

Model tobitowy

- Modele tobitowy używany jest w dwóch różnych kontekstach:
 - cenzurowania danych (*censored data*)
 - rozwiązań brzegowych (*corner solutions*)
- **Dane ocenzone:**
 - dane dla których znamy wartość zmiennej poniżej danego progu a powyżej tego progu wiemy jedynie, że zmienna jest większa niż ten próg

Przykład Udało nam się uzyskać dane dotyczące charakterystyk demograficzno społecznych wszystkich pracowników pewnej firmy. Dla płac otrzymaliśmy dane dotyczące jedynie szeregowych pracowników, w odniesieniu do osób na kierowniczych stanowiskach wiemy jedynie, że zarabiają więcej niż XXX.

- **Rozwiązania brzegowe**

- w próbie obserwujemy dodatnie wartości zmiennej zależnej oraz znaczącą liczbę zer (np. wydatki na dobro trwałego użytku, wydatki na ubezpieczenie, wydatki na badania i rozwój)
 - zara są wynikiem racjonalnych decyzji podmiotów ekonomicznych
 - wynik optymalizacji funkcji celu, który wypada na brzegu zbioru możliwych decyzji (a więc np. wydatki równe zeru) nazywamy rozwiązaniem brzegowym
- W przypadku takiego zbioru danych zastosowanie liniowej bądź nieliniowej regresji da nieprawidłowe wyniki, ponieważ:
 - w regresji liniowej $y = x\beta + \varepsilon$ nie ma sposobu by zagwarantować, by wszystkie wartości dopasowane były dodatnie
 - za pomocą regresji $\ln y = x\beta + \varepsilon$ nie da się modelować obserwacji $y = 0$
 - w regresji nieliniowej $y = \exp(x\beta) + \varepsilon$, zazwyczaj wystąpi heteroskedastyczność

- w żadnej z tych regresji nie da się oszacować dwóch ważnych własności y :
 - * $P(y > 0 | x)$ - prawdopodobieństwo dodatnich y
 - * $E(y | x, y > 0)$ - warunkowej wartości oczekiwanej y jeśli $y > 0$ (nie jest rozwiązaniem brzegowym)
- Ważne jest zauważenie, że choć dane otrzymane w wyniku ocenzonego i dane związane z rozwiązaniami brzegowymi wyglądają identycznie, to jednak są różnią się sposobem powstawania. W pierwszym przypadku obserwujemy prawdziwe wartości zmiennej zależnej jedynie dla części obserwacji, w drugim przypadku obserwujemy prawdziwe wyniki decyzji dla wszystkich obserwacji.
- Model tobitowy definiuje się w sposób następujący:

$$y^* = x\beta + \varepsilon$$

$$\varepsilon \sim N(0, 1)$$

$$y = \begin{cases} 0 & \text{dla } y^* \leq 0 \\ y^* & \text{dla } y^* > 0 \end{cases}$$

gdzie y^* nazywamy zmienną ukrytą (*latent variable*)

- Uwaga: model tobitowy można łatwo uogolnić na przypadek, w którym miejsce ucięcia przypada w innym punkcie niż zero.
- Model tobitowy szacujemy za pomocą *MNW*. Funkcja gęstości dla poszczególnych obserwacji ma postać:

$$f(y_i | \mathbf{x}_i) = \begin{cases} 1 - \Phi(\mathbf{x}_i \boldsymbol{\beta} / \sigma) & \text{dla } y_i = 0 \\ \frac{1}{\sigma} \phi\left(\frac{y_i - \mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right) & \text{dla } y_i > 0 \end{cases}$$

$$= \{1 - \Phi(\mathbf{x}_i \boldsymbol{\beta} / \sigma)\}^{1[y_i=0]} \left[\frac{1}{\sigma} \phi\left(\frac{y_i - \mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right) \right]^{1[y_i>0]}$$

- W zależności od przypadku różne interesować nas będą różne własności tego modelu:
 - dla zmiennych ocenzonego będącej nas interesować $E(y^*|x) = x\beta$ (model dla zmiennej ukrytej)
 - dla zmiennej związanej z rozwiązaniami brzegowymi, interesuje nas zazwyczaj $P(y > 0|x)$ lub $E(y|x, y > 0)$, ponieważ zmienna ukryta w tym kontekście nie ma zazwyczaj interpretacji
- Zmienna y^* powinna być z grubsza ciągła, homoskedastyczna i mieć rozkład normalny - aby uzyskać rozkład zbliżony do normalnego można zastosować przekształcenie logarytmiczne do $y_i > 0$

