DYNAMICZNE MODELE EKONOMETRYCZNE
IX Ogólnopolskie Seminarium Naukowe, 6-8 września 2005 w Toruniu
Katedra Ekonometrii i Statystyki, Uniwersytet Mikołaja Kopernika w Toruniu

Joanna Małgorzata Gwiazda
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie

Zastosowanie modeli hazardu do szacowania długości czasu pozostawania bez pracy w Niemczech i w Polsce

1. Wstęp

W pracy dokonamy estymacji modeli hazardu mających na celu określenie wpływu takich wielkości jak wiek, płeć, wykształcenie, czy narodowość na indywidualną długość czasu pozostawania bez pracy w Niemczech i w Polsce.

Bazę danych stanowi niemiecki Panel Socjoekonomiczny (SOEP) oraz polskie Badanie Aktywności Ekonomicznej Ludności (BAEL).

Zmienną zależną będzie prawdopodobieństwo przejścia ze stanu pozostawania bezrobotnym w stan zatrudnienia. Analizę czynników wywierających wpływ na ową zmienną przeprowadzili Uhlendorff (2003) oraz Hujer, Schneider (1996), my zaś stawiamy następujące hipotezy, których zasadność zweryfikujemy za pomocą modeli:

1. Starszy wiek zmniejsza prawdopodobieństwo ponownego zatrudnienia.
2. Kobiety są zatrudniane z mniejszym prawdopodobieństwem niż mężczyźni.
3. Dłuższa edukacja zwiększa prawdopodobieństwo podjęcia zatrudnienia.
4. Bezrobotni obcego pochodzenia podejmują zatrudnienie z mniejszym prawdopodobieństwem.

2. Wybór i opis właściwej metody analizy

Dane empiryczne dla zmiennej zależnej mogą przyjmować wyłącznie dodatnie wartości – ujemne czasy trwania nie istnieją! Fakt ten oraz to, że długość czasu trwania zjawiska często tylko częściowo może być obserwowana i mierzona (*censoring problem*), wyklucza możliwość stosowania klasycznych mo-

deli regresji. Doprowadziło to do rozwoju tzw. analizy czasu przeżycia (*Survival Analysis*) [1].

Celem *Survival Analysis* jest wyrażenie zaobserwowanego rozkładu długości czasu trwania zjawiska za pomocą tzw. funkcji hazardu. *Survival Analysis* znajduje zastosowanie w różnych dziedzinach nauk technicznych i społecznych. *Hazard models* są budowane wszędzie tam, gdzie celem jest prognozowanie momentu, w którym dojdzie do pewnego wydarzenia. Prognoza długości czasu pozostawania przez bezrobotnych bez pracy, aż do chwili podjęcia przez nich zatrudnienia, stanowi typowy przykład zastosowań[2].

W ramach *Survival Analysis* rozwinięto całą gamę modeli, które pomiędzy sobą różnią się założeniami dotyczącymi rozkładu indywidualnego czasu wystąpienia wydarzenia *T*. Możliwe są następujące formy prezentujące rozkład prawdopodobieństwa dla czasu *T*.

Funkcja rozkładu *F(t)* dostarcza prawdopodobieństwo tego, że moment wystąpienia wydarzenia (np. znalezienie pracy) leży przed albo dokładnie w wybranym punkcie czasu *t*; $F(t) = Pr[T \le t]$. Symbolem *f(t)* będziemy oznaczać funkcję gęstości.

W zastosowaniach *Survival Analysis* dużą uwagę przywiązuje się zazwyczaj możliwościom przetrwania procesu ponad pewien wybrany punkt czasu. Prawdopodobieństwo przeżycia chwili czasu *t* wyraża *Survival Function S(t),* $S(t) = Pr[T > t] = 1 - F(t)$. Funkcja ta może przyjmować wartości od *0* do *1*. Graficzny przebieg tej funkcji zależy od właściwości obserwowanego procesu, z reguły jest to jednak krzywa opadająca w dół.

