Zmienne Binarne w Pakiecie `Stata`

Karol Kuhl

Zbiór (hipotetyczny) `dummy.dta` zawiera dane, na podstawie których prowadzono analizy opisane poniżej. Nazwy zmiennych oznaczają: `doch` – dochód w jednostkach pieniężnych; `plec` – płeć: kobieta (0), mężczyzna (1); `wiek` – wiek w latach; `eduk` – poziom wykształcenia: podstawowe (1), zawodowe (2), średnie (3), wyższe (4); `stan` – stan cywilny: panna/kawaler (1), zamężna/żonaty (2), wdowa/wdowiec (3).

1 Dyskretne zmienne nominalne

Najprostszym przykładem użycia zmiennej binarnej (zwanej również zmienną zerojedynkową) w analizie regresji jest sytuacja, w której regresor jest zmienną nominalną o dwóch kategoriach, np.: tak-nie, miasto-wieś, kobieta-mężczyzna. W przeciwieństwie do zmiennej porządkowej, nie ma znaczenia, która z tych kategorii będzie zakodowana za pomocą zera, a która za pomocą jedynki. Przykładem takiej zmiennej jest w zbiorze danych zmienna `plec`:

$$plec_i = \begin{cases} 0 \text{ dla kobiet,} \\ 1 \text{ dla mężczyzn.} \end{cases}$$

Teoretycznie nie ma żadnego znaczenia, w jaki sposób zakodowane zostaną poszczególne kategorie tej zmiennej. Można za pomoca polecenia „`generate sex=1-plec`" wygenerować nową zmienną `sex`:

$$sex_i = \begin{cases} 0 \text{ dla mężczyzn,} \\ 1 \text{ dla kobiet.} \end{cases}$$

Obydwie zmienne zawierają te same informacje. W takiej sytuacji, różnica pomiędzy modelami:

$$doch_i = \alpha_1 + \alpha_2 plec_i + \epsilon_{\alpha i},$$
$$doch_i = \beta_1 + \beta_2 sex_i + \epsilon_{\beta i},$$

sprowadza się do interpretacji współczynników regresji. W modelu α, wpółczynnik α_1 to średni dochód kobiet, a współczynnik α_2 to różnica pomiędzy średnim dochodem mężczyzn, a średnim dochodem kobiet. Średni dochód mężczyzn to $\alpha_1 + \alpha_2$. W modelu β, wpółczynnik pierwszy (β_1) to średni dochód mężczyzn, a współczynnik drugi (β_2) to różnica pomiędzy średnim dochodem kobiet, a średnim dochodem mężczyzn. Średni dochód kobiet to $\beta_1 + \beta_2$. Wyniki estymacji modeli α i β są następujące:

```
. regress doch plec
      Source |       SS       df       MS              Number of obs =     400
-------------+------------------------------           F(  1,   398) =   17.76
       Model |  40.5789154     1  40.5789154           Prob > F      =  0.0000
    Residual |  909.613846   398  2.28546193           R-squared     =  0.0427
-------------+------------------------------           Adj R-squared =  0.0403
       Total |  950.192762   399  2.38143549           Root MSE      =  1.5118
------------------------------------------------------------------------------
        doch |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
```

```
------------+----------------------------------------------------------------
       plec |   .6370158   .1511774     4.21   0.000     .3398097     .934222
      _cons |   12.89881   .1068986   120.66   0.000     12.68866    13.10897
------------------------------------------------------------------------------
. regress doch sex
      Source |       SS       df       MS              Number of obs =     400
-------------+------------------------------           F(  1,   398) =   17.76
       Model |  40.5789154     1  40.5789154           Prob > F      =  0.0000
    Residual |  909.613846   398  2.28546193           R-squared     =  0.0427
-------------+------------------------------           Adj R-squared =  0.0403
       Total |  950.192762   399  2.38143549           Root MSE      =  1.5118
------------------------------------------------------------------------------
       doch |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sex |  -.6370158   .1511774    -4.21   0.000     -.934222   -.3398097
      _cons |   13.53583   .1068986   126.62   0.000     13.32567    13.74598
------------------------------------------------------------------------------
```

