EKONOMETRIA ECONOMETRICS 1(51) • 2016

ISSN 1507-3866
e-ISSN 2449-9994

Dorota Raczkiewicz
Szkoła Główna Handlowa w Warszawie
e-mail: dbartos@sgh.waw.pl

ZASTOSOWANIE ANALIZY REGRESJI W REPREZENTACYJNYCH BADANIACH SPOŁECZNO-GOSPODARCZYCH

APPLICATION OF REGRESSION ANALYSIS IN SOCIO-ECONOMIC SAMPLE SURVEYS

DOI: 10.15611/ekt.2016.1.03
JEL Classification: B23

Streszczenie: Celem badawczym jest zaprezentowanie, w jaki sposób przeprowadzać analizy regresji klasycznej i logistycznej w badaniach reprezentacyjnych, opisujących zjawiska społeczno-gospodarcze, do których wylosowano próbę złożoną. Obiektem badań są gospodarstwa domowe w Polsce, ankietowane w badaniu budżetów gospodarstw domowych przeprowadzanym co roku przez Główny Urząd Statystyczny. Istota stosowanych metod analizy prób złożonych polega na uwzględnianiu odpowiedniego schematu losowania próby w estymacji, który obejmuje warstwowanie, ważenie, wielostopniowość losowania i korekty wynikające z błędów nielosowych. Oszacowania parametrów oraz oszacowania ich wariancji, mierzącej precyzję oszacowań tych parametrów, różnią się przy zastosowaniu odpowiednich procedur dla prób złożonych od wyników, które otrzymano by, gdyby zastosować procedury dla próby prostej. W artykule wykorzystano procedury SAS do regresji z prób złożonych. Było to możliwe ze względu na znaczny postęp w technikach obliczeniowych, w tym rozwój oprogramowania do modelowania, a także wzrost jego dostępności dla użytkowników.

Słowa kluczowe: analiza regresji, badania reprezentacyjne, próba złożona.

Summary: The aim of the article is to present how to carry out the classical and logistic regression analyses in sample surveys, describing the socio-economic phenomena, to which complex a sample was drawn. The object of the study are households in Poland, surveyed in the household budgets survey, conducted annually by the Central Statistical Office. The essence of the methods analysis of complex samples is based on taking into account an appropriate design sampling scheme in the estimation which includes stratification, weighing, multistage sampling and adjustments for non-sampling errors. Parameters' estimates and their variances' estimates which measure the precision of the parameters' estimates are different when using appropriate procedures for complex samples from the results which would be obtained if the procedures for simple sample were used. The SAS procedures for regression from complex samples were applied. It was possible due to the significant advances in computational techniques including the development of modeling software as well as increase its availability to users.

Keywords: regression analysis, sample surveys, complex sample.

1. Wstęp

Nauczając metod ilościowych, zajmujemy się głównie populacjami nieskończonymi, z których losuje się próby proste (losowanie niezależne). Na podstawie wyników uzyskanych z tych prób szacujemy parametry rozkładu badanej cechy w populacji lub typ jej rozkładu. Zakładamy, że nie występują błędy nielosowe. Rozpatrujemy tylko błędy losowe, zakładając identyczne i niezależne rozkłady zmiennych losowych. Tymczasem prawdziwy świat – obojętnie: fizyczny, biologiczny czy społeczny – rzadko odpowiada wymaganiom losowania próby prostej [Kish 1996]. W praktyce mamy często do czynienia z populacjami skończonymi, szczególnie w badaniach społeczno-gospodarczych: gospodarstw domowych, aktywności ekonomicznej ludności, dochodów i warunków życia, przedsiębiorstw, rolniczych. Populacjami skończonymi są więc na przykład: populacja ludności danego kraju, populacja gospodarstw domowych, gospodarstw rolnych, przedsiębiorstw, wyborców mających prawo wyborcze w danym kraju itd. Z takich populacji losuje się próby bezzwrotne, nie ankietujemy wielokrotnie tej samej jednostki. W przypadku losowania niezależnego każda z prób ma takie same szanse wylosowania. W skrajnym przypadku do próby może zostać wylosowana wielokrotnie (n razy) tylko jedna jednostka. Może się zdarzyć, że do próby trafią na przykład tylko jednostki małe albo tylko jednostki duże, pewne jednostki mogą nie być reprezentowane w próbie, mimo że ich udział w populacji jest znaczny. Próby takie byłyby niereprezentatywne. Aby temu zaradzić, przeprowadza się warstwowanie populacji przed wylosowaniem próby.

W wielu podręcznikach dużo miejsca poświęca się błędom losowym, ignoruje się zaś błędy nielosowe. Błędy losowe występują tylko w badaniach próbkowych i powstają wskutek poddania badaniu wylosowanej próby zamiast populacji. Oszacowanie ich wielkości jest stosunkowo łatwe, łatwo też jest je kontrolować, zależą one bowiem m.in. od wielkości próbki. Natomiast błędy nielosowe występują w każdym badaniu statystycznym – zarówno pełnym, jak i częściowym, ich udział w całkowitym błędzie badania może być znaczny i znacznie trudniej jest oszacować ich wielkość. W praktyce badań statystycznych, zwłaszcza społeczno-gospodarczych, nie można zakładać, że błędy nielosowe nie występują, trudno jest uniknąć błędów związanych z obserwacją, z brakiem odpowiedzi podczas przetwarzania danych. W badaniach społeczno-gospodarczych istnieje szczególnie duże ryzyko nieuzyskania danych; nawet do 50% respondentów odmawia udzielenia informacji. Błędy nielosowe wpływają więc znacznie na dokładność i jakość danych statystycznych.

W praktyce społeczno-gospodarczej często traktuje się zagregowane wyniki badań reprezentacyjnych, jakby pochodziły one z badania pełnego, to znaczy nie uwzględnia się w ogóle błędów losowych estymacji parametrów. A jeśli są one brane pod uwagę, to zwykle korzysta się ze standardowych procedur w komputerowych pakietach statystycznych, które są odpowiednie dla prób prostych uzyskanych w wyniku losowania niezależnego. W praktyce badań statystycznych, przy

ograniczonych zasobach (możliwościach techniczno-organizacyjnych, rzeczowych, finansowych i nakładach pracy ludzkiej), aby uzyskać precyzyjne oszacowania parametrów populacji, stosuje się złożone procedury losowania prób. Wiele badań nie opiera się na prostych próbach losowych, a zwykle na próbach złożonych: z nierównymi prawdopodobieństwami wyboru jednostek, z warstwowaniem, z zespołami i wieloma stopniami losowania. Standardowe procedury w komputerowych pakietach statystycznych, zastosowane dla prób złożonych, dają obciążone wyniki i zniekształcają wariancje estymatorów parametrów, czyli wnioskowanie statystyczne o populacji (estymacja parametrów i weryfikacja hipotez) może być niepoprawne.

