PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 7 (1207) 2008
TAKSONOMIA 15
Klasyfikacja i analiza danych – teoria i zastosowania

Jadwiga Kostrzewska
Uniwersytet Ekonomiczny w Krakowie

ZASTOSOWANIE WYBRANYCH MODELI TOBITOWYCH DO OPISU TYGODNIOWEJ LICZBY GODZIN PRACY

1. Wstęp

W badaniach statystycznych często występują zmienne, których wartości nie są znane, gdyż z pewnych powodów nie są możliwe do zaobserwowania lub nie są dostępne dla badacza. Analizie poddaje się jedynie obserwowalną realizację tych zmiennych. Taki charakter mają m.in. zmienne ograniczone: ucięte, cenzurowane lub podlegające nieprzypadkowej selekcji. Odpowiednim narzędziem od ich opisu są modele tobitowe.

Standardowy model tobitowy został wprowadzony w 1958 r. przez J. Tobina, który rozważył zmienną zależną cenzurowaną w kontekście regresji. Zauważył on, że wnioskowanie na podstawie danych bez uwzględnienia cenzurowania lub ucięcia, w przypadku gdy ono występuje, zawsze jest obciążone pewnym błędem (por. [Tobin 1958]). Na podstawie tego modelu zostały skonstruowane kolejne, bardziej złożone modele, tworzące rodzinę modeli zmiennej zależnej ograniczonej[1] (*limited dependent variable model* – LDV). W zależności od tego, czy zmienna zależna podlega ucięciu, cenzurowaniu czy nieprzypadkowej selekcji, można rozważać: modele regresji uciętej (*truncated regression model*) oraz różne cenzurowane modele regresji (*censored regression model*), czyli modele tobitowe (*tobit model*).

2. Podstawowe definicje

Niech Y^* będzie zmienną ukrytą. Można zdefiniować następujące obserwowalne realizacje Y tej zmiennej[2].

[1] Szeroki opis modeli zmiennej zależnej ograniczonej można znaleźć w [Maddala 1983], podstawowe modele wraz ze sposobami ich estymacji – np. w [Greene 2003; Verbeek 2001].

[2] Dla ustalenia uwagi w artykule podano definicje dla ograniczenia lewostronnego. Bez straty ogólności w definicjach (1)–(3) można przyjąć $a = 0$ oraz $c = 0$, uwzględniając ewentualnie przesunięcie zmiennych.

Zmienna ucięta (*truncated variable*). Obserwowane są tylko te wartości zmiennej Y^*, które są powyżej wartości progowej c (por. rys. 1), tzn.:

$$y_i = \begin{cases} y_i^* & \text{gdy} \quad y_i^* > c \\ \text{brak} & \text{poza tym} \end{cases}, \text{ dla } i = 1, ..., n. \tag{1}$$

Przykładem zmiennej uciętej jest zmienna określająca wysokość dochodów w grupie osób zarabiających poniżej średniej krajowej. Dokładne wysokości dochodów obserwowane są tylko wówczas, gdy nie przekraczają wyznaczonego poziomu, i tylko te wartości są rozważane w dalszej analizie (ucięcie prawostronne). Nie są znane (lub nie są dostępne) wartości zmiennej dotyczące dochodów przekraczających średnią krajową.

Zmienna cenzurowana (*censored variable*). Obserwowane są tylko te wartości zmiennej Y^*, które są powyżej wartości progowej c. Natomiast nie są znane wartości zmiennej Y^* występujące poniżej progu i jest im przypisana umowna wartość, najczęściej wartość progowa (por. rys. 1). Poniżej podano formalną definicję:

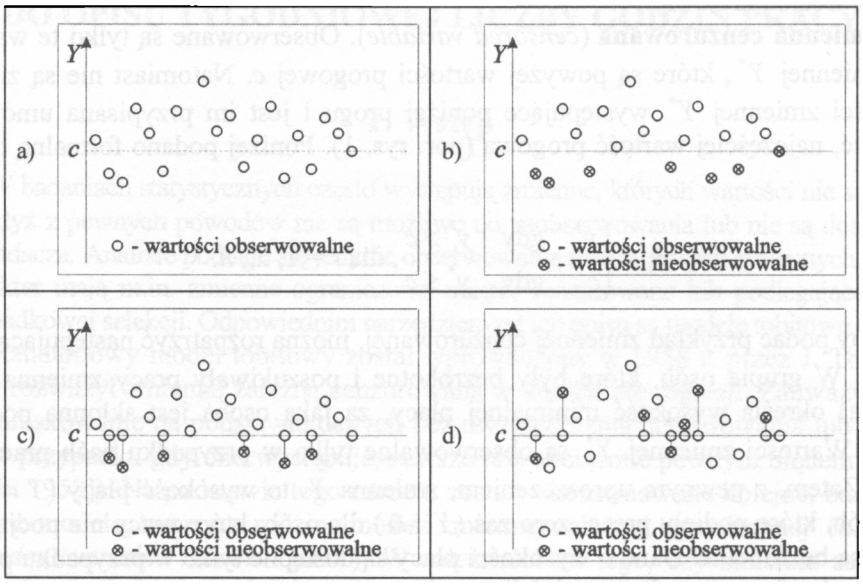
$$y_i = \begin{cases} y_i^* & \text{gdy} \quad y_i^* > c \\ c & \text{gdy} \quad y_i^* \leq c \end{cases}, \text{ dla } i = 1, ..., n. \tag{2}$$

Aby podać przykład zmiennej cenzurowanej, można rozpatrzyć następującą sytuację. W grupie osób, które były bezrobotne i poszukiwały pracy zmienna Y^* (ukryta) określa wysokość minimalnej płacy, za jaką osoba jest skłonna podjąć pracę. Wartości zmiennej Y^* są obserwowalne tylko w przypadku osób pracujących. Zatem, z pewnym uproszczeniem, zmienna Y to wysokość płacy ($Y = Y^*$) dla osób, które podjęły pracę; zero zaś ($Y = 0$) dla osób, które pracy nie podjęły i nadal są bezrobotne. Dane o wysokości płacy są dostępne tylko w przypadku prac, w których oferowana płaca spełniała oczekiwania przynajmniej jednej z osób szukających pracy. Jeżeli oferowana praca jest bardzo nisko wynagradzana, może się zdarzyć, że nie znajdą się chętni do jej podjęcia. Będzie to skutkowało tym, że ta niska płaca nie będzie obserwowana i będzie reprezentowana przez cyfrę zero (płaca była niezerowa, ale tak niska, że w ocenie bezrobotnych równa zeru).

Zmienna podlegająca nieprzypadkowej selekcji (*self-selection*, *sample selection*, *sample selectivity*). Obserwowalność wartości zmiennej Y^* zależy od warunków narzuconych na inną, skorelowaną z nią zmienną Z^*, która również może być zmienną ukrytą (nieobserwowalną) – por. rys. 1. Przyjmuje się, że obserwowalną realizacją zmiennej Z^* może być zmienna binarna, ucięta albo cenzurowana. Poniżej podano formalną definicję:

$$y_i = \begin{cases} y_i^* & \text{gdy} \quad z_i^* > a \\ c \text{ (lub } brak) & \text{gdy} \quad z_i^* \leq a \end{cases}, \text{ dla } i = 1, ..., n. \tag{3}$$

Przykładem zmiennej podlegającej selekcji jest zmienna opisująca tygodniową liczbę godzin pracy np. kobiet aktywnych zawodowo. Można obserwować liczbę godzin pracy tylko tych kobiet, które pracują, tzn. odpowiada im zaoferowana przez pracodawcę liczba godzin pracy. Kobiety te, przez swoje decyzje o podjęciu pracy, tworzą próbę kobiet pracujących. Próba ta nie jest losowa, gdyż to, czy kobieta podjęła pracę, czy nie, zależy od pewnych jej indywidualnych cech (np. wiek, miejsce zamieszkania, fakt, czy współmałżonek pracuje, wysokość wynagrodzenia itp.).



Rys. 1. a) zmienna losowa, b) zmienna ucięta,
c) zmienna cenzurowana, d) zmienna podlegająca nieprzypadkowej selekcji
Źródło: opracowanie własne.

Przede wszystkim ucięcie i cenzurowanie podlegają nieprzypadkowej selekcji.

Przy analizowaniu zmiennych ograniczonych istotne jest, że gęstość, dystrybuanta oraz momenty zmiennej ograniczonej Y różnią się od gęstości, dystrybuanty oraz momentów zmiennej Y^* (por. np. [Maddala 1983; Greene 2003]).