Statystyki opisujące całość modelu (tabela analizy wariancji i inne – znajdujące się w górnej części) są w obydwu przypadkach identyczne. Oceny współczynników różnią się między sobą, ale zachowane zostały opisane wcześniej związki:

- Średni dochód kobiet wynosi: $\hat{\alpha}_1 \approx 12.90 = 13.54 - 0.64 = \hat{\beta}_1 + \hat{\beta}_2$.
- Średni dochód mężczyzn wynosi: $\hat{\alpha}_1 + \hat{\alpha}_2 \approx 12.90 + 0.64 = 13.54 = \hat{\beta}_1$.
- Różnica pomiędzy średnim dochodem mężczyzn, a średnim dochodem kobiet wynosi: $\hat{\alpha}_2 \approx 0.64 = -\hat{\beta}_2$.

Pomimo tego, że z perspektywy obliczeń, sposób zakodowania zmiennej binarnej jest nieistotny, należy to robić „z głową". W powyższym przykładzie (modelu analizującego wpływ płci na dochody) można było oczekiwać, że średnie dochody mężczyzn są wyższe od średnich dochodów kobiet. W związku z tym, wygodniej jest użyć zmiennej `plec`, ponieważ ocena współczynnika przy tej zmiennej, zgodnie z oczekiwaniami, powinna być dodatnia.

Czasami zmienne binarne nie są kodowane za pomocą zer i jedynek. Przykładowo można (za pomocą polecenia „`generate qqq=plec+1`") zdefiniować zmienną:

$$qqq_i = plec_i + 1 = \begin{cases} 1 \text{ dla kobiet,} \\ 2 \text{ dla mężczyzn.} \end{cases}$$

Oszacowanie modelu z tą zmienną da następujący rezultat:

```
. regress doch qqq
      Source |       SS       df       MS              Number of obs =     400
-------------+------------------------------           F(  1,   398) =   17.76
       Model |  40.5789154     1  40.5789154           Prob > F      =  0.0000
    Residual |  909.613846   398  2.28546193           R-squared     =  0.0427
-------------+------------------------------           Adj R-squared =  0.0403
       Total |  950.192762   399  2.38143549           Root MSE      =  1.5118
------------------------------------------------------------------------------
       doch |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        qqq |   .6370158   .1511774     4.21   0.000     .3398097     .934222
      _cons |    12.2618   .2390325    51.30   0.000     11.79187    12.73172
------------------------------------------------------------------------------
```

Pomimo tego, że wyniki oszacowania różnicy pomiędzy średnimi dochodami mężczyzn i kobiet są takie same, to nie należy w ten sposób postępować, ponieważ oszacowanie stałej jest niewłaściwe. Prawidłowym rozwiązaniem byłoby samodzielne zrekodowanie zmiennej `qqq` na zmienną `plec`, albo skorzystanie z polecenia „`xi:`", które ułatwia tego typu operacje:

```
. xi: regress doch i.qqq
i.qqq             _Iqqq_1-2           (naturally coded; _Iqqq_1 omitted)
      Source |       SS       df       MS              Number of obs =     400
-------------+------------------------------           F(  1,   398) =   17.76
       Model |  40.5789154     1  40.5789154           Prob > F      =  0.0000
    Residual |  909.613846   398  2.28546193           R-squared     =  0.0427
-------------+------------------------------           Adj R-squared =  0.0403
       Total |  950.192762   399  2.38143549           Root MSE      =  1.5118
------------------------------------------------------------------------------
        doch |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Iqqq_2 |   .6370158   .1511774     4.21   0.000     .3398097     .934222
       _cons |   12.89881   .1068986   120.66   0.000     12.68866    13.10897
------------------------------------------------------------------------------
```

Polecenie „`xi:`" powoduje, że każda zmienna objaśniająca, którą poprzedzi „`i.`", zostanie zamieniona na zestaw zmiennych binarnych. Liczba nowych zmiennych binarnych jest równa liczbie kategorii zmiennej objaśniającej minus jeden (opuszczana jest kategoria pierwsza w kolejności), w celu uniknięcia współliniowości (stąd komunikat: „`naturally coded; _Iqqq_1 omitted`"). Nazwy nowych zmiennych binarnych zawierają w sobie nazwę rekodowanej zmiennej i kody poszczególnych kategorii. Dlatego wyniki są identyczne z otrzymanymi podczas szacowania modelu α. Zastosowanie polecenia „`xi`" w sytuacji, gdy zmienna objaśniająca zakodowana jest w sposób właściwy („`xi: regress doch i.plec`") jest poprawne. W związku z tym, dobrą praktyką jest stosowanie tego polecenia zawsze, zamiast samodzielnego rekodowania.