W większości podręczników dotyczących metody reprezentacyjnej najwięcej miejsca poświęca się szacowaniu wartości średnich i wartości globalnych cechy badanej. Mniejszą wagę przywiązuje się do szacowania liczby i frakcji elementów wyróżnionych, co ma uzasadnienie w tym, że frakcja elementów wyróżnionych może być szacowana jako średnia wartości zmiennej zero-jedynkowej (przy przyporządkowaniu 1 – elementom wyróżnionym i 0 – elementom niewyróżnionym), liczba elementów wyróżnionych zaś może być szacowana jako wartość globalna tej zmiennej [Bracha 1996; Zasępa 1962]. Natomiast bardzo rzadko podejmuje się tematykę zależności pomiędzy zmiennymi, która może być szacowana przy wykorzystaniu różnych form analizy regresji. Pozwala ona na znalezienie mechanizmu powiązań między zmiennymi (przybliżenie go za pomocą funkcji matematycznej), a także na predykcję wartości zmiennej objaśnianej dla jednostek w przyszłości, jak również dla jednostek spoza próby. Ta tematyka występuje w pracach: [Bracha 1983; Koninj 1962; Kott 2007; Pfeffermann 1993]. Do oszacowania modeli regresji na podstawie danych z prób złożonych nie można wykorzystać klasycznej metody najmniejszych kwadratów, ponieważ:

- Macierz obserwacji na zmiennych objaśniających X jest macierzą losową i nie można zakładać, że w każdej możliwej próbie otrzymamy taką samą macierz X. Nie możemy więc rozpatrywać wektora β przy ustalonej macierzy X.
- Poszczególne obserwacje próby nie są niezależne oraz mają różne rozkłady (zależy to od zastosowanego schematu losowania próby).
- Należy uwzględnić schemat losowania próby i warstwowanie oraz zastosować odpowiednie wagi, wynikające z zastosowanego schematu losowania próby i warstwowania przed wylosowaniem próby oraz skorygowane ze względu na braki odpowiedzi i błędy pokrycia.

Celem artykułu jest zaprezentowanie, w jaki sposób przeprowadzać analizy regresji klasycznej i logistycznej w badaniach reprezentacyjnych, opisujących zjawiska społeczno-gospodarcze, do których wylosowano próbę złożoną.

2. Regresja z jedną zmienną objaśniającą

2.1. Regresja z jedną zmienną objaśniającą – próba prosta

Rozpatrujemy zależność zmiennej objaśnianej Y od zmiennej objaśniającej X w postaci prostej o równaniu:

$$Y_i/x_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

gdzie: Y_i – wartość zmiennej Y dla i-tej jednostki, x_i – wartość zmiennej X dla i-tej jednostki, β_0 i β_1 – nieznane parametry, Y_i są zmiennymi losowymi, dane zebrane w próbie prostej o liczebności n, wylosowanej z populacji nieskończonej, są jedną z realizacji tych n zmiennych losowych $\{y_i, i \in S\}$, ε_i są odchyleniami zmiennej objaśnianej wokół prostej opisanej przez model, przy następujących założeniach:

Z1: $E[\varepsilon_i] = 0$ dla wszystkich i, czyli $E(Y_i/x_i) = \beta_0 + \beta_1 x_i$.

Z2: $V[\varepsilon_i] = \sigma^2$ dla wszystkich i, czyli wariancja wokół prostej regresji jest taka sama dla wszystkich wartości X.

Z3: $cov[\varepsilon_i, \varepsilon_j] = 0$ dla $i \neq j$, tzn. obserwacje są nieskorelowane.

Estymatorami metody najmniejszych kwadratów parametrów β_0 i β_1 są $\hat{\beta}_0$ i $\hat{\beta}_1$, które minimalizują sumę kwadratów reszt $\sum[y_i - (\beta_0 + \beta_1 x_i)]^2$. Estymatory współczynnika regresji β_0 i wyrazu wolnego β_1 uzyskuje się poprzez wyznaczenie pierwszych pochodnych powyższej sumy i następnie rozwiązanie następującego układu równań normalnych:

$$\beta_0 n + \beta_1 \sum x_i = \sum y_i,$$

$$\beta_0 \sum x_i + \beta_1 \sum x_i^2 = \sum x_i y_i,$$

na podstawie czego otrzymuje się:

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i)}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2},$$

$$\hat{\beta}_0 = \frac{1}{n}(\sum y_i) - \hat{\beta}_1 \frac{1}{n}(\sum x_i).$$

Zarówno $\hat{\beta}_0$, jak i $\hat{\beta}_1$ są liniowe względem y, więc każde z nich można zapisać jako $\sum a_i y_i$ dla znanej stałej a_i:

$$\hat{\beta}_1 = \sum\nolimits_{i \epsilon S} \left[\frac{x_i - \frac{1}{n}(\sum x_i)}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} \right] y_i,$$

$$\hat{\beta}_0 = \sum_{i\epsilon S} \frac{1}{n}\left[1 - \frac{x_i \sum x_i - \frac{1}{n}(\sum x_i)^2}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}\right] y_i.$$

Jeśli spełnione są założenia: Z1, Z2 i Z3, wtedy $\hat{\beta}_0$ i $\hat{\beta}_1$ są najlepszymi nieobciążonymi estymatorami β_0 i β_1, tzn. że mają one najmniejszą wariancję wśród wszystkich liniowych estymatorów, które są nieobciążone.

Nieobciążonym estymatorem wariancji estymatora $\hat{\beta}_1$ jest:

$$\hat{V}(\hat{\beta}_1) = \frac{\frac{\sum_{i\epsilon S}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}}{\sum_{i\epsilon S}(x_i - \bar{x})^2}.$$

2.2. Regresja z jedną zmienną objaśniającą – próba złożona

W teorii metody reprezentacyjnej zajmujemy się szacowaniem takich parametrów populacji skończonej, jak wartość globalna $t_y = \sum_{i=1}^{N} y_i$ i średnia $\bar{y}_U = \frac{t_y}{N}$, gdzie N – liczebność populacji, U – populacja.

Wtedy parametry regresji dla skończonej populacji są parametrami najmniejszych kwadratów B_0 i B_1, które minimalizują sumę kwadratów reszt $\sum[y_i - B_0 - B_1 x_i]^2$ dla całej skończonej populacji. Wzory podano za [Lohr 2010].

