badacz B przypisał wartość jeden wszystkim poziomom wykształcenia, które osoba osiągnęła. Czyli np. dla osoby o wykształceniu średnim wartość 1 przyjmują zmienne dla wykształcenia podstawowego, zawodowego oraz średniego. Badacze otrzymali następujące wyniki:

```
                    Model badacza A                 Model Badacza B
                    Number of obs =   25794         Number of obs =   25794
                    F(  8, 25785) =  789.17         F(  8, 25785) =  789.17
                    Prob > F      =  0.0000         Prob > F      =  0.0000
                    R-squared     =  0.1967         R-squared     =  0.1967
                    Adj R-squared =  0.1964         Adj R-squared =  0.1964
                    Root MSE      =  216.12         Root MSE      =  216.12
------------------------------------------------------------------------------
     zarobki |      Coef.   Std. Err.  P>|t|       Coef.   Std. Err.  P>|t|
-------------+----------------------------------------------------------------
        plec |   64.78646   2.780403   0.000    64.78646   2.780403   0.000
        staz |   7.713932    .340798   0.000    7.713932    .340798   0.000
       staz2 |   -.192435    .007008   0.000    -.192435    .007008   0.000
 duze miasto |   78.40807   3.201374   0.000    78.40807   3.201374   0.000
wyksztal_wyz |   250.2458   5.429926   0.000    92.88095   8.530816   0.000
wyksztal_pol |   157.3648   8.019582   0.000    31.82745   7.728078   0.000
wyksztal_sre |   125.5374   4.024915   0.000    51.31021   3.479491   0.000
wyksztal_zaw |   74.22715   3.922634   0.000    74.22715   3.922634   0.000
       _cons |    50.6661   4.896975   0.000     50.6661   4.896975   0.000
------------------------------------------------------------------------------
```

1. Oceń właściwości statystyczne obu modeli oraz ich dopasowanie do danych empirycznych.

2. Zinterpretuj współczynnik dla wykształcenia średniego w obu modelach

3. Policz o ile przeciętnie więcej zarabia osoba z wykształceniem wyższym od osoby z wykształceniem zawodowym według modelu A, a o ile według modelu B?

4. Czy sposób kodowania zmiennej *wykształcenie* ma istotny wpływ na osiągane wyniki?

Odpowiedź

1. W obu modelach wszystkie zmienne objaśniające są pojedynczo istotne oraz łącznie istotne. Współczynnik $Adj-R^2$ świadczy o tym, że zmienne objaśniające w prawie 20 % wyjaśniają zmienność zarobków.

2. W modelu A współczynnik dla zmiennej wykształcenie średnie mówi o ile przeciętnie więcej zarabia osoba z wykształceniem średnim w stosunku do osoby o wykształceniu podstawowym. W modelu B współczynnik dla zmiennej wykształcenie średnie mówi o ile złotych więcej zarabia osoba o wykształceniu średnim od osoby o wykształceniu zawodowym.

3. W modelu A każdy współczynnik oznacza premię do zarobków z posiadania danego poziomu wykształcenia w stosunku do wykształcenia podstawowego. Wobec tego różnica w zarobkach między osobą o wykształceniu wyższym a wykształceniu zawodowym jest to różnica: $\beta_{wyzsze} - \beta_{zawodowe} = 250.25 - 74.23 = 176.02$. W modelu B każdy współczynnik oznacza premię do zarobków w stosunku do wykształcenia o „jeden stopień" niższego. Więc szukana różnica jest sumą $\beta_{srednie} + \beta_{policealne} + \beta_{wyzsze} = 51.31 + 32.83 + 92.88 = 176.02$.

4. Sposób rozkodowania zmiennej *wykształcenie* nie zmienia w żaden sposób wyników modelu. Zmieniają się wartości i interpretacja parametrów. Sposób liczenia przeciętnych różnic w zarobkach powodowanych przez różne poziomy wykształcenia jest inny dla każdego modelu, ale ostateczne rezultaty takie same.

Literatura

[1] William H. Greene (2003) *Econometric Analysis*, 5th edition.

[2] Jerzy Mycielski (2000) *Notatki do ćwiczeń z ekonometrii*, WNE.

[3] Aleksander Welfe (1998) *Zbiór zadań z ekonometrii*, PWE