Polecenie „`xi:`" jest szczególnie pomocne w sytuacji, gdy zmienna objaśniająca typu nominalnego ma więcej niż dwie kategorie, np. stan cywilny. W takim przypadku konieczne byłoby utworzenie $k-1$ zmiennych binarnych (gdzie k to liczba kategorii). Niech

$$stan_i = \begin{cases} 1 \text{ dla panny/kawalera} \\ 2 \text{ dla zamężnej/żonatego} \\ 3 \text{ dla wdowy/wdowca} \end{cases} .$$

Oszacowanie modelu, w którym dochód objaśniany jest stanem cywilnym odbywa się w sposób następujący:

```
. xi: regress doch i.stan
i.stan            _Istan_1-3          (naturally coded; _Istan_1 omitted)
      Source |       SS       df       MS              Number of obs =     400
-------------+------------------------------           F(  2,   397) =    0.29
       Model |  1.39765939     2  .698829697           Prob > F      =  0.7466
    Residual |  948.795103   397   2.3899121           R-squared     =  0.0015
-------------+------------------------------           Adj R-squared = -0.0036
       Total |  950.192762   399  2.38143549           Root MSE      =  1.5459
------------------------------------------------------------------------------
        doch |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Istan_2 |   .0533545   .1649276     0.32   0.746    -.2708862    .3775952
    _Istan_3 |  -.1370867   .2528464    -0.54   0.588     -.634172    .3599987
       _cons |   13.20936    .118918   111.08   0.000     12.97557    13.44315
------------------------------------------------------------------------------
```

W zbiorze danych pojawiły się 3 nowe zmienne, ale tylko dwie z nich zostały włączone do modelu. Wyniki oszacowania wskazują na to, że stan cywilny nie ma wpływu na dochody. Polecenie „xi:" może jednocześnie zrekodować więcej niż jedną zmienna nominalną, w związku z czym możliwe jest oszacowanie jednoczesnego wpływu stanu cywilnego i płci na wysokość dochodów:

```
. xi: reg doch i.stan i.plec
i.stan            _Istan_1-3          (naturally coded; _Istan_1 omitted)
i.plec            _Iplec_0-1          (naturally coded; _Iplec_0 omitted)
      Source |       SS       df       MS              Number of obs =     400
-------------+------------------------------           F(  3,   396) =    6.07
       Model |  41.7975269     3   13.932509           Prob > F      =  0.0005
    Residual |  908.395235   396  2.29392736           R-squared     =  0.0440
-------------+------------------------------           Adj R-squared =  0.0367
       Total |  950.192762   399  2.38143549           Root MSE      =  1.5146

------------------------------------------------------------------------------
        doch |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Istan_2 |   .0717434   .1616411     0.44   0.657    -.2460387    .3895254
    _Istan_3 |  -.0974061   .2478973    -0.39   0.695    -.5847655    .3899533
    _Iplec_1 |   .6361438   .1515846     4.20   0.000     .3381326    .9341551
       _cons |   12.87811   .1407258    91.51   0.000     12.60145    13.15478
------------------------------------------------------------------------------
```

W tym przypadku, raport ze zrekodowania zawiera informacje nt. każdej ze zmiennych i, co najważniejsze, informuje o tym, które kategorie zostały zostały uznane za tzw. kategorie bazowe (lub referencyjne). Ponownie, zmienne opisujące stan cywilny okazały się statystycznie nieistotne. Ostatni model można zapisać w następujący sposób:

$$doch_i = \delta_1 + \delta_2 stan_{2i} + \delta_3 stan_{3i} + \delta_4 sex_i + \epsilon_{\delta i}.$$