Wtedy układ równań normalnych z podpunktu 1.1 można zapisać następująco:

$$B_0 N + B_1 \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i,$$

$$B_0 \sum_{i=1}^{N} x_i + B_1 \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i y_i,$$

a B_0 i B_1 można wyrazić jako funkcje wartości globalnych:

$$B_1 = \frac{\sum_{i=1}^{N} x_i y_i - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)}{\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2} = \frac{t_{xy} - \frac{t_x t_y}{N}}{t_{x^2} - \frac{(t_x)^2}{N}}$$

$$B_0 = \frac{1}{N}\left(\sum_{i=1}^{N} y_i\right) - B_1 \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right) = t_y - \frac{B_1 t_x}{N}.$$

Możemy oszacować osobno wartości globalne, wykorzystując wagi w sposób znany z metody reprezentacyjnej, czyli $\hat{N} = \sum_{i\epsilon S} w_i$, $\hat{t}_y = \sum_{i\epsilon S} w_i y_i$, $\hat{t}_x = \sum_{i\epsilon S} w_i x_i$, $\hat{t}_{xy} = \sum_{i\epsilon S} w_i x_i y_i$, $\hat{t}_{x^2} = \sum_{i\epsilon S} w_i x_i^2$; gdzie w_i – waga i-tej jednostki w próbie, która oznacza, że i-ta jednostka w próbie reprezentuje w_i jednostek w populacji.

Wtedy estymatory $\hat{B}_1$ i $\hat{B}_0$ parametrów regresji B_1 i B_0 mają postać:

$$\hat{B}_1 = \frac{\sum_{i\epsilon S} w_i\, x_i y_i - \frac{1}{\sum_{i\epsilon S} w_i}(\sum_{i\epsilon S} w_i\, x_i)(\sum_{i\epsilon S} w_i\, y_i)}{\sum_{i\epsilon S} w_i\, x_i^2 - \frac{1}{\sum_{i\epsilon S} w_i}(\sum_{i\epsilon S} w_i\, x_i)^2},$$

$$\hat{B}_0 = \frac{\sum_{i\epsilon S} w_i y_i - \hat{B}_1 \sum_{i\epsilon S} w_i x_i}{\sum_{i\epsilon S} w_i}.$$

Ze względu na to, że $\hat{B}_1$ i $\hat{B}_0$ są funkcjami wartości globalnych w populacji, szacowanymi na podstawie próby złożonej, oszacowanie wariancji złożonych estymatorów parametrów B_0 i B_1 w sposób analityczny jest bardzo utrudnione. Stąd wykorzystuje się pośrednie metody szacowania wariancji [Wolter 1985; Jakubowski, Bracha 2001], spośród których najbardziej rozpowszechniona jest linearyzacja Taylora.

Można tu zastosować linearyzację Taylora do estymacji wariancji estymatora $\hat{B}_1$, ponieważ B_1 jest funkcją pięciu wartości globalnych: $B_1 = h(t_{xy}, t_x, t_y, t_{x^2}, N)$, gdzie $h = (a, b, c, d, e) = \frac{a - \frac{bc}{e}}{d - \frac{b^2}{e}} = \frac{ea - bc}{ed - b^2}$.

Stąd estymator wariancji estymatora $\hat{B}_1$ ma postać:

$$V(\hat{B}_1) \approx V\left[\frac{\partial h}{\partial a}(\hat{t}_{xy} - t_{xy}) + \frac{\partial h}{\partial b}(\hat{t}_x - t_x) + \frac{\partial h}{\partial c}(\hat{t}_y - t_y) + \frac{\partial h}{\partial d}(\hat{t}_{x^2} - t_{x^2}) + \frac{\partial h}{\partial e}(\hat{N} - N)\right] = V\left[\left\{t_{x^2} - \frac{(t_x)^2}{N}\right\}^{-1} \sum_{i\epsilon S} w_i (y_i - B_0 - B_1 x_i)(x_i - \bar{x}_U)\right].$$

Definiując $q_i = (y_i - \hat{B}_0 - \hat{B}_1 x_i)(x_i - \hat{\bar{x}})$, gdzie: $\hat{\bar{x}} = \frac{\hat{t}_x}{\hat{N}}$, mamy:

$$\hat{V}_L(\hat{B}_1) = \frac{\hat{V}(\sum_{i\epsilon S} w_i\, q_i)}{\left[\sum_{i\epsilon S} w_i\, x_i^2 - \frac{(\sum_{i\epsilon S} w_i\, x_i)^2}{\sum_{i\epsilon S} w_i}\right]^2}.$$

3. Regresja wieloraka

3.1. Regresja wieloraka – próba prosta

Rozpatrujemy zależność cechy Y od p cech $X_1, X_2, \ldots, X_p$, wyrażoną w zapisie macierzowym następująco:

$$\mathbf{y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

gdzie: cechy $Y, X_1, X_2, \ldots, X_p$ przyjmują wartości $Y_i, X_{i1}, X_{i2}, \ldots, X_{ip}$ dla każdego i, gdzie i oznacza i-tą badaną jednostkę w próbie prostej o liczebności n, wylosowaną z populacji nieskończonej (i=1, 2,…, n),

oraz

$$\mathbf{y}= [Y_i]_{n*1}\ ,\ \mathbf{X_j}= [X_{ij}]_{1*p}\ ,\ \mathbf{X}= [X_{ij}]_{n*p}.$$

Przyjmując, że:

X jest $N * p$-wymiarową macierzą zaobserwowanych wartości zmiennych objaśniających, które są ustalone; $\boldsymbol{\beta} = [\beta_1 \quad \beta_2 \quad \dots \quad \beta_p]^T$ jest p-wymiarowym wektorem nieznanych współczynników regresji; $\boldsymbol{\epsilon}$ jest n-wymiarowym wektorem losowym, którego wektor wartości oczekiwanych wynosi $\boldsymbol{E}(\boldsymbol{\epsilon}) = 0$, a macierz kowariancji jest postaci:

$$\mathrm{V}(\boldsymbol{\epsilon}) = \mathrm{E}\,(\boldsymbol{\epsilon}\,\boldsymbol{\epsilon}^{\mathrm{T}}) = \sigma^2\,\mathbf{I},$$

wektor **β** szacuje się klasyczną metodą najmniejszych kwadratów za pomocą nieobciążonego estymatora o postaci:

$$\widehat{\boldsymbol{\beta}}= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\,(\mathbf{X}^{\mathrm{T}}\mathbf{y}),$$

którego macierz kowariancji estymatorów parametrów regresji ma postać:

$$\boldsymbol{V}\left(\widehat{\boldsymbol{\beta}}\right) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1},$$

a jej estymator przybiera postać:

$$\widehat{\boldsymbol{V}}\left(\widehat{\boldsymbol{\beta}}\right) = \hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}.$$

3.2. Regresja wieloraka – próba złożona

Rozpatrujemy zależność zmiennej objaśnianej y_i od p-wymiarowego wektora zmiennych objaśniających $\boldsymbol{x}_i$, gdzie: $\boldsymbol{x}_i = \left[x_{i1}, x_{i2}, \dots, x_{ip}\right]^T$. Chcemy oszacować p-wymiarowy wektor $\boldsymbol{B}$ parametrów populacji skończonej w modelu: $y = \boldsymbol{x}^T\boldsymbol{B}$. Zdefiniujmy:

$$\boldsymbol{y}_U = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{bmatrix} \text{ i } \boldsymbol{X}_U = \begin{bmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \dots \\ \boldsymbol{x}_N^T \end{bmatrix}.$$