Najczęściej stosowanym przedstawieniem rozkładu czasu trwania jest jednak funkcja hazardu *h(t)* (*conditional failure function, hazard function*). Szacuje ona „prawdopodobieństwo" lub raczej bezpośrednie ryzyko tego, że wydarzenie nastąpi w przedziale czasowym pomiędzy *t* i *t+dt*, pod warunkiem, że do danego momentu zajście jego nie nastąpiło.

$$h(t) = \frac{f(t)}{S(t)} = \lim_{dt \to 0} \frac{Pr[t < T < t + dt \mid T > t]}{dt}$$

Stopy hazardu – wartości funkcji hazardu – nie są prawdopodobieństwami we właściwym tego słowa znaczeniu, lecz ukrytymi zmiennymi, które można scharakteryzować jako intensywności przejścia od jednego stanu do drugiego. Im wyższa jest w tym wypadku wartość funkcji, tym szybciej przeciętnie ma miejsce przejście ze stanu A do stanu B.

[1] Wprowadzenie do omawianej tematyki oferują: Kalbfleisch, Prentice (1980), Cox, Oakes (1984), Heckman, Singer (1984), Blossfeld i inni (1986), Kiefer (1988), Lancaster (1990), Green (2000).

[2] Por. Devine, Kiefer (1991), Lancaster (1979).

3. Modele hazardu

Zazwyczaj modele hazardu uwzględniają nie tylko aktualny czas trwania zjawiska jako istotną determinantę dla prawdopodobieństwa zajścia zdarzenia, ale również inne wielkości. Wśród modeli czasu przeżycia, które umożliwiają oszacowanie wpływu różnych determinant, wyróżnia się następujące modele parametryczne (*parametric hazard models*): modele proporcjonalnego hazardu PH (*proportional hazard models*) oraz *accelerated failure-time models* AFT.

W *proportional hazard models* dodatkowe zmienne objaśniające są zdefiniowane jako funkcja wektora zmiennych X, która działa w sposób multiplikatywny na podstawową funkcję hazardu $h_0(t)$ (*baseline hazard*):

$$h(t|X) = h_0(t)g_0(X) = h_0(t)exp(X\beta).$$

Po zlogarytmowaniu otrzymamy $log\ h(t|X) = \alpha(t) + \beta_1 x_1 + \ldots + \beta_k x_k$.

Własności funkcji hazardu zmieniają się więc w sposób proporcjonalny do wpływu zmiennych objaśniających.

Do klasy *Proportional Hazard* należy cały szereg modeli, które wykazują istotne różnice jeśli chodzi o założenia dotyczące rozkładu *baseline hazard*. Graficzny kształt funkcji hazardu dla konkretnych procesów trwania w czasie może przybierać różne formy.

W modelach *accelerated failure-time* AFT dokonuje się parametryzowania zmiennej τ_i, $\tau_i = t_i exp(-X\beta)$. Człon $exp(-X\beta)$ nosi miano tzw. parametru przyspieszającego (*acceleration parameter*). Jeśli $exp(-X\beta) = 1$, to $\tau_i = t_i$ i czas „biegnie normalnie". W wypadku $exp(-X\beta) > 1$ czas jest przyspieszany; dla danej obserwacji zdarzenie wystąpi szybciej. Natomiast $exp(-X\beta) < 1$ oznacza, iż czas jest spowolniony. W modelach AFT uwaga koncentruje się na zmiennej t_i, $t_i = exp(X\beta)\tau_i$, a dokładniej na $E[ln(t_i)|x_i]$ dla różnych x_i.

Poniżej przedstawione zostaną oszacowane modele opisujące czas bycia bezrobotnym prezentujące dwa powyższe podejścia.

4. Zakres czasowy badania oraz obiekty badań

Dane użyte w analizie niemieckiego rynku pracy zaczerpnięto z socjoekonomicznego panelu SOEP (*The German Socio Economic Panel Study*). Budowę panelu SOEP zapoczątkowano w 1984r. na terenie BRD (4528 gospodarstw), a w 1990r. poszerzono go o próbę z byłego DDR (2179 gospodarstw). Na pytania dotyczące wydarzeń z ubiegłego roku (np. przebieg pracy), jak również sytuacji bieżącej (np. ilość dzieci) lub poglądów, odpowiadają co najmniej 16-letni respondenci. Bazę danych do niniejszej analizy stanowi siedem „fal" panelu SOEP obejmujących okres od I.1996 do XII.2000 (dane miesięczne).