Interpretacja jego parametrów jest następująca:

- δ_1 to średni dochód panny, czyli osoby o charakterystykach bazowych (płci i stanie cywilnym).
- δ_2 to różnica pomiędzy średnim dochodem osób zamężnych/żonatych, a średnim dochodem panien, niezależnie od płci.
- δ_3 to różnica pomiędzy średnim dochodem wdów/wdowców, a średnim dochodem panien, niezależnie od płci.
- δ_4 to różnica pomiędzy średnim dochodem mężczyzn, a średnim dochodem panien, niezależnie od stanu cywilnego.

Ważnym zagadnieniem w kontekście zmiennych dyskretnych nominalnych o więcej niż dwóch kategoriach staje się testowanie istotności wpływu takich zmiennych na zmienną objaśnianą. Statystyki t przy zmiennych `stan_2` i `stan_3` służą do oddzielnej weryfikacji hipotez mówiących o nieistotności współczynników δ_2 i δ_3. Aby zweryfikować hipotezę $H_0: \delta_2 = \delta_3 = 0$ należy zastosować inny test. Robi się to po wyestymowaniu modelu, za pomocą polecenia „`test (_Istan_2=0) (_Istan_3=0)`", w wyniku czego otrzymuje się:

```
. test (_Istan_2=0) (_Istan_3=0)
 ( 1)  _Istan_2 = 0
 ( 2)  _Istan_3 = 0
       F(  2,   396) =    0.27
            Prob > F =    0.7669
```

Hipoteza zerowa tego typu testów mówi o tym, że łącznie obowiązują wszystkie ograniczenia na współczynniki. Dlatego niska wartość statystyki testującej F i towarzyszące jej prawdopodobieństwo większe od 5% powodują, że nie ma podstaw, żeby uznać, że te ograniczenia nie obowiązują. Zatem stan cywilny nie ma wpływu na wysokość dochodów. Po wyestymowaniu modelu, zmienne `_Istan_2` i `_Istan_3` są nadal dostępne. W poleceniu „`test`" (w przypadku KMRL) w każdym nawiasie wpisuje się jedno ograniczenie na kombinację liniową współczynników regresji, reprezentowanych przez nazwy zmiennych, przy których stoją. Liczba ograniczeń jest dowolna (w granicach zdrowego rozsądku), a w powyższym przykładzie testowano dwa ograniczenia.

2 Dyskretne zmienne porządkowe

Dla zmiennych dyskretnych porządkowych, możliwe jest jednoznaczne uporządkowanie kategorii, ale niemożliwe jest określenie ile razy kategoria wyższa różni się od kategorii niższej. Przykładem takiej zmiennej jest poziom wykształcenia. Można 4 poziomy uporządkować od najniższego (wykształcenie podstawowe) do najwyższego (wykształcenie wyższe), ale nie można np. stwierdzić ile razy wykształenie wyższe jest „lepsze" od wykształcenia średniego. Pomimo tej różnicy względem zmiennych dyskretnych nominalnych, sposób postępowania jest identyczny – używa się polecenia „`xi:`":

```
. xi: regress doch i.eduk
i.eduk            _Ieduk_1-4          (naturally coded; _Ieduk_1 omitted)
      Source |       SS       df       MS              Number of obs =     400
-------------+------------------------------           F(  3,   396) =    2.20
       Model |  15.5956551     3  5.19855171           Prob > F      =  0.0873
    Residual |  934.597107   396   2.3600937           R-squared     =  0.0164
-------------+------------------------------           Adj R-squared =  0.0090
       Total |  950.192762   399  2.38143549           Root MSE      =  1.5363
------------------------------------------------------------------------------
        doch |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Ieduk_2 |   .0715904     .22174     0.32   0.747    -.3643443    .5075251
    _Ieduk_3 |   .3286282    .210361     1.56   0.119    -.0849358    .7421922
    _Ieduk_4 |   .6441291   .2974954     2.17   0.031     .0592613    1.228997
       _cons |   12.99998   .1717591    75.69   0.000      12.6623    13.33765
------------------------------------------------------------------------------
```