Układ równań normalnych dla całej populacji skończonej ma postać:

$$\boldsymbol{X}_U^T\boldsymbol{X}_U\boldsymbol{B} = \boldsymbol{X}_U^T\boldsymbol{y}_U.$$

Zakładając, że istnieje $(\boldsymbol{X}_U^T\boldsymbol{X}_U)^{-1}$, dla skończonej populacji otrzymujemy:

$$\boldsymbol{B} = (\boldsymbol{X}_U^T\boldsymbol{X}_U)^{-1}\boldsymbol{X}_U^T\boldsymbol{y}_U.$$

Zarówno $\boldsymbol{X}_U^T\boldsymbol{X}_U$, jak i $\boldsymbol{X}_U^T\boldsymbol{y}_U$ są macierzami wartości globalnej w populacji:

$$\boldsymbol{X}_U^T\boldsymbol{X}_U = \sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i^T \text{ i } \boldsymbol{X}_U^T\boldsymbol{y}_U = \sum_{i=1}^{N}\boldsymbol{x}_i y_i.$$

Element o współrzędnych (*j*, k) macierzy o wymiarach p*p $\boldsymbol{X}_U^T\boldsymbol{X}_U$ wynosi $\sum_{i=1}^{N} x_{ij}x_{ik}$, a *k*-ty element *p*-wymiarowego wektora $\boldsymbol{X}_U^T\boldsymbol{y}_U$ wynosi $\sum_{i=1}^{N} x_{ik}y_i$.

W przypadku próby złożonej, wylosowanej z populacji skończonej, macierze $\boldsymbol{X}_U^T\boldsymbol{X}_U$ i $\boldsymbol{X}_U^T\boldsymbol{y}_U$ szacujemy przy użyciu wag. Szacujemy $\boldsymbol{X}_U^T\boldsymbol{X}_U = \sum_{i=1}^{N} \boldsymbol{x}_i\boldsymbol{x}_i^T$ przez $\sum_{i\epsilon S} w_i\boldsymbol{x}_i\boldsymbol{x}_i^T$ oraz $\boldsymbol{X}_U^T\boldsymbol{y}_U = \sum_{i=1}^{N} \boldsymbol{x}_i y_i$ przez $\sum_{i\epsilon S} w_i\boldsymbol{x}_i y_i$.

Wtedy estymator parametru **B** ma postać:

$$\widehat{\boldsymbol{B}} = \left(\sum_{i\epsilon S} w_i\boldsymbol{x}_i\boldsymbol{x}_i^T\right)^{-1} \sum_{i\epsilon S} w_i\boldsymbol{x}_i y_i.$$

Niech $\boldsymbol{q}_i = \boldsymbol{x}_i\left(y_i - \boldsymbol{x}_i^T\widehat{\boldsymbol{B}}\right)$, wtedy stosując linearyzację Taylora, otrzymujemy estymator macierzy kowariancji estymatorów parametrów regresji o postaci:

$$\hat{V}(\widehat{\boldsymbol{B}}) = \left(\sum_{i\epsilon S} w_i\boldsymbol{x}_i\boldsymbol{x}_i^T\right)^{-1} \hat{V}\left(\sum_{i\epsilon S} w_i\boldsymbol{q}_i\right)\left(\sum_{i\epsilon S} w_i\boldsymbol{x}_i\boldsymbol{x}_i^T\right)^{-1}.$$

4. Regresja logistyczna

4.1. Regresja logistyczna – próba prosta

Regresja logistyczna jest często wykorzystywana, aby przewidzieć prawdopodobieństwo otrzymania wartości 1 dla zmiennej dwukategorialnej y_i, która przyjmuje tylko dwie wartości: 1 (tak) lub 0 (nie).

Niech $\boldsymbol{x}$ będzie wektorem zmiennych objaśniających, a $\boldsymbol{\beta}$ wektorem nieznanych parametrów. Wtedy model regresji logistycznej ma postać:

$$p(\boldsymbol{x}) = \frac{exp(\boldsymbol{x}^T\boldsymbol{\beta})}{1 + exp(\boldsymbol{x}^T\boldsymbol{\beta})},$$

gdzie $p(\boldsymbol{x})$ oznacza prawdopodobieństwo, że jednostka ze zmiennymi objaśniającymi $\boldsymbol{x}$ przyjmie wartość 1. Model ten można zapisać również w postaci logitu, gdzie $logit(p) = ln\left[\frac{p}{1-p}\right]$, więc $logit[p(\boldsymbol{x})] = \boldsymbol{x}^T\boldsymbol{\beta}$.

Nieznane parametry wektora $\boldsymbol{\beta}$ szacujemy na podstawie próby metodą największej wiarygodności. Jest to metoda iteracyjna.

Niech $y_1,\ldots,y_n$ (i=1,...n) będą zaobserwowanymi wartościami zmiennej zależnej w *n*-elementowej próbie i niech $x_{1j},\ldots,x_{nj}$ będą wartościami *j*-tej zmiennej objaśniającej (j=1,...,k). Wynik próby możemy zapisać w postaci macierzowej jako:

$$Y = \begin{bmatrix} y_1 \\ . \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} x_{10} & x_{11} & \ldots & x_{1k} \\ . & . & \ldots & . \\ x_{n0} & x_{n1} & \ldots & x_{nk} \end{bmatrix},$$

gdzie: $x_{oi} = 1$ ($i = 1, \ldots, n$).

Zaobserwowane w próbie wartości $y_1,\ldots, y_n$ są realizacją n-wymiarowej zmiennej losowej $(Y_1,\ldots,Y_n)$. Każda ze zmiennych Y_i (i=1,...n) ma rozkład zero-jedynkowy o wartości średniej $pi = P(Y_i = 1)$, gdzie

$$P(Y_i = 1) = \frac{1}{1 - e^{-(\beta_0 x_{i0} + \beta_1 x_{i1} + ..+ \beta_k x_{ik})}}.$$

Jeżeli wartości zmiennych objaśniających są ustalone, to rozkład zmiennej losowej $(Y_1,\ldots,Y_n)$ zależy jedynie od parametrów $\beta_1,\ldots\ \beta_k$. Ponieważ zmienne losowe $Y_1,\ldots,Y_n$ są niezależne, prawdopodobieństwo otrzymania zaobserwowanych wartości $y_1,\ldots,y_n$ w próbie wynosi

$$P\left(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n\right) =$$
$$P\left(Y_1 = y_1\right) P\left(Y_1 = y_1\right) \cdot ... \cdot P\left(Y_1 = y_1\right) = \prod_{i=1}^{n} \mu_i^{y_i} \left(1 - \mu_i\right)^{1-y_i}.$$

Dla ustalonej próby powyższe prawdopodobieństwo jest funkcją parametrów $\beta_0,\ldots\ \beta_k$, zwaną funkcją wiarogodności próby:

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}.$$

Metoda największej wiarogodności (MNW) polega na szukaniu takich wartości nieznanych parametrów, dla których funkcja L przyjmuje wartość maksymalną. Bierze się to z założenia, że w wyniku wylosowania próby powinno zrealizować się zdarzenie o największym prawdopodobieństwie. Wartości estymatorów dla $\beta_0,\ldots\ \beta_k$ otrzymane metodą największej wiarogodności oznaczamy $b_0,\ldots,b_k$.