3. Modele tobitowe

Szczególnie wtedy, gdy na podstawie kryterium obserwowalności: ucięcia, cenzurowania lub nieprzypadkowej selekcji, zostanie odrzucony stosunkowo duży

procent obserwacji zmiennej Y^*, do opisu jej obserwowalnej realizacji Y, jako zmiennej zależnej, nie należy stosować klasycznego modelu regresji, gdyż estymatory parametrów będą obciążone. Właściwym narzędziem do modelowania zmiennej zależnej ograniczonej są modele tobitowe. Ogólnie modele te można zapisać w następującej postaci:

$$\text{a) } y_i = \begin{cases} y_i^* & \text{gdy } z_i^* > 0 \\ 0 & \text{gdy } z_i^* = 0 \end{cases}, \quad \text{b) } y_i^* = X_i\beta + u_i, \quad \text{c) } z_i^* = W_i\gamma + v_i, \qquad (4)$$

gdzie: $i = 1, ..., n$,

β, γ – wektory parametrów modelu,

X_i, W_i – wektory zmiennych objaśniających, mogące mieć ze sobą część wspólną lub całkowicie się pokrywać,

u_i, v_i – składniki losowe, niezależne, o wartości oczekiwanej równej zeru oraz stałych wariancjach odpowiednio σ_u oraz σ_v.

W praktyce najczęściej zakłada się, że składniki losowe podlegają rozkładowi normalnemu: $(u_i, v_i) \sim N(0,0; \begin{bmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{bmatrix})$, przy czym ρ jest współczynnikiem korelacji między u_i oraz v_i.

Równanie (4b) nazywane jest *równaniem regresji* opisującym zmienną Y^*. Równanie (4c) nazywane jest *równaniem selekcji* lub *kryterium obserwowalności* opisującym zmienną Z^*, od której zależy obserwowalność zmiennej Y^*. Uwzględniając sposób, w jaki zmienna Z^* dopuszcza obserwowalność wartości zmiennej Y^*, można skonstruować poszczególne modele tobitowe. Poniżej przedstawiono dwa z nich[3].

Model tobitowy typu I (standardowy model tobitowy) służy do opisu zmiennej zależnej Y podlegającej cenzurowaniu. Jest to model (4a)–(4c), w którym $Y^* \equiv Z^*$ oraz $X \equiv W$. Mechanizmy selekcji oraz wartości zmiennej Y^* zależą od identycznego zbioru zmiennych objaśniających ($X \equiv W$) i w ten sam sposób ($\beta \equiv \gamma$). Estymację parametrów modelu przeprowadza się za pomocą metody największej wiarygodności.

Model tobitowy typu II (inaczej: *model selekcji*, *model heckitowy*) pozwala na opis zmiennych Y^* i Z^* za pomocą różnych zbiorów zmiennych objaśniających (X, W) oraz w różny sposób (β, γ). Mechanizm selekcji dokonywany jest przez zmienną binarną $I_i = \begin{cases} 1 & \text{gdy } z_i^* > 0 \\ 0 & \text{gdy } z_i^* \le 0 \end{cases}$. Zakłada się możliwość istnienia korelacji między u_i oraz v_i. Estymację parametrów modelu przeprowadza się za pomocą dwustopniowej metody Heckmana lub metody największej wiarygodności.

[3] Autorka rozważała standardowy model tobitowy w pracy [Kostrzewska 2003], modele selekcji zaś w pracy [Kostrzewska 2004].

Inne modele tobitowe dopuszczają m.in. zmienną cenzurowaną do opisu równania selekcji i/lub więcej zmiennych zależnych.

Interpretacja parametrów w modelach tobitowych różnego typu nie jest natychmiastowa. Bardzo ważne jest, że cząstkowe wpływy (*marginal effects*) poszczególnych zmiennych objaśniających zazwyczaj różnią się od wartości oszacowanych parametrów. Należy je wyznaczyć, korzystając z odpowiednich formuł (por. np. [Greene 2003]).

4. Opis badania

Podstawą badań są dane pochodzące z badania aktywności ekonomicznej ludności (BAEL) dla II kwartału 2005 r., przeprowadzanego przez GUS. Próba składa się z 5414 zamężnych kobiet, aktywnych zawodowo (według definicji BAEL), dla których dostępne są także dane dotyczące współmałżonka. Wśród badanych znalazło się 1033 (19,08%) kobiet bezrobotnych oraz 4381 kobiet pracujących.