Oszacowania współczynników modelu regresji opisują różnice pomiędzy średnimi dochodami poszczególnych poziomów wykształcenia, a poziomem podstawowym:

- średni dochód osób z wykształceniem podstawowym wynosi 13.00;
- średni dochód osób z wykształceniem zawodowym wynosi 13.00+0.07=13.07;
- średni dochód osób z wykształceniem średnim wynosi 13.00+0.33=13.33;
- średni dochód osób z wykształceniem wyższym wynosi 13.00+0.64=13.64.

W tym przykładzie, w macierzy danych $\mathbf{X}$ wiersze wyglądają następująco:

- $(1, 0, 0, 0)$ dla osób z wykształceniem podstawowym;
- $(1, 1, 0, 0)$ dla osób z wykształceniem zawodowym;
- $(1, 0, 1, 0)$ dla osób z wykształceniem średnim;
- $(1, 0, 0, 1)$ dla osób z wykształceniem średnim.

Możliwe są inne sposoby (wzorce) zakodowania zmiennych binarnych reprezentujących poziomy wykształcenia. Oczywiście inna będzie wtedy interpretacja współczynników. Przykładowo, można oszacować model z tzw. efektami progowymi. W tym przypadku w macierzy danych $\mathbf{X}$ wiersze wyglądają następująco:

- $(1, 0, 0, 0)$ dla osób z wykształceniem podstawowym;
- $(1, 1, 0, 0)$ dla osób z wykształceniem zawodowym;
- $(1, 1, 1, 0)$ dla osób z wykształceniem średnim;
- $(1, 1, 1, 1)$ dla osób z wykształceniem średnim.

Aby taki model wyestymować, należy zdefiniować odpowiednie zmienne:

```
generate d2=0
replace d2=1 if eduk>=2
generate d3=0
replace d3=1 if eduk>=3
generate d4=0
replace d4=1 if eduk>=4
```

W charakterze zmiennej `d1` wystąpi stała w modelu:

```
. reg doch d2 d3 d4
      Source |       SS       df       MS              Number of obs =     400
-------------+------------------------------           F(  3,   396) =    2.20
       Model |  15.5956551     3  5.19855171           Prob > F      =  0.0873
    Residual |  934.597107   396   2.3600937           R-squared     =  0.0164
-------------+------------------------------           Adj R-squared =  0.0090
       Total |  950.192762   399  2.38143549           Root MSE      =  1.5363
------------------------------------------------------------------------------
        doch |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          d2 |   .0715904     .22174     0.32   0.747    -.3643443    .5075251
          d3 |   .2570378    .185521     1.39   0.167    -.1076914     .621767
          d4 |   .3155009   .2715749     1.16   0.246    -.2184079    .8494097
       _cons |   12.99998   .1717591    75.69   0.000      12.6623    13.33765
```

Oszacowania współczynników tego modelu regresji opisują wysokości progów dochodowych, czyli różnice pomiędzy średnim dochodem osób z o pewnym poziomie wykształćeniem i średnim dochodem osób z wykształceniem o poziom niższym. :

- średni dochód osób z wykształceniem podstawowym wynosi 13.00;
- średni dochód osób z wykształceniem zawodowym wynosi 13.00+0.07=13.07;
- średni dochód osób z wykształceniem średnim wynosi 13.00+0.07+0.26=13.33;
- średni dochód osób z wykształceniem wyższym wynosi 13.00+0.07+0.26+0.62=13.65;