Ponieważ funkcja L osiąga maksimum w tych samych punktach co jej logarytm (tj. funkcja ln$L$), w praktyce wyznacza się maksimum funkcji ln$L$. Maksimum to znajduje się metodami rachunku różniczkowego, rozwiązując układ równań:

$$\frac{\partial \ln L}{\partial \beta_j} = 0,$$

$$j = 0, \ldots, k.$$

W tym przypadku $\ln L = \sum_{i=1}^{n} \left(y_i \ln\left(p_i\right) + \left(1 - y_i\right) \ln\left(1 - p_i\right)\right)$

$$\frac{\partial \ln L}{\beta_j} = \sum_{i=1}^{n} \left(y_i - p_i\right) x_{ij}.$$

Układ k+1 równań $\sum_{i=1}^{n} \left(y_i - p_i\right) x_{ij} = 0$ jest układem równań nieliniowych, który można rozwiązać, stosując iteracyjny algorytm Newtona-Raphsona.

4.2. Regresja logistyczna – próba złożona

Wykorzystując funkcję największej wiarygodności dla całej populacji skończonej o liczbie jednostek *N,* otrzymujemy funkcję wiarygodności próby [Lohr 2010]:

$$\mathcal{L}(\beta) = \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{1-y_i},$$

gdzie: $p_i = \frac{exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{1+exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}$ oznacza prawdopodobieństwo, że jednostka ze zmiennymi objaśniającymi $\boldsymbol{x}_i$ przyjmie wartość 1.

Parametr **B** populacji skończonej jest wtedy zdefiniowany jako estymator największej wiarygodności **β**. Parametr **B** jest rozwiązaniem układu równań: $\sum_{i=1}^{N} x_{ij}\left[y_i - \frac{exp(\boldsymbol{x}_i^T\boldsymbol{B})}{1+exp(\boldsymbol{x}_i^T\boldsymbol{B})}\right] = 0$ dla j = 1, …, *p*, jeśli wszystkie jednostki w populacji są obserwowane.

Estymatorem **B** jest $\widehat{\boldsymbol{B}}$ dany przez rozwiązanie układu równań:

$\sum_{i\epsilon S} w_i x_{ij}\left[y_i - \frac{exp(\boldsymbol{x}_i^T\widehat{\boldsymbol{B}})}{1+exp(\boldsymbol{x}_i^T\widehat{\boldsymbol{B}})}\right] = 0$ dla $j = 1, \ldots, p$, gdzie: *S* oznacza jednostki włączone do próby.

Estymację wariancji regresji logistycznej w przypadku próby złożonej można przeprowadzić, wykorzystując m.in. linearyzację Taylora.

W regresji logistycznej wyraz wolny jest jedynym szacowanym parametrem, na który wpływa schemat losowania próby. Natomiast w regresji liniowej schemat losowania próby wpływa na wszystkie szacowane parametry.

5.1. Materiał i metodyka

5.1. Opis próby

W opracowaniu wykorzystano dane z badania budżetów gospodarstw domowych, przeprowadzanego co roku przez Główny Urząd Statystyczny [GUS 2011]. Celem tego badania jest analiza poziomu życia ludności. Stanowi ono źródło informacji o przychodach, rozchodach, spożyciu ilościowym żywności, warunkach mieszkaniowych, wyposażeniu gospodarstw domowych w dobra trwałego użytku, a także ich subiektywnej ocenie sytuacji materialnej.

Badanie budżetów gospodarstw domowych jest przeprowadzane metodą reprezentacyjną na próbie gospodarstw domowych wylosowanych z populacji gospodarstw domowych w Polsce. Liczebność populacji wynosi około $N = 13$ mln gospodarstw domowych. Liczebność próby wynosi około $n = 37$ tys. gospodarstw domowych i stanowi 0,3% populacji.

Do wylosowania próby stosowany jest schemat dwustopniowy, warstwowy, z różnymi prawdopodobieństwami wyboru jednostek na I stopniu losowania. Jed-

nostkami losowania I stopnia są terenowe punkty badań, a jednostkami losowania II stopnia są mieszkania w wylosowanych terenowych punktach badań.

Parametry populacji w badaniach budżetów gospodarstw domowych są szacowane dla kraju jako całości. Stosuje się estymatory złożone z wykorzystaniem wag. Wagi pierwotne są odwrotnościami prawdopodobieństw wyboru do próby poszczególnych jednostek. W przypadku losowania wielostopniowego prawdopodobieństwa wylosowania jednostek do próby oraz wagi pierwotne otrzymuje się przez przemnożenie odpowiednio prawdopodobieństw i wag ze wszystkich stopni losowania próby, w tym przypadku z dwóch. Następnie z powodu nieprzystąpienia do badań około połowy gospodarstw domowych wagi pierwotne są korygowane danymi o strukturze gospodarstw domowych według liczby osób w podziale na miasto i wieś, pochodzącymi z Narodowego Spisu Powszechnego.

W badaniach budżetów gospodarstw domowych, podobnie jak w innych badaniach statystyki publicznej, szacowanymi parametrami populacji są średnie wartości globalne (sumy wartości zmiennej), frakcje i liczby elementów wyróżnionych oraz ilorazy dwóch średnich bądź dwóch wartości globalnych. Natomiast nie szacuje się żadnych modeli regresji zmiennych, stąd podjęto taką próbę w niniejszym opracowaniu.

5.2. Opis procedur SAS do modelowania na podstawie prób złożonych

Do oszacowania modeli regresji wykorzystano procedury dla prób złożonych, dostępne w oprogramowaniu SAS System: procedurę SURVEYREG i procedurę SURVEYLOGISTIC [SAS 2014]. Procedury te umożliwiają włączenie złożonego schematu losowania próby do analizy, zawierającego warstwowanie, wagi, wielostopniowość losowania za pomocą instrukcji odpowiednio STRATA, WEIGHT i CLUSTER.