- Można pokazać, że:

$$\begin{aligned}P(y > 0 | x) &= \Phi(x\beta / \sigma) \\E(y | x, y > 0) &= x\beta + \sigma \lambda(x\beta / \sigma) \\E(y | x) &= [x\beta + \sigma \phi(x\beta / \sigma) / \Phi(x\beta / \sigma)] \Phi(x\beta / \sigma)\end{aligned}$$

gdzie $\lambda(c) = \frac{\phi(c)}{\Phi(c)}$ jest nazywane odwrotnością współczynnika Millsa (*inverse Mills ratio*).

- Efekty cząstkowe dla poszczególnych tych wartości są równe:

$$\begin{aligned}\frac{\partial P(y > 0 | x)}{\partial x} &= \phi(x\beta / \sigma) \beta / \sigma \\ \frac{\partial E(y | x, y > 0)}{\partial x} &= \{1 - \lambda(x\beta / \sigma) [x\beta / \sigma + \lambda(x\beta / \sigma)]\} \beta \\ \frac{\partial E(y | x)}{\partial x} &= \Phi(x\beta / \sigma) \beta\end{aligned}$$

- Znak β_k jest taki sam jak znak wszystkich rodzajów efektów cząstkowych - można pokazać, że $1 - \lambda(x\beta / \sigma) [x\beta / \sigma + \lambda(x\beta / \sigma)] > 0$
- Ostatni wynik ma następującą interpretację: jeśli prawdopodobieństwo uzyskania $y = 0$ jest bliskie zeru $P(y > 0 | x) = \Phi(x\beta / \sigma) \approx 1$, wtedy model praktycznie nie różni się od modelu liniowego. Dla przypadków ciekawych, to jest takich, w których liczba zer jest znacząca relacja między

współczynnikami uzyskanymi z modelu liniowego i tobitowego będzie w przybliżeniu równa $\beta_{MNK} \approx \Phi(\bar{x}\beta/\sigma)\beta \approx \frac{\text{liczba obserwacji } y_i > 0}{\text{liczba obserwacji}} \times \beta$.

- Oszacowania uzyskane z modelu liniowego zastosowanego do danych ocenzonego dają obciążony estymator

prawdziwego parametru β przy czym obciążenie to rośnie wraz ze wzrostem udziału zer w próbie.

- Estymator MNK nie jest zgodnym estymatorem β zarówno dla przypadku, kiedy używamy wszystkich obserwacji jak i dla przypadku, kiedy szacujemy model używając jedynie $y_i > 0$.

Przykład Wydatki na zakup obuwia (zlogarytmowane) w danym miesiącu a logarytm dochodu, liczba osób i miesiąc, dane z budżetów gospodarstw domowych.

- Oszacowania parametrów

Tobit estimates

Number of obs = 35952

LR chi2(13) = 3420.59

Prob > chi2 = 0.0000

Log likelihood = -53588.396

Pseudo R2 = 0.0309

lobuwie	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ldochg	1.321404	.0439397	30.07	0.000	1.235281	1.407528
los	.4160746	.0165307	25.17	0.000	.383674	.4484753
_Imo_2	.2583854	.1320263	1.96	0.050	-.0003901	.5171609
_Imo_3	1.342601	.1281683	10.48	0.000	1.091387	1.593815
_Imo_4	2.706463	.1253771	21.59	0.000	2.46072	2.952206
_Imo_5	2.609419	.1252601	20.83	0.000	2.363906	2.854933
_Imo_6	1.770684	.1263784	14.01	0.000	1.522978	2.018389
_Imo_7	1.369659	.1275474	10.74	0.000	1.119662	1.619656
_Imo_8	2.043053	.1262433	16.18	0.000	1.795612	2.290494
_Imo_9	2.016629	.1260279	16.00	0.000	1.76961	2.263647
_Imo_10	2.13176	.1257238	16.96	0.000	1.885338	2.378183
_Imo_11	1.816731	.1265315	14.36	0.000	1.568725	2.064736
_Imo_12	1.20642	.1274091	9.47	0.000	.9566939	1.456145

_cons		-13.2753	.3282685	-40.44	0.000	-13.91872	-12.63189
-----+-----							
_se		4.014929	.0276227	(Ancillary parameter)			

Obs. summary: 21467 left-censored observations at lobuwie<=0
 14485 uncensored observations