3 Interakcje zmiennych i regresja „łamana”

W modelu funkcji dochodów:

$$doch_i = \gamma_1 + \gamma_2 wiek_i + \epsilon_{\gamma i},$$

można przyjąć, że zarówno stała (γ_1), jak i współczynnik kierunkowy (γ_2) mogą się różnić w przypadku kobiet i mężczyzn. W takiej sytuacji można oszacować oddzielne modele dla kobiet i dla mężczyzn:

```
. regress doch wiek if plec==0
      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  1,   198) =  544.32
       Model |  200.948519     1  200.948519            Prob > F      =  0.0000
    Residual |   73.095755   198   .36917048           R-squared     =  0.7333
-------------+------------------------------           Adj R-squared =  0.7319
       Total |  274.044274   199   1.3771069           Root MSE      =  .60759
------------------------------------------------------------------------------
        doch |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wiek |   .0914526   .0039198    23.33   0.000     .0837226    .0991826
       _cons |   9.228819   .1630644    56.60   0.000     8.907253    9.550385
------------------------------------------------------------------------------
. regress doch wiek if plec==1
      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  1,   198) =  894.89
       Model |  520.422668     1  520.422668            Prob > F      =  0.0000
    Residual |  115.146905   198  .581550023           R-squared     =  0.8188
-------------+------------------------------           Adj R-squared =  0.8179
       Total |  635.569573   199  3.19381695           Root MSE      =  .76259
------------------------------------------------------------------------------
        doch |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wiek |   .1416249   .0047343    29.91   0.000     .1322888     .150961
       _cons |   7.595372    .205771    36.91   0.000     7.189588    8.001156
------------------------------------------------------------------------------
```

Można również oszacować na wszystkich obserwacjach model (zakładając jednakowe wariancje składnika losowego dla kobiet i dla mężczyzn):

$$doch_i = \lambda_1 + \lambda_2 plec_i + \lambda_3 wiek_i + \lambda_4 (plec_i * wiek_i) + \epsilon_{\lambda i}.$$

Iloczyn zmiennych `plec` i `plec` jest interakcją zmiennych. Model ten można rozpisać w sposób następujący:

$$doch_i = \begin{cases} \lambda_1 + \lambda_3 wiek_i + \epsilon_{\lambda i} & \text{dla kobiet,} \\ (\lambda_1 + \lambda_2) + (\lambda_3 + \lambda_4) wiek_i + \epsilon_{\lambda i} & \text{dla mężczyzn.} \end{cases}$$

Przykładem takich zależności jest następujący wynik estymacji:

```
. xi: regress doch i.plec*wiek
i.plec            _Iplec_0-1          (naturally coded; _Iplec_0 omitted)
i.plec*wiek       _IpleXwiek_#        (coded as above)
      Source |       SS       df       MS              Number of obs =     400
-------------+------------------------------           F(  3,   396) =  534.30
       Model |  761.950102     3  253.983367            Prob > F      =  0.0000
    Residual |  188.242659   396  .475360251           R-squared     =  0.8019
-------------+------------------------------           Adj R-squared =  0.8004
       Total |  950.192762   399  2.38143549           Root MSE      =  .68946

------------------------------------------------------------------------------
        doch |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Iplec_1 |  -1.633447   .2623903    -6.23   0.000      -2.1493   -1.117595
        wiek |   .0914526    .004448    20.56   0.000     .0827079    .1001973
```

```
_IpleXwiek_1 |   .0501723    .006173     8.13   0.000     .0380364    .0623082
       _cons |   9.228819   .1850364    49.88   0.000     8.865043    9.592595
------------------------------------------------------------------------------
```

Wyniki oszacowania potwierdzają opisane wyżej zależności:

- Stała dla kobiet wynosi: $\hat{\gamma}_1 = \hat{\lambda}_1 = 9.23$;
- Stała dla mężczyzn wynosi: $\hat{\gamma}_1 = \hat{\lambda}_1 + \hat{\lambda}_2 = 9.23 - 1.63 = 7.60$;
- Współczynnik kierunkowy dla kobiet wynosi: $\hat{\gamma}_2 = \hat{\lambda}_3 = 0.09$;
- Współczynnik kierunkowy dla mężczyzn wynosi: $\hat{\gamma}_2 = \hat{\lambda}_3 + \hat{\lambda}_4 = 0.09 - 0.05 = 0.14$.

Interakcje mogą zachodzić pomiędzy zmiennymi różnego typu i są sposobem na urozmaicenie postaci analizowanej funkcji w KMRL.