Procedura SURVEYREG wykonuje analizę regresji dla danych z próby złożonej. Dopasowuje ona modele liniowe do danych z próby złożonej, obliczając m.in. współczynniki regresji oraz ich macierz wariancji i kowariancji, przedziały ufności dla parametrów regresji, a także wartości teoretyczne zmiennej objaśnianej przez model. Procedura ta wykorzystuje uogólnioną metodę najmniejszych kwadratów. Zakłada się, że współczynniki regresji są takie same w warstwach i jednostkach losowania pierwszego stopnia. Do oszacowania macierzy wariancji i kowariancji parametrów regresji używana jest domyślnie metoda linearyzacji Taylora, która zakłada liniowe przybliżenie estymatora parametrów regresji.

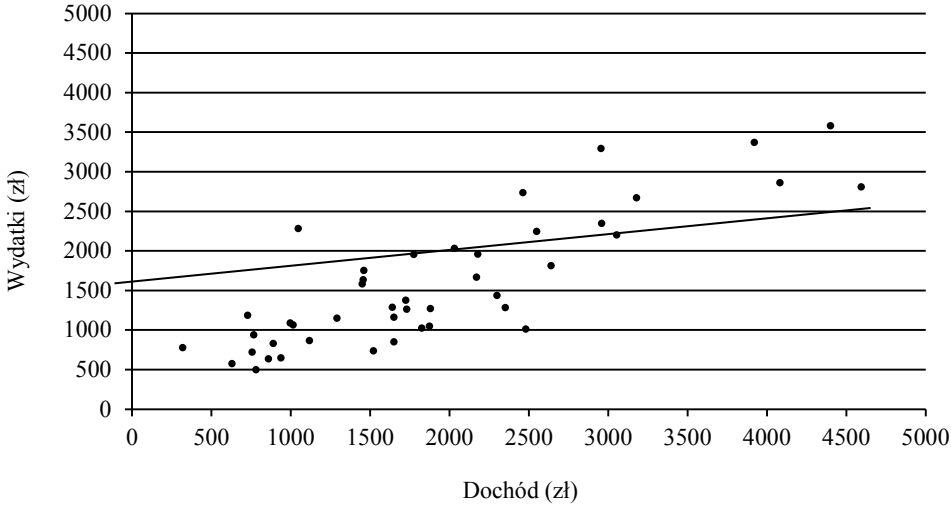
Procedura SURVEYLOGISTIC służy do badania związku między zmienną jakościową a zbiorem zmiennych objaśniających na podstawie danych z próby złożonej. Procedura ta dopasowuje modele liniowej regresji logistycznej dla jakościowej zmiennej z próby złożonej metodą największej wiarygodności.

6. Przykłady zastosowań procedur SAS do modelowania na podstawie prób złożonych

6.1. Przykład 1. Zastosowanie procedury SURVEYREG

Rozważamy zależność miesięcznych wydatków ogółem w zależności od miesięcznych dochodów ogółem w gospodarstwach domowych na podstawie danych z próby wylosowanej do badania budżetów gospodarstw domowych.

Gdyby była to próba prosta, czyli każda jednostka w próbie miałaby takie samo prawdopodobieństwo włączenia jej do próby, uzyskalibyśmy wykres rozrzutu punktów empirycznych jak na rys. 1 (dla lepszej widoczności na wykresie zamieszczono tylko 46 pierwszych obserwacji z nieuporządkowanego zbioru danych).

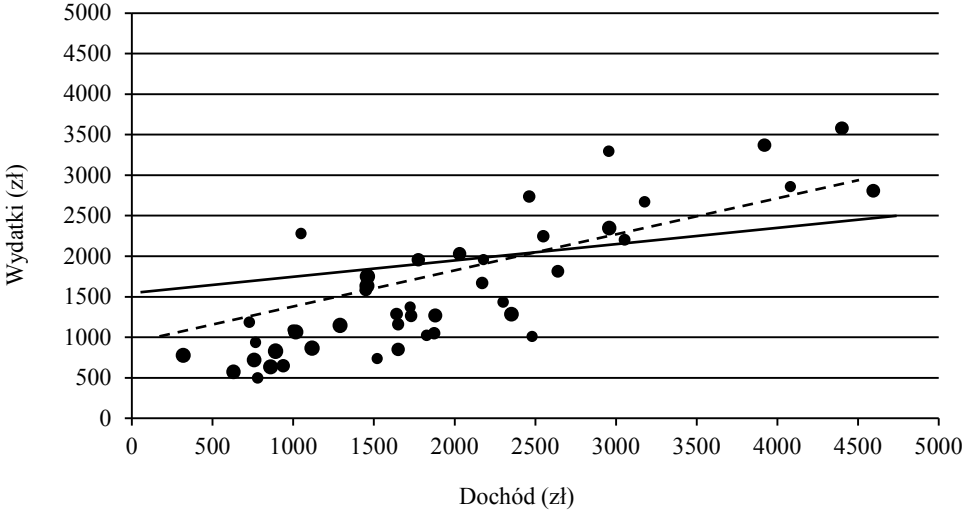


Rys. 1. Wykres rozrzutu miesięcznych wydatków ogółem względem miesięcznych dochodów ogółem w próbie gospodarstw domowych, bez uwzględnienia wag

Źródło: opracowanie własne.

Jednakże poszczególne jednostki w próbie mają różne prawdopodobieństwa włączenia ich do próby i w związku z tym różne wagi, próba więc nie jest próbą prostą, lecz złożoną. Stąd wykres rozrzutu wydatków ogółem względem miesięcznych dochodów ogółem powinien wyglądać jak na rys. 2 (również zamieszczono tylko 46 pierwszych obserwacji z nieuporządkowanego zbioru danych, tych samych co na rys. 1).

Rysunek 2 różni się od rysunku 1 tym, że wszystkie punkty empiryczne na rysunku 1 mają taką samą wielkość, podczas gdy wielkość punktów empirycznych na rysunku 2 jest różna i proporcjonalna do wielkości wag dla poszczególnych jednostek próby.



Rys. 2. Wykres rozrzutu miesięcznych wydatków ogółem względem miesięcznych dochodów ogółem w próbie gospodarstw domowych, z uwzględnieniem wag

Źródło: opracowanie własne.