Informacje na temat stanów, w których znaleźli się ankietowani (np. zatrudnienie, kształcenie, bezrobocie, emerytura) są archiwizowane w postaci tzw.

epizodów. Długość pojedynczego epizodu wyznacza ilość sąsiadujących ze sobą miesięcy, w trakcie których osoba znajdowała się w danym stanie.

W analizie zostały uwzględnione osoby w wieku od 16 do 59 lat, które w trakcie badania wykazały co najmniej jeden epizod oznaczający bycie bezrobotnym (przynajmniej w jednym miesiącu były zameldowane jako bezrobotne). Z analizy wyłączono dane dotyczące osób bezrobotnych już przed 1.I.1996 (*left censoring*). Natomiast, gdy w rzeczywistości rozgrywający się proces był dłuższy niż ten obserwowany w czasie trwania badania, uznano, że mamy do czynienia z *right censoring*. W wyniku opisanych kryteriów selekcji wyłoniono próbę 1455 osób. Uszeregowane długości trwania ostatnich epizodów pozostawania bezrobotnym dla poszczególnych osób tworzą zmienną *bezrob*. Ze wstępnej analizy wynika, że średni czas trwania w bezrobociu wyniósł w Niemczech 10,918 miesiąca. Za zmienne objaśniające przyjęto: *rokur* – rok urodzenia, *pl* – płeć (1 = mężczyzna, 0 = kobieta), *latanauk* – czas trwania edukacji w latach, *narod* – narodowość (1 = niemiecka, 0 = inna).

Do estymacji modeli hazardu dotyczących długości pozostawania bezrobotnym w Polsce użyto danych pochodzących z reprezentacyjnego Badania Aktywności Ekonomicznej Ludności (BAEL) przeprowadzonego w 4. kwartale 2000 roku. Próba BAEL jest panelem rotacyjnym. Wylosowane do próby gospodarstwo domowe jest badane czterokrotnie: w 1. i 2. kwartale badania, oraz w 5. i 6. kwartale. Przy każdym nowym badaniu jedna czwarta próby jest na nowo losowana. W rezultacie na podstawie BAEL można skonstruować panele dla gospodarstw z ilością przekrojów równą tylko 4. Nie da się więc zbadać problemu długości trwania w stanie bezrobocia, opierając się tylko na obserwacjach dotyczących zmian w czasie wybranych charakterystyk respondenta. Jednak ankieta BAEL-a zawiera również pytania retrospektywne, dotyczące przeszłości zawodowej. Na ich podstawie można ustalić, jak długo aktualnie bezrobotna osoba szuka pracy (ilość miesięcy), lub – jeśli ankietowana osoba była bezrobotna w przeszłości – jak długo szukała pracy.

Pełną próbę BAEL ($\approx$ 55000 obserwacji) ograniczono w badaniu do podpróby (3951 obserwacji) dotyczącej mieszkańców województwa mazowieckiego. Dokonując dalszej selekcji danych wyłoniono próbę 363 aktywnie poszukujących pracy osób w wieku od 16 do 59 lat, które były zarejestrowane w urzędzie pracy jako bezrobotne (170 osób) lub pozostawały bezrobotne przed rozpoczęciem aktualnie wykonywanej pracy (193 osoby). Długości trwania w stanie bezrobocia (w miesiącach) dla poszczególnych osób tworzą zmienną *bezrob*. Średni czas trwania w bezrobociu wyniósł w województwie mazowieckim 13,017 miesiąca. Zmienne, za pomocą których opisano długość pozostawania w stanie bezrobocia, są następujące: *wiek* – wiek respondenta w latach (wiek w chwili podjęcia pracy dla danych cenzurowanych; dla pozostałych danych wiek w chwili ankietowania), *pl* – płeć (1 = mężczyzna, 0 = kobieta), *wyksz* – wykształcenie (1 = wyższe, 2 = policealne, 3 = średnie zawodowe, 4 = średnie ogólnokształcące, 5 = zasadnicze zawodowe, 6 = podstawowe, 7 = niepełne podstawowe).