- W tym przypadku parametry nie mają interpretacji, ponieważ mamy do czynienia z przypadkiem rozwiązań brzegowych a więc zmienna ukryta nie ma, żadnej sensownej interpretacji.
- Na podstawie tego wydruku można jedynie powiedzieć, że wszystkie zmienne w modelu są istotne

- Oszacowania efektów cząstkowych

Marginal Effects: Latent Variable

variable	dF/dx	Std. Err.	z	P> z	X_at	[	95% C.I.	]
ldochg	1.321404	.0439397	30.07	0.000	7.37311	1.23528	1.40752	
los	.4160746	.0165307	25.17	0.000	3.15553	.383675	.448474	
_Imo_2*	.2583854	.1320263	1.96	0.050	0 --> 1	-.000381	.517152	
_Imo_3*	1.342601	.1281683	10.48	0.000	0 --> 1	1.0914	1.59381	
_Imo_4*	2.706463	.1253771	21.59	0.000	0 --> 1	2.46073	2.9522	
_Imo_5*	2.609419	.1252601	20.83	0.000	0 --> 1	2.36391	2.85492	
_Imo_6*	1.770684	.1263784	14.01	0.000	0 --> 1	1.52299	2.01838	
_Imo_7*	1.369659	.1275474	10.74	0.000	0 --> 1	1.11967	1.61965	
_Imo_8*	2.043053	.1262433	16.18	0.000	0 --> 1	1.79562	2.29049	
_Imo_9*	2.016629	.1260279	16.00	0.000	0 --> 1	1.76962	2.26364	
_Imo_10*	2.13176	.1257238	16.96	0.000	0 --> 1	1.88535	2.37817	
_Imo_11*	1.816731	.1265315	14.36	0.000	0 --> 1	1.56873	2.06473	
_Imo_12*	1.20642	.1274091	9.47	0.000	0 --> 1	.956702	1.45614	
_cons	-13.2753	.3282685	-40.44	0.000	1	-13.9187	-12.6319	

Marginal Effects: Unconditional Expected Value

variable	dF/dx	Std. Err.	z	P> z	X_at	[95% C.I.]
ldochg	.5803899	.0192993	30.07	0.000	7.37311	.542564 .618216
los	.1827491	.0072606	25.17	0.000	3.15553	.168519 .19698
_Imo_2*	.1162293	.0579889	2.00	0.045	0 --> 1	.002573 .229885
_Imo_3*	.6641303	.0562944	11.80	0.000	0 --> 1	.553795 .774465
_Imo_4*	1.489383	.0550684	27.05	0.000	0 --> 1	1.38145 1.59732
_Imo_5*	1.425377	.055017	25.91	0.000	0 --> 1	1.31755 1.53321
_Imo_6*	.9066672	.0555082	16.33	0.000	0 --> 1	.797873 1.01546
_Imo_7*	.6789829	.0560216	12.12	0.000	0 --> 1	.569182 .788783
_Imo_8*	1.069416	.0554489	19.29	0.000	0 --> 1	.960738 1.17809
_Imo_9*	1.052992	.0553542	19.02	0.000	0 --> 1	.9445 1.16148
_Imo_10*	1.123004	.0552207	20.34	0.000	0 --> 1	1.01477 1.23123
_Imo_11*	.9338237	.0555754	16.80	0.000	0 --> 1	.824898 1.04275
_Imo_12*	.5895898	.0559609	10.54	0.000	0 --> 1	.479908 .699271
_cons	-5.830805	.1441827	-40.44	0.000	1	-6.1134 -5.54821

Marginal Effects: Conditional on being Uncensored

variable	dF/dx	Std. Err.	z	P> z	X_at	[	95% C.I.	]
ldochg	.4378849	.0145607	30.07	0.000	7.37311	.409347	.466423	
los	.1378782	.0054779	25.17	0.000	3.15553	.127142	.148615	
_Imo_2*	.0870288	.0437507	1.99	0.047	0 --> 1	.001279	.172779	
_Imo_3*	.4846688	.0424722	11.41	0.000	0 --> 1	.401425	.567913	
_Imo_4*	1.067789	.0415473	25.70	0.000	0 --> 1	.986358	1.14922	
_Imo_5*	1.022704	.0415085	24.64	0.000	0 --> 1	.941349	1.10406	
_Imo_6*	.6569141	.0418791	15.69	0.000	0 --> 1	.574833	.738996	
_Imo_7*	.4952665	.0422665	11.72	0.000	0 --> 1	.412426	.578107	
_Imo_8*	.771793	.0418343	18.45	0.000	0 --> 1	.689799	.853787	
_Imo_9*	.7602637	.0417629	18.20	0.000	0 --> 1	.67841	.842118	
_Imo_10*	.8096546	.0416622	19.43	0.000	0 --> 1	.727998	.891311	
_Imo_11*	.6760974	.0419298	16.12	0.000	0 --> 1	.593917	.758278	
_Imo_12*	.4314933	.0422206	10.22	0.000	0 --> 1	.348742	.514244	
_cons	-4.399149	.1087811	-40.44	0.000	1	-4.61236	-4.18594	