Uwzględnienie wag wpływa zarówno na wygląd wykresu rozrzutu punktów empirycznych regresji, jak i na położenie prostej regresji (por. prostą zaznaczoną linią ciągłą z prostą zaznaczoną linią przerywaną na rys. 2) oraz na wszystkie wyniki oszacowań modelu regresji. W tabeli 1 dokonano porównania wyników analizy regresji dla próby złożonej (z uwzględnieniem wag, jednostek losowania pierwszego stopnia i warstwowania jednostek losowania pierwszego stopnia) przy wykorzystaniu procedury SURVEYREG w SAS z wynikami analizy regresji, jakie otrzymano

Tabela 1. Porównanie wyników analizy regresji wydatków (zł) względem dochodów (zł) na podstawie próby złożonej gospodarstw domowych z hipotetyczną próbą prostą

Próba	Prosta	Złożona
Procedura SAS	REG	SURVEYREG
Równanie modelu regresji	$\widehat{wyd}=1608,17+0,33056doch$	$\widehat{wyd}=1429,11+0,38178doch$
Błędy standardowe oszacowań parametrów strukturalnych regresji	*13,48 0,00319*	*216,70 0,07396*
t	*119,34 103,47*	*6,56 5,16*
p	*<0,0001 <0,0001*	*<0,0001 <0,0001*
Współczynnik korelacji	*0,4719*	*0,5090*
Współczynnik determinacji	*0,2227*	*0,2591*
F	*10705,416*	
p	*<0,0001*	
Błąd modelu regresji	*1826,45*	*1806,97*
Współczynnik zmienności resztowej	*70,20%*	*70,63%*

Źródło: opracowanie własne.

by, gdyby potraktowano tę próbę jako prostą i zastosowano procedurę REG w SAS. W procedurze SURVEYREG wykorzystano linearyzację Taylora do oszacowania wariancji estymatorów parametrów regresji.

Punktowe oszacowanie parametrów regresji przy wykorzystaniu wag jest następujące: współczynnik regresji wynosi 0,38178 dla próby złożonej i jest wyższy o 0,05122, niż byłby dla próby prostej (0,33056). Oznacza on, że wraz ze wzrostem dochodu gospodarstwa domowego o 1 zł wydatki ogółem rosną średnio o 38 groszy przy innych warunkach niezmienionych.

Oszacowanie wyrazu wolnego regresji wynosi 1429,11 dla próby złożonej i jest niższe o 179,06 zł, niż byłoby dla próby prostej (1608,17). Oznacza ono wydatki autonomiczne, które musi ponieść gospodarstwo domowe, aby przeżyć, niezależnie od posiadanego dochodu.

6.2. Przykład 2. Zastosowanie procedury SURVEYLOGISTIC

Rozważamy zależność posiadania bądź nieposiadania komputera z dostępem do Internetu w zależności od:

- klasy miejscowości zamieszkania gospodarstwa domowego,
- liczby dzieci do 17 roku życia w gospodarstwie domowym,
- stanu cywilnego głowy gospodarstwa domowego,
- poziomu wykształcenia głowy gospodarstwa domowego.

Dane pochodzą z próby wylosowanej do badania budżetów gospodarstw domowych, jak w przykładzie poprzednim.

Zmienna objaśniana jest jakościowa, zero-jedynkowa i przyjmuje wartość 1, jeśli gospodarstwo domowe ma komputer z dostępem do Internetu, oraz wartość 0, jeśli gospodarstwo domowe nie ma komputera z dostępem do Internetu. Zmienne objaśniające są jakościowe z więcej niż dwiema kategoriami.

Zastosowano procedurę SURVEYLOGISTIC do oszacowania modelu regresji logistycznej na podstawie danych z próby złożonej (z uwzględnieniem wag, jednostek losowania pierwszego stopnia i ich warstwowania). Do oszacowania wariancji estymatorów parametrów regresji logistycznej zastosowano linearyzację Taylora.

W tabeli 2 zamieszczono szczegółowe wyniki analizy regresji logistycznej posiadania bądź nieposiadania komputera z dostępem do Internetu względem czterech rozpatrywanych zmiennych objaśniających dla próby złożonej (z uwzględnieniem wag, jednostek losowania pierwszego stopnia i warstwowania jednostek losowania pierwszego stopnia) przy wykorzystaniu procedury SURVEYLOGISTIC. Wszystkie rozpatrywane zmienne objaśniające wpływają istotnie na fakt posiadania bądź nieposiadania komputera z dostępem do Internetu przez ogół gospodarstw domowych w Polsce (p<0,05).

W tabeli 3 zamieszczono wyniki analizy regresji logistycznej posiadania bądź nieposiadania komputera z dostępem do Internetu względem czterech rozpatrywanych zmiennych objaśniających, jakie otrzymano by, gdyby potraktowano tę próbę

jako prostą i zastosowano procedurę LOGISTIC. Wyniki te różnią się od wyników zamieszczonych w tab. 2.

W zastosowanym sposobie kodowania zmiennych objaśniających efekt dla każdej kategorii zmiennej objaśniającej jest porównywany z całkowitym efektem tej zmiennej. Współczynnik regresji dla kategorii odniesienia nie jest wyświetlany w tablicy wyników i musimy go samodzielnie obliczyć poza programem, wiedząc, że współczynniki regresji dla wszystkich kategorii danej zmiennej sumują się do zera, więc jest to suma wyświetlanych współczynników ze zmienionym znakiem.

Na rysunku 3 zilustrowano ilorazy szans nieposiadania komputera z dostępem do Internetu względem czterech rozpatrywanych zmiennych objaśniających. Szanse nieposiadania komputera z dostępem do Internetu są większe w gospodarstwach domowych na wsi, bez dzieci do 17. roku życia, z owdowiałą głową gospodarstwa domowego, z wykształceniem podstawowym lub niższym.

Tabela 2. Wyniki analizy regresji logistycznej posiadania bądź nieposiadania komputera z dostępem do Internetu względem czterech rozpatrywanych zmiennych objaśniających – próba złożona