5. Konstrukcja i estymacja modeli hazardu

Parametry modeli oszacowano metodą największej wiarygodności. Ocena dobroci modeli obejmuje ocenę wiarygodności uzyskanych wyników. Miarą wiarygodności jest ujemny dwukrotny logarytm funkcji wiarygodności *(-2lnL)*, który jest tym mniejszy, im większa jest wiarygodność wyników. Poprawność modelu oceniana jest również w tzw. *Likelihood-Ratio-Test* za pomocą różnicy między ujemnym dwukrotnym logarytmem wiarygodności dla oszacowanego modelu oraz dla modelu zawierającego tylko stałą: $\chi^2 \approx$ *-2ln L(zmienne,stała)-[-2lnL(stała)]*. Dodatkowo oceniana jest istotność poszczególnych współczynników regresji. Obliczenia wykonano w programie STATA.

5.1. Model PH: Regresja Weibulla

Założony w tym wypadku rozkład Weibulla dla stopy hazardu znajduje zastosowanie w modelowaniu danych z monotonicznymi stopami hazardu, które albo rosną, albo maleją z biegiem czasu.

Baseline hazard definiuje się następująco: $h_0(t) = pt^{p-1}$, gdzie p to tzw. *shape parameter*, estymowany na podstawie danych. Funkcja hazardu jest postaci: $h(t) = pt^{p-1} exp(X\beta)$, a *survival function*: $S(t) = exp(-exp(X\beta)t^p)$.

Uzyskane wyniki estymacji modelu dla danych pochodzących z Niemiec, w którym funkcja hazardu jest opisywana rokiem urodzenia, płcią, ilością lat nauki oraz narodowością (X = (*rokur*, *pl*, *latanauk*, *narod*)), są następujące:

Tabela 1. Wyniki estymacji modelu regresji Weibulla opartego na danych SOEP

Number of obs = 1455			LR chi2(4) = 347.52		
Log likelihood = -1940.7108			Prob > chi2 = 0.0000		
t	Coef.	Std.Err.	z	P>z	[95% Conf. Interval]
rokur	0.03887	0.00248	15.67	0.000	0.03401 0.04373
pl	0.12867	0.06141	2.10	0.036	0.00831 0.24903
latanauk	0.12220	0.01274	9.59	0.000	0.09724 0.14717
narod	0.28203	0.09681	2.91	0.004	0.09228 0.47177
cons	-80.64718	4.90190	-16.45	0.000	-90.25472 -71.03964
p	1.05964	0.02432			1.01303 1.10838

Źródło: obliczenia własne.

Interpretując oceny parametrów modelu, możemy stwierdzić, że niższy wiek respondenta o 1 rok prowadzi do 4% wzrostu "ryzyka" znalezienia pracy ($exp(0{,}039) = 1{,}04$). "Ryzyko" znalezienia pracy przez mężczyznę jest o 13,7% większe niż przez kobietę ($exp(0{,}129) = 1{,}137$). Zwiększenie czasu nauki o 1 rok prowadzi do 13% wzrostu "ryzyka" znalezienia pracy ($exp(0{,}122) = 1{,}13$). "Ryzyko" znalezienia pracy przez Niemca jest o 32,6% większe niż przez osobę

o innej narodowości ($exp(0{,}282) = 1{,}326$). Oszacowany parametr $p>1$ wskazuje na rosnącą w czasie stopę hazardu, której uzasadnieniem może być większa skłonność do szukania pracy przez bezrobotnych, jeśli pomoc dla nich ze strony państwa (np. zasiłek) wyczerpie się.