Podjęto próbę zastosowania modeli tobitowych do opisu tygodniowej liczby godzin pracy w zależności od określonych cech badanych kobiet i ich współmałżonków. Ponieważ liczbę godzin pracy można obserwować tylko w przypadku kobiet pracujących, tę zmienną uznano za podlegającą nieprzypadkowej selekcji. Tym samym rozważono dwa problemy: „czy badana kobieta pracuje, czy jest osobą bezrobotną (I_i)" oraz „ile godzin badana kobieta poświęca tygodniowo na pracę, o ile pracuje (y_i)". Przyjęto, że problemy te mogą zależeć od dwóch różnych zbiorów zmiennych objaśniających.

Jaki model należy wybrać do opisu tygodniowej liczby godzin pracy? *Model probitowy* nie wykorzystuje informacji o liczbie godzin pracy. Można go zastosować tylko w celu modelowania prawdopodobieństwa stanu „osoba pracująca" i stanu „osoba bezrobotna". *Model regresji* na podstawie wszystkich obserwacji nie uwzględnia faktu, że w próbie są osoby pracujące i bezrobotne. Inne powinno być podejście do modelowania badanego zjawiska w przypadku kobiet pracujących, inne w przypadku kobiet bezrobotnych, gdyż fakt posiadania lub nieposiadania pracy zależy od nielosowych czynników. Model regresji na podstawie danych dotyczących tylko pracujących nie uwzględnia faktu, że taka próba nie jest losowa. *Model tobitowy typu I* wykorzystuje zarówno informację, czy badana osoba pracuje, czy jest bezrobotna, jak i liczbę przepracowanych godzin. Jednak zakłada, że oba problemy zależą od tych samych czynników i w ten sam sposób. *Model tobitowy typu II* wykorzystuje wszystkie dostępne informacje. Ponadto pozwala na opis równania selekcji i równania regresji za pomocą różnych zbiorów zmiennych objaśniających oraz dopuszcza istnienie korelacji między składnikami losowymi tych równań.

Na tej podstawie wybrano model tobitowy typu II jako najbardziej adekwatny do opisu badanego zjawiska. Ostateczny zbiór zmiennych objaśniających ujętych w tym modelu zestawiono w tab. 1.

Tabela 1. Ostateczny zbiór zmiennych objaśniających

KLM	Klasa miejscowości; zmienna binarna (wartość 1 – wieś, wartość 0 – miasto)
PW	Poziom wykształcenia. Bazując na tej zmiennej, utworzono pięć zmiennych binarnych: PW_{01} – wykształcenie wyższe; PW_{02} – policealne; PW_{03} – średnie ogólnokształcące; PW_{04} – zasadnicze zawodowe; PW_{05} – gimnazjalne i niżej
RT	Sytuacja na rynku pracy przed rokiem. Bazując na tej zmiennej, utworzono trzy zmienne binarne: RT_{01} – pracująca; RT_{02} – bezrobotna; RT_{03} – bierna zawodowo
RODZP	Rodzaj wykonywanej pracy (tylko dla pracujących). Bazując na tej zmiennej, utworzono trzy zmienne binarne: $RODZP_{01}$ – pracujący na własny rachunek; $RODZP_{02}$ – pracownik najemny; $RODZP_{03}$ – pomagający członek rodziny
mLGP	Tygodniowa liczba godzin pracy męża
PD	Praca dodatkowa; zmienna binarna (wartość 1 – posiada pracę dodatkową, wartość 0 – nie)
mPD	Praca dodatkowa męża; zmienna binarna (jw.)

Źródło: opracowanie własne.