Specjalnym przypadkiem interakcji jest tzw. regresja „łamana". W modelu:

$$doch_i = \gamma_1 + \gamma_2 wiek_i + \epsilon_{\gamma i},$$

może być tak, że od pewnej granicznej wartości ($wiek^* = 40$) współczynnik nachylenia zmienia się powodując „złamanie" prostej regresji. W takiej sytuacji możliwe są dwa rozwiązania:

1. Można (za pomocą polecenia `generate w=0, replace w=1 if wiek>40`) do modelu wprowadzić zmienną binarną:

$$w_i = \begin{cases} 0 & \text{dla } wiek_i \leq wiek^*, \\ 1 & \text{dla } wiek_i > wiek^* \end{cases}$$

   i oszacować model z trzema zmiennymi objaśniającymi: `w` i `wiek` oraz interakcją tych zmiennych. Jednak w tym przypadku „złamanie" funkcji regresji może być jej przerwaniem – w punkcie $wiek^* = 40$ funkcja regresji może nie być ciągła. Wyniki takiego oszacowania są następujące:

```
. xi: regress doch i.w*wiek
i.w               _Iw_0-1             (naturally coded; _Iw_0 omitted)
i.w*wiek          _IwXwiek_#          (coded as above)
    Source |       SS       df       MS              Number of obs =     400
-----------+------------------------------           F(  3,   396) =  626.03
     Model | 784.731352     3  261.577117            Prob > F      =  0.0000
  Residual |  165.46141   396  .417831844            R-squared     =  0.8259
-----------+------------------------------           Adj R-squared =  0.8245
     Total | 950.192762   399  2.38143549            Root MSE      =   .6464
------------------------------------------------------------------------------
      doch |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
     _Iw_1 |   5.469265   .4732379    11.56   0.000     4.538892    6.399638
      wiek |   .2150592    .008584    25.05   0.000     .1981834     .231935
_IwXwiek_1 |  -.1477594   .0115297   -12.82   0.000    -.1704264   -.1250924
     _cons |   5.419094   .2689128    20.15   0.000     4.890419     5.94777
------------------------------------------------------------------------------
```

   W tym przypadku rzeczywiście następuje przerwanie wykresu funkcji:

$$doch_{w=0}(40) = 5.42 + 0.22 * 40 = 14.22 \neq$$
$$\neq 13.69 = 10.89 + 0.07 * 40 = (5.42 + 5.47) + (0.22 - 0.15) * 40 = doch_{w=1}(40).$$

2. Można do modelu wprowadzić zmienną ciągłą:

$$v_i = \begin{cases} 0 & \text{dla } wiek_i \leq wiek^*, \\ wiek_i - wiek^* & \text{dla } wiek_i > wiek^*, \end{cases}$$

utworzoną za pomocą polecenia „`mkspline u 40 v = wiek`", które automatycznie tworzy również zmienną `u`:

$$u_i = \begin{cases} wiek & \text{dla } wiek_i \leq wiek^*, \\ wiek^* & \text{dla } wiek_i > wiek^*. \end{cases}$$

Wtedy model będzie zawierać dwie zmienne objaśniające: `v` i `wiek`, a wyniki estymacji będą następujące:

```
. regress doch v wiek
   Source |       SS       df       MS              Number of obs =     400
----------+------------------------------           F(  2,   397) =  908.64
    Model |  779.832181     2  389.916091           Prob > F      =  0.0000
 Residual |   170.36058   397   .42911985           R-squared     =  0.8207
----------+------------------------------           Adj R-squared =  0.8198
    Total |  950.192762   399  2.38143549           Root MSE      =  .65507
------------------------------------------------------------------------------
     doch |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------+-------------------------------------------------------------------
        v |   -.145987   .0116726   -12.51   0.000    -.1689349   -.1230392
     wiek |   .1971328    .006894    28.59   0.000     .1835795    .2106861
    _cons |   5.913622   .2298843    25.72   0.000      5.46168    6.365565
------------------------------------------------------------------------------
```

W tym przypadku funkcja regresji będzie „złamana", a punkt tego złamania ($wiek^* = 40$) nazywa się węzłem.

Regresja może być „łamana" w wielu punktach i w ten sposób przybliżać dowolną nieliniową funkcję.