Wyniki estymacji modelu na podstawie danych z bazy BAEL (dane dla województwa mazowieckiego) są następujące:

Tabela 2. Wyniki estymacji modelu regresji Weibulla opartego na danych BAEL

Number of obs = 363			LR chi2(4) = 27.75			
Log likelihood = -369.03173			Prob > chi2 = 0.0000			
T	Coef.	Std.Err.	z	P>z	[95% Conf.	Interval]
wiek	-0.02362	0.00768	-3.08	0.002	-0.03866	-0.00858
pl	0.51011	0.15061	3.39	0.001	0.21491	0.80531
wyksz	-0.11741	0.05520	-2.13	0.033	-0.22559	-0.00926
cons	-3.09608	0.35231	-8.79	0.000	-3.78660	-2.40556
p	1.31718	0.07320			1.18124	1.46876

Źródło: obliczenia własne.

Oceny parametrów modelu wskazują na to, że niższy wiek respondenta o 1 rok prowadzi do 2,4% wzrostu "ryzyka" znalezienia pracy ($exp(-0{,}024) = 0{,}976$). Istotną różnicę w porównaniu z oszacowanym modelem na podstawie niemieckich danych stanowi fakt, iż tym razem "ryzyko" znalezienia pracy przez mężczyznę jest aż o 66,6% większe niż przez kobietę ($exp(0{,}51) = 1{,}666$). Osiągnięcie wyższego stopnia wykształcenia prowadzi do 11,1% wzrostu "ryzyka" znalezienia pracy ($exp(-0{,}117) = 0{,}889$).

5.2. Log-normalny model AFT

W modelu przyjmuje się, że logarytm naturalny dla czasu trwania, $ln\, t_i = X\beta + ln\, \tau_i$, ma rozkład normalny. *Survival function* oraz funkcja gęstości są następujące:

$$S(t) = 1 - \Phi\left\{\frac{ln(t) - X\beta}{\sigma}\right\}, \qquad f(t) = \frac{1}{t\sigma\sqrt{2\pi}} exp\left[\frac{-1}{2\sigma^2}\{ln(t) - X\beta\}^2\right],$$

gdzie $\Phi(z)$ jest funkcją rozkładu dla standardowego rozkładu normalnego.

Ujemne wartości ocen parametrów strukturalnych w modelu AFT dla Niemiec oznaczają, że czas jest „przyspieszaczem". Wzrost wartości zmiennej objaśniającej o 1 jednostkę wywołuje przyspieszenie zajścia zdarzenia „zakończenie trwania w bezrobociu".

Tabela 3. Wyniki estymacji log-normalnego modelu AFT opartego na danych SOEP

Number of obs = 1455			LR chi2(4) = 328.55		
Log likelihood = -1887.6586			Prob > chi2 = 0.0000		
t	Coef.	Std.Err.	z	P>z	[95% Conf. Interval]
rokur	-0.03786	0.00241	-15.68	0.000	-0.04259 -0.03313
pl	-0.15447	0.06157	-2.51	0.012	-0.27515 -0.03379
latanauk	-0.11020	0.01320	-8.35	0.000	-0.13608 -0.08433
narod	-0.33841	0.09703	-3.49	0.000	-0.52859 -0.14823
cons	77.94248	4.72727	16.49	0.000	68.67720 87.20775
sigma	1.10115	0.02401			1.05509 1.14922

Źródło: obliczenia własne.

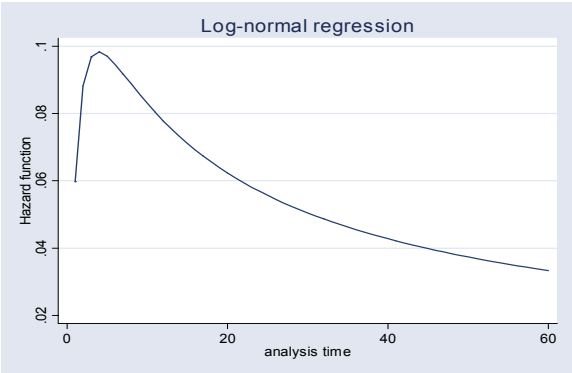
Młodszy wiek, płeć męska oraz wyższy stopień wykształcenia wywołują również i w województwie mazowieckim spadek oczekiwanego czasu, jaki upłynie do chwili zmiany stanu z bezrobocia w jakiś inny.