Marginal Effects: Probability Uncensored

variable	dF/dx	Std. Err.	z	P> z	X_at	[95% C.I.]
ldochg	.1297743	.0043153	30.07	0.000	7.37311	.121316 .138232
los	.0408624	.0016235	25.17	0.000	3.15553	.03768 .044044
_Imo_2*	.0254667	.0129662	1.96	0.050	0 --> 1	.000053 .05088
_Imo_3*	.1327763	.0125873	10.55	0.000	0 --> 1	.108106 .157447
_Imo_4*	.2618206	.0123132	21.26	0.000	0 --> 1	.237687 .285954
_Imo_5*	.2530791	.0123017	20.57	0.000	0 --> 1	.228968 .27719
_Imo_6*	.1744484	.0124115	14.06	0.000	0 --> 1	.150122 .198775
_Imo_7*	.135431	.0125264	10.81	0.000	0 --> 1	.11088 .159982
_Imo_8*	.2004915	.0123983	16.17	0.000	0 --> 1	.176191 .224792
_Imo_9*	.1979893	.0123771	16.00	0.000	0 --> 1	.173731 .222248
_Imo_10*	.2088873	.0123473	16.92	0.000	0 --> 1	.184687 .233087
_Imo_11*	.1788802	.0124266	14.39	0.000	0 --> 1	.154525 .203236
_Imo_12*	.1193783	.0125128	9.54	0.000	0 --> 1	.094854 .143903
_cons	-1.303759	.032239	-40.44	0.000	1	-1.36695 -1.24057

(*) dF/dx is for discrete change of dummy variable from 0 to 1

- Tabele:

1. efekty cząstkowe dla zmiennej ukrytej
 - nieinterpretowalne w tym przypadku
2. efekty cząstkowe dla bezwarunkowej wartości oczekiwanej $\frac{\partial E(y|\mathbf{x})}{\partial \mathbf{x}}$
 - oczekiwane wydatki gospodarstwa na obuwiu wzrosną o 0.58% jeśli dochód wzrośnie o 1%
3. efekty cząstkowe dla warunkowej wartości oczekiwanej $\frac{\partial E(y|\mathbf{x}, y>0)}{\partial \mathbf{x}}$
 - dla gospodarstw, które poniosły w danym miesiącu jakiegokolwiek wydatki na obuwiu, oczekiwane wydatki gospodarstwa na obuwiu wzrosną o 0.44% jeśli dochód wzrośnie o 1%
4. prawdopodobieństwo, że gospodarstwo poniesie jakiegokolwiek wydatki na obuwiu wzrośnie o 0.13% jeśli dochód wzrośnie o 1%

- Uwagi: zauważmy że całkowity wpływ wzrostu dochodu na wydatki gospodarstw na obuwiu jest w przybliżeniu równy sumie wzrostu wydatków tych gospodarstw, które wydają coś na obuwiu i wzrostu proporcji ych

gospodarstw w ogóle populacji.

Model selekcji próby Heckmana

- Model Heckmana jest modelem samoselekcji (self-selection) próby - podmioty częściowo same decydują o znalezieniu się w wyselekcjonowanej próbie
- Zmienną zależną obserwujemy tylko dla wyselekcjonowanej części próby, zmienne niezależne obserwujemy dla całej próby
- Problem:
 - obserwacje, którymi dysponujemy są nietypowe na tle całej populacji
 - decyzje o przynależności do próby mogą być oparte są na nieobserwowalnych charakterystykach podmiotów

- jeśli nieobserwowalne charakterystyki wpływające na zmienną, którą modelujemy są skorelowane są z nieobserwowalnymi charakterystykami wpływającymi na selekcje do próby, to otrzymane wyniki estymacji mogą być obciążone