Tabela 2. Wyniki estymacji parametrów modeli: regresji, tobitowego typu I oraz tobitowego typu II

		Model regresji (n = 5414)	Model regresji (n = 4381)	Model tobitowy I		Model tobitowy II		
	Zmienna	–	–	–	–	parametr	wpływ pośredni	wpływ całkowity
Równanie selekcji	Wyraz wolny	–	–	–	–	–0,2909	–0,3257	–0,3257
	KLM	–	–	–	–	0,2584	0,2893	–0,8031
	PW_{02}	–	–	–	–	–0,5545	–0,6208	5,0861
	PW_{03}	–	–	–	–	–0,7060	–0,7904	4,9118
	PW_{04}	–	–	–	–	–0,9113	–1,0203	5,1504
	PW_{05}	–	–	–	–	–0,9244	–1,0350	5,1301
	RT_{01}	–	–	–	–	2,7087	3,0328	–2,2899*
	RT_{03}	–	–	–	–	0,5217	0,5841	–4,5196
	Zmienna	parametr	parametr	parametr	wpływ	parametr	wpływ bezpośredni	wpływ całkowity
Równanie regresji	Wyraz wolny	–12,5264	13,3183	–33,8014	–33,3604	19,1023	19,1023	19,1023
	KLM	1,7328	–0,9123	2,5731	2,5395	–1,0924	–1,0924	–0,8031
	PW_{02}	5,3943	5,4927	5,5905	5,5176	5,7069	5,7069	5,0861
	PW_{03}	5,3577	5,3554	5,5376	5,4653	5,7023	5,7023	4,9118
	PW_{04}	5,6885	5,6409	5,8213	5,7453	6,1707	6,1707	5,1504
	PW_{05}	7,1123	5,6280	8,0007	7,8963	6,1651	6,1651	5,1301
	RT_{01}	15,0637	0,4288*	23,6556	23,3469	–5,3227	–5,3227	–2,2899*
	RT_{03}	2,7752	–3,1037	7,8861	7,7832	–5,1037	–5,1037	–4,5196
	$RODZP_{01}$	17,5519	5,4536	21,9609	21,6744	5,4782	5,4782	5,4782
	$RODZP_{02}$	17,5342	3,6758	23,7843	23,4740	3,6808	3,6808	3,6808
	mLGP	0,3309	0,3902	0,4852	0,4789	0,3910	0,3910	0,3910
	PD	12,3900	13,0454	11,5461	11,3955	13,0964	13,0964	13,0964
	mPD	–5,6757	–6,3574	–7,2316	–7,1373	–6,3749	–6,3749	–6,3749
	$\hat{\sigma}_u$	11,4426	10,5973	12,9359	–	10,7186	–	–
	$\hat{\rho}$	–	–	–	–	–0,3977	–	–
	$R^2_{dekomp.}$	0,6408	0,2712	0,6628	–	0,2725	–	–

Uwaga: (*) oznacza, że parametr lub wpływ cząstkowy jest nieistotny statystycznie (p-value > 0,1).

Źródło: obliczenia własne z użyciem oprogramowania LimDep.

W tab. 2 zestawiono wyniki oszacowania parametrów modelu tobitowego typu II. Dla porównania w tabeli tej zamieszczono również wyniki estymacji parametrów modelu tobitowego typu I, a także modelu regresji na podstawie całej próby (n = 5414) i tylko osób pracujących (n = 4381). Dla modeli tobitowych wyznaczono także wpływy cząstkowe poszczególnych zmiennych objaśniających na tygodniową liczbę godzin pracy zamężnych kobiet. W przypadku modelu tobitowego II wpływy te podzielono na dwa rodzaje. Wpływy pośrednie, tzn. przez równanie selekcji, to wpływy zmiennej objaśniającej na liczbę godzin pracy przez jej wpływ na prawdopodobieństwo, że dana osoba jest pracująca. Wpływy bezpośrednie to wpływy na liczbę godzin pracy bezpośrednio przez równanie regresji. Sumą wpływów pośrednich i bezpośrednich są wpływy całkowite. Aby porównać wartości oszacowań parametrów modeli, należy porównać parametry modeli regresji z wpływami w modelu tobitowym typu I oraz z wpływami całkowitymi w modelu tobitowym typu II. Odpowiednie kolumny w tabeli oznaczono szarym kolorem.