Tabela 4. Wyniki estymacji log-normalnego modelu AFT opartego na danych BAEL

Number of obs = 363			LR chi2(4) = 16.44		
Log likelihood = -364.71719			Prob > chi2 = 0.0009		
t	Coef.	Std.Err.	z	P>z	[95% Conf. Interval]
wiek	0.01457	0.00578	2.52	0.012	0.00324 0.02590
pl	-0.24956	0.12324	-2.03	0.043	-0.49111 -0.00802
wyksz	0.11161	0.04754	2.35	0.019	0.01843 0.20479
cons	1.93517	0.26095	7.42	0.000	1.42371 2.44663
sigma	1.00823	0.05248			0.91044 1.11652

Źródło: obliczenia własne.

Stosowanie rozkładu typu log-normalnego jest zalecane w wypadku, gdy dane wskazują na niemonotoniczne stopy hazardu.



Wykres 1. Funkcja hazardu w wypadku rozkładu log-normalnego.

Powyższy wykres oszacowanej na podstawie danych SOEP funkcji hazardu wskazuje na następujące zależności: a) podjęcie pracy jest najbardziej prawdo-

podobne w krótkim czasie po zaprzestaniu ostatniego zatrudnienia, b) szanse na podjęcie pracy stają się tym mniejsze, im dłużej jest się bezrobotnym (jest to tzw. „lagged duration dependence“ opisana w Heckman, Borjas (1980)).

6. Wnioski

Celem dokonanej w pracy estymacji modeli hazardu było określenie wpływu takich wielkości jak wiek, płeć, poziom wykształcenia oraz narodowość na indywidualną długość czasu pozostawania bez pracy. Uzyskane wyniki - zarówno dla obywateli niemieckich, jak i mieszkańców polskiego województwa mazowieckiego - wskazują na to, że młody wiek i wysoki poziom wykształcenia istotnie podwyższają stopę hazardu określającą prawdopodobieństwo ”wyjścia” ze stanu bezrobocia. “Ryzyko” znalezienia pracy przez bezrobotną kobietę jest zazwyczaj niższe niż przez mężczyznę.

Otrzymane rezultaty potwierdzają, że modele hazardu mogą być odpowiednimi narzędziami do analizy czasu pozostawania w stanie bezrobocia.

Literatura

Blossfeld, H.-P., Hamerle, A., Mayer, K. (1986), *Ereignisanalyse. Statistische Theorie und Anwendung in den Wirtschafts- und Sozialwissenschaften*, Frankfurt/Main.

Cox, D., Oakes, D. (1984), *Analysis of Survival Data*, London.

Devine, T., Kiefer N. (1991), *Empirical Labor Economics*, New York, Oxford.

Green, W. (2000), *Econometric Analysis*, New York.

Heckman, J., Borjas G. (1980), Does Unemployment Cause Future Unemployment? Definitions, Questions and Answers from a Continuous Time Model of Heterogenity and State Dependence, *Economica*, 47, 247–283.

Heckman, J., Singer B. (1984), Econometric duration analysis, *Journal of Econometrics*, 24, 63–132.

Hujer, R., Schneider H. (1996), Institutionelle und strukturelle Determinanten der Arbeitslosigkeit in Westdeutschland. Eine mikroökonometrische Analyse mit Paneldaten, w: B. Gahlen, H. Hesse, H. Ramser (red.), *Arbeitslosigkeit und Möglichkeiten ihrer Überwindung*, Tübingen, 53–76.

Kalbfleisch, J., Prentice R. (1980), *The Statistical Analysis of Failure Time Data*, New York.

Kiefer, N. (1988), Economic Duration Data and Hazard Functions, *Journal of Economic Literature*, 26, 646–679.

Lancaster, T. (1979), Econometric Methods for the Duration of Unemployment, *Econometrica*, 47, 939–956.

Lancaster, T. (1990), *The Econometric Analysis of Transition Data*, Cambridge, New York, Melbourne.

Uhlendorff, A. (2003), *Der Einfluss von Persönlichkeitseigenschaften und sozialen Ressourcen auf die Arbeitslosigkeitsdauer*, Berlin.