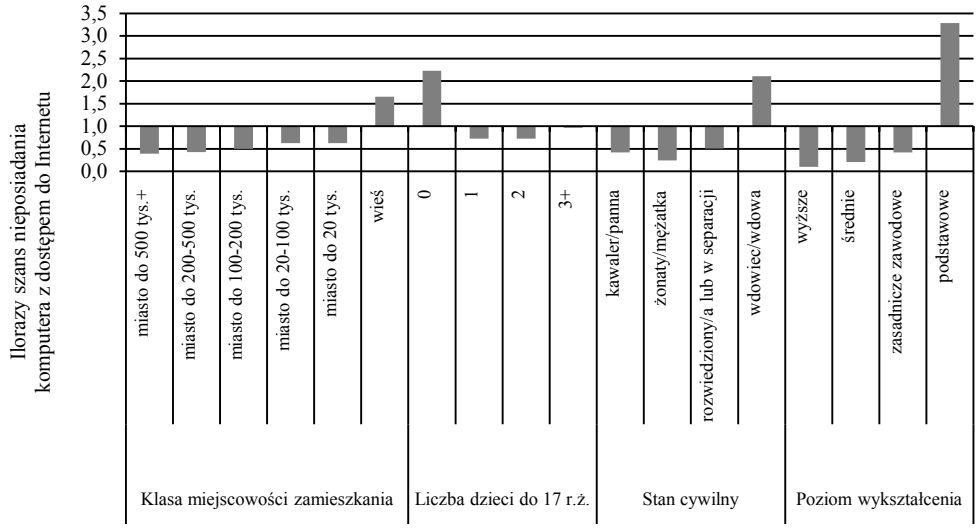
Zmienna	Kategoria	Ocena	Błąd standardowy	χ^2 Walda	*p*	Iloraz szans
Wyraz wolny	–	0,0707	0,0277	6,4925	0,0108	–
Klasa miejscowości zamieszkania	miasto do 500 tys.+	–0,3740	0,038	97,0309	<0,0001	0,387
	miasto do 200-500 tys.	–0,2710	0,0472	33,0189	<0,0001	0,429
	miasto do 100-200 tys.	–0,1289	0,0552	5,4442	0,0196	0,495
	miasto do 20-100 tys.	0,1016	0,0376	7,3023	0,0069	0,623
	miasto do 20 tys.	0,0978	0,0458	4,5642	0,0326	0,621
	wieś	0,5038	–	–	–	1,655
Liczba dzieci do 17. Roku życia	0	0,7639	0,0219	1220,5006	<0,0001	2,231
	1	–0,3608	0,0255	199,628	<0,0001	0,724
	2	–0,3648	0,0289	159,2528	<0,0001	0,722
	3+	–0,0383	–	–	–	0,962
Stan cywilny głowy	kawaler/panna	–0,1169	0,0343	11,6463	0,0006	0,422
	żonaty/mężatka	–0,6816	0,0222	944,5894	<0,0001	0,24
	Rozwiedziony/a lub w separacji	0,0536	0,0352	2,323	0,1275	0,501
	wdowiec/wdowa	0,7449	–	–	–	2,106
Poziom wykształcenia głowy	wyższe	–1,1309	0,0285	1576,6501	<0,0001	0,098
	średnie	–0,3748	0,02	352,7242	<0,0001	0,209
	zasadnicze zawodowe	0,3168	0,0214	219,951	<0,0001	0,418
	podstawowe lub niższe	1,1889	–	–	–	3,283

Źródło: opracowanie własne.

Tabela 3. Wyniki analizy regresji logistycznej posiadania bądź nieposiadania komputera z dostępem do Internetu względem czterech rozpatrywanych zmiennych objaśniających – hipotetyczna próba prosta

Zmienna	Kategoria	Ocena	Błąd standar-dowy	χ^2 Walda	p	Iloraz szans
Wyraz wolny	–	–	0,0237	1,5546	0,2125	
Klasa miejscowości zamieszkania	miasto do 500 tys.+	–0,3488	0,0313	124,4035	<0,0001	0,706
	miasto do 200-500 tys.	–0,2454	0,0345	50,5765	<0,0001	0,782
	miasto do 100-200 tys.	–0,1429	0,037	14,9482	0,0001	0,867
	miasto do 20-100 tys.	0,1116	0,0263	17,9989	<0,0001	1,118
	miasto do 20 tys.	0,1165	0,0322	13,068	0,0003	1,124
	wieś	0,4794	–	–	–	1,615
Liczba dzieci do 17. roku życia	0	0,7274	0,0207	1231,5955	<0,0001	2,070
	1	–0,3490	0,0243	205,6397	<0,0001	0,705
	2	–0,3527	0,0271	169,2008	<0,0001	0,703
	3+	–0,0257	–	–	–	0,975
Stan cywilny głowy	kawaler/panna	–0,0397	0,0318	1,5591	0,2118	0,961
	żonaty/mężatka	–0,563	0,0209	723,4908	<0,0001	0,569
	rozwiedziony/a lub w separacji	0,0303	0,0346	0,7706	0,3800	1,031
	wdowiec/wdowa	0,5724	–	–	–	1,773
Poziom wykształcenia głowy	wyższe	–1,1486	0,0271	1796,7001	<0,0001	0,317
	średnie	–0,3800	0,0194	385,2212	<0,0001	0,684
	zasadnicze zawodowe	0,3109	0,0201	239,8211	<0,0001	1,365
	podstawowe lub niższe	1,2177	–	–	–	3,379

Źródło: opracowanie własne.



Klasa miejscowości zamieszkania, liczba dzieci do 17 r.ż., stan cywilny, poziom wykształcenia

Rys. 3. Ilorazy szans nieposiadania komputera z dostępem do Internetu

Źródło: opracowanie własne.

7. Zakończenie

W opracowaniu zaprezentowano metody szacowania parametrów regresji na podstawie prób złożonych, losowanych najczęściej w badaniach zjawisk społeczno-gospodarczych. W analizach takich należy uwzględniać strukturę danych, tj. schemat losowania próby, w tym warstwowanie, wagi i wielostopniowość losowania, a także korekty wynikające z braków danych, których udział w badaniach społeczno-gospodarczych jest znaczny.

Rozwój oprogramowania statystycznego w ostatnich latach umożliwia praktyczne zastosowanie analizy regresji w badaniach, w których zastosowano złożone schematy losowania próby. Ponadto zwiększa się dostępność takiego oprogramowania dla użytkowników. Wcześniej odpowiednie procedury były tworzone na użytek wewnętrzny badaczy i niedostępne dla szerszej rzeszy użytkowników, w tym studentów i badaczy z innych dziedzin niż statystyka. Obecnie procedury te są dołączane do powszechnie używanych statystycznych pakietów komputerowych, takich jak m.in.: SAS, SPSS i STATA.

Literatura

Bracha C., 1983, *Regresja liniowa w badaniach reprezentacyjnych*, SGPiS, Warszawa.

Bracha C., 1996, *Teoretyczne podstawy metody reprezentacyjnej*, WN PWN, Warszawa.

GUS, 2011, *Metodologia badań budżetów gospodarstw domowych*, Zeszyty Metodyczne i Klasyfikacje, Warszawa.

Jakubowski J., Bracha C., 2001, *Przybliżone szacowanie wariancji w przypadku złożonych schematów losowania,* Z Prac Zakładu Badań Statystyczno-Ekonomicznych, zeszyt 273, GUS, Warszawa.

Kish L., 1996, *Stuletnie zmagania o badania reprezentacyjne*, Wiadomości Statystyczne, nr 8, s. 3-16.

Koninj H.S., 1962, *Regression analysis in sample surveys*, Jasa, vol. 57, no. 299, s. 590-606.

Kott P.S., 2007, *Clarifying some issues in the regression analysis of survey data*, Survey Research Methods, vol. 1, no. 1, s. 11-18.

Lohr S.L., 2010, *Sampling: Design and Analysis*, Brooks/Cole, Cengage Learning.

Materiały SAS Institute dostępne na stronie: support.sas.com, SAS, 2014.

Pfeffermann D., 1993, *The role of sampling weights when modeling survey data*, International Statistical Review, vol. 61, no. 2, s. 317-337.

Wolter K.M., 1985, *Introduction to Variance Estimation*, Springer-Verlag, New York Berlin Heidelberg Tokyo.

Zasępa R., 1962, *Badania statystyczne metodą reprezentacyjną*, PWN, Warszawa.