Przykład (klasyczny) Decyzje o podjęciu pracy o oferowana płaca. Chcemy oszacować równanie wyjaśniające zależność oferowanej płacy od charakterystyk respondenta. Niestety obserwujemy oferty płacy jedynie dla pracujących (w postaci płac które otrzymują). Oczywiste jest, że oferowana płaca w_i^o zależy od charakterystyk społeczno ekonomicznych respondenta danych przez zmienne x_{1i} (np. od wykształcenia i płci). Zakładamy, że respondent podejmie pracę jedynie wtedy, gdy oferowana płaca jest wyższa od pewnego minimum w_i^r (reservation wage). Te minimum płacowe jest także zależne od charakterystyk respondenta (np. osoby, które otrzymują rodzinę są silniej motywowane do podjęcia pracy nawet przy niskich stawkach).

1. założmy, że równanie oferowanej płacy ma postać

$$\ln w_i^o = \mathbf{x}_{1i}\boldsymbol{\beta}_1 + u_{1i}$$

2. respondent podejmie pracę jeśli

$$w_i^o > w_i^r$$

założmy, że:

$$\ln w_i^r = \mathbf{x}_{2i}\boldsymbol{\beta}_2 + u_{2i}$$

Podstawiając to równanie do poprzedniej nierówności i logarytmując stronami dochodzimy do wniosku, że respondent podejmie pracę jeśli:

$$\ln w_i^o - \ln w_i^r = \mathbf{x}_{1i}\boldsymbol{\beta}_1 + u_{1i} - \mathbf{x}_{2i}\boldsymbol{\beta}_2 - u_{2i} = \mathbf{x}_i\boldsymbol{\delta}_2 + v_{2i} > 0$$

gdzie $\mathbf{x}_i = (\mathbf{x}_{1i}, \mathbf{x}_{2i})$

Uwagi:

- ponieważ $x_i = (x_{1i}, x_{2i})$, więc wszystkie zmienne wpływające na równanie płac wpływają też równanie selekcji do próby
- skoro $v_{2i} = u_{1i} - u_{2i}$ więc oczekujemy dodatniej korelacji między u_{1i} i v_{2i}
- różnica między przypadkiem analizowanym w przypadku modelu tobitowego: nie znamy dokładnie miejsca cenzurowania $\log w_i^r$
- ponieważ parametry równania $\log w_i^o - \log w_i^r = x_i \delta_2 + v_{2i}$ są mieszaniną parametrów równania płac i równania selekcji więc nie można ich interpretować jako parametrów równania podaży pracy.

- Podstawowy model:

$$y_1 = \mathbf{x}_1\boldsymbol{\beta} + u_1 \quad (\text{Równanie regresji})$$

$$y^* = \mathbf{x}\boldsymbol{\delta} + u_2 \quad (\text{Równanie selekcji})$$

$$y_2 = \begin{cases} 0 & \text{dla } y^* \leq 0 \\ 1 & \text{dla } y^* > 0 \end{cases}$$

$$(u_1, u_2) \sim N\left(0, \begin{bmatrix} \sigma^2 & \rho \\ \rho & 1 \end{bmatrix}\right)$$

$(\mathbf{x}, y_2)$ są zawsze obserwowalne, y_1 jest obserwowalne tylko dla $y_2 = 1$

- Warunkowa wartość oczekiwana dla obserwacji wyselekcjonowanych

$$E(y_1 | \mathbf{x}, y_2 = 1) = \mathbf{x}_1\boldsymbol{\beta} + \sigma\rho\lambda(\mathbf{x}\boldsymbol{\delta})$$

gdzie $\lambda(\cdot) = \frac{\phi(\cdot)}{\Phi(\cdot)}$ jest nazywany odwrotnością ilorazu Mills'a

- Zwykłą regresję liniową regresję liniową y_1 na x_1 , da niezgodne estymatory parametru β ponieważ pominięty został element $\sigma\rho\lambda(x\delta)$ (problem zmiennej pominiętej: $\sigma\rho\lambda(x\delta)$ jest zazwyczaj skorelowane z x_1)
- Uwaga:
 - w szczególnym przypadku, gdy błędy losowe (charakterystyki nieobserwowalne) w obu równaniach są nieskorelowane ($\rho = 0$) estymator MNK jest zgodny
- Zauważmy, że:
 - równanie selekcji opisuje zwykły model probitowy - z modelu probitowego można uzyskać oszacowania δ
 - w równaniu $y_1 = x_1\beta + \sigma\rho\lambda(x\delta) + \varepsilon$ błąd losowy ε jest nieskorelowany z x_1 i $\lambda(x\delta)$ - parametr β można zgodnie oszacować MNK