Wyniki zawarte w tab. 2 w pewnym stopniu obrazują, jakie mogą być skutki zastosowania klasycznego modelu regresji do opisu zmiennej zależnej ograniczonej. W zaprezentowanym przykładzie oszacowania wpływów poszczególnych zmiennych objaśniających w modelu tobitowym typu II są inne niż oszacowania parametrów modelu regresji wielorakiej, chociaż, ze względu na słabe dopasowanie modelu do danych[4], nie są bardzo dalekie. Porównując oszacowania wpływów w modelu tobitowym typu II oraz parametrów modeli regresji, można zauważyć, że mniejsze rozbieżności wystąpiły, gdy rozważono model regresji na podstawie obserwacji dotyczących tylko pracujących. Różnice widać przede wszystkim w przypadku tych zmiennych objaśniających, które występują równocześnie w równaniu selekcji i równaniu regresji modelu tobitowego typu II. Wynika to właśnie z faktu, że model regresji nie uwzględnia problemu: „czy badana kobieta pracuje, czy jest osobą bezrobotną".

W praktyce różnice w oszacowaniach są tym większe, im większy jest odsetek obserwacji niespełniających kryterium obserwowalności, a także im większa jest korelacja między składnikami losowymi równania regresji (4b) i równania selekcji (4c). W badaniach empirycznych należy mieć na uwadze, czy próba będąca przedmiotem analizy jest rzeczywiście próbą losową, a zatem czy można do niej zastosować klasyczny model regresji. Może się okazać, że jest to próba ucięta, cenzurowana lub podlegająca nieprzypadkowej selekcji, wówczas należy sięgnąć po modele tobitowe.

[4] Na podstawie dostępnych informacji nie udało się oszacować parametrów modelu tak, by wystarczająco dobrze opisywał tygodniową liczbę godzin pracy. Głównym powodem może być brak zmiennych objaśniających badane zjawisko nieuwzględnionych w ankietach BAEL (np. takich jak: liczba dzieci w gospodarstwie domowym; całkowity staż pracy; zagregowana zmienna dotycząca wynagrodzeń o bardzo dużej liczbie brakujących danych), ale także przyjęte założenie rozkładu normalnego składnika losowego. Być może lepszy model można będzie otrzymać przy zastosowaniu semiparametrycznych metod estymacji, co będzie przedmiotem dalszych badań autorki.

5. Kierunki dalszych badań

Przeprowadzone badanie miało na celu wskazanie modeli tobitowych jako odpowiedniego narzędzia do opisu zmiennej zależnej ograniczonej. Zaprezentowany przykład zastosowania uwidacznia, że oszacowania parametrów (wpływów cząstkowych) modelu tobitowego typu I oraz II są różne od oszacowań parametrów modelu regresji wielorakiej.

Dalsze prace badawcze mogą objąć rozszerzenie analizy m.in. przez uwzględnienie osób biernych zawodowo, powtórzenie wnioskowania dla prób odnoszących się do I, III lub IV kwartału 2005 r., rozpatrzenie modeli przy założeniu innego rozkładu składnika losowego, a także z wykorzystaniem metod semiparametrycznych, co może doprowadzić do uzyskania lepszego opisu tygodniowej liczby godzin pracy zamężnych kobiet.

Literatura

Greene W. (2003), *Econometric Analysis*, 5th ed., Prentice-Hall, Inc., New Jersey.

Kostrzewska J. (2004), *Model regresji dla danych dobieranych do próby według nielosowego kryterium*, Zeszyty Naukowe Akademii Ekonomicznej w Krakowie nr 666, AE, Kraków, s. 111-117.

Kostrzewska J. (2003), *Model tobitowy jako szczególny przypadek cenzurowanego modelu regresji*, [w:] *Przestrzenno-czasowe modelowanie i prognozowanie zjawisk gospodarczych*, red. A. Zeliaś, AE, Kraków, s. 397-404.

Maddala G.S. (1983), *Limited-dependent and Qualitative Variables in Econometrics*, Cambridge University Press, Cambridge.

Tobin J. (1958), *Estimation of Relationships for Limited Dependent Variables*, „Econometrica", vol. 26, s. 24-36.

Verbeek M. (2001), *A Guide to Modern Econometrics*, John Wiley & Sons, Chichester.

AN APPLICATION OF TOBIT MODELS TO A DESCRIPTION OF A NUMBER OF HOURS WORKED PER WEEK

Summary

The aim of the paper is to present some types of tobit models and their application on the Polish labour market. There is made a trial of a description of a number of hours worked per week by married women in dependence on some characteristics which influence on being employed or unemployed. The results of an estimation of tobit models type I and II are compared with results of an estimation of classical regression models. The analysis is based on a data set from the labour force survey (BAEL) conducted by the CSO in Poland.