- Dwustopniowa procedura Heckmana:
 1. estymujemy najpierw regresję probitową y_{2i} na x_i , liczymy $\hat{\lambda}_i = \lambda(x_i\hat{\delta})$
 2. przeprowadzamy zwykłą regresję y_{2i} na x_{1i} i $\hat{\lambda}_i$
- Uzyskane w ten sposób estymatory parametru β są zgodne, trudniej uzyskać jest prawidłowe oszacowania jego macierzy wariancji
- Istnienie problemu selekcji próby można przetestować testując istotność parametru $\gamma = \sigma\rho$
- Estymator parametrów w modelu Heckmana można także uzyskać za pomocą *MNW*
- Uwaga:
 - formalnie biorąc model Heckmana można wyestymować nawet wtedy, gdy $x = x_1$ (w równaniu regresji i selekcji występują te same zmienne)

- w literaturze pokazuje się jednak, że dla uzyskania precyzyjnych oszacowań w równaniu selekcji powinny pojawiać się zmienne, które nie pojawiają się w równaniu regresji

Przykład Równanie płac a płać i wykształcenie - dane PGSS. W równaniu selekcji poza charakterystykami respondentów pojawiają się dochody pozostałych członków gospodarstwa. Zakładamy, że wysokie dochody pozostałych członków gospodarstwa wpływają negatywnie na motywację do podjęcia pracy ale nie mają wpływu na poziom płacy. Do obu równań dodano zmienne zerojedynkowe związane z rokiem badania.

- Wyniki regresji metodą dwustopniową

```

Heckman selection model -- two-step estimates
(regression model with sample selection)

Number of obs      =      7895
Censored obs       =      4490
Uncensored obs     =      3405

Wald chi2(28)      =      5115.02
Prob > chi2        =      0.0000

```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
lrincome						
_Isex_2	-.4161068	.0203238	-20.47	0.000	-.4559408	-.3762729
_Idegree_1	-.1686797	.2480776	-0.68	0.497	-.6549029	.3175435
_Idegree_2	.403593	.2313048	1.74	0.081	-.0497561	.8569421
_Idegree_3	.7092761	.2327722	3.05	0.002	.253051	1.165501
_Idegree_4	.7318259	.2381942	3.07	0.002	.2649738	1.198678
_Idegree_5	.9678059	.2342956	4.13	0.000	.508595	1.427017
_Idegree_6	.9905437	.2332893	4.25	0.000	.533305	1.447782
_Idegree_7	1.145547	.2367375	4.84	0.000	.6815499	1.609544
_Idegree_8	.8485571	.2389341	3.55	0.000	.3802549	1.316859
_Idegree_9	1.398745	.2353774	5.94	0.000	.9374135	1.860076

_Ipgssy~1993		.2621812	.0295216	8.88	0.000	.2043199	.3200425
_Ipgssy~1994		.514156	.0297156	17.30	0.000	.4559145	.5723975
_Ipgssy~1995		.7689225	.0298112	25.79	0.000	.7104937	.8273513
_Ipgssy~1997		1.289651	.0275999	46.73	0.000	1.235556	1.343745
_cons		4.306924	.2376954	18.12	0.000	3.841049	4.772798
-----+-----							
select							
_Isex_2		-.2238364	.0318749	-7.02	0.000	-.2863101	-.1613627
_Idegree_1		-.1948326	.2648298	-0.74	0.462	-.7138894	.3242242
_Idegree_2		.6486607	.2498993	2.60	0.009	.1588671	1.138454
_Idegree_3		1.49557	.2498923	5.98	0.000	1.00579	1.98535
_Idegree_4		.9601074	.262451	3.66	0.000	.4457129	1.474502
_Idegree_5		1.35211	.2551189	5.30	0.000	.8520858	1.852133
_Idegree_6		1.721247	.2505348	6.87	0.000	1.230208	2.212286
_Idegree_7		2.183448	.2593931	8.42	0.000	1.675047	2.691849
_Idegree_8		1.382837	.2656166	5.21	0.000	.8622379	1.903436
_Idegree_9		2.316108	.2549755	9.08	0.000	1.816366	2.815851
lextrincome		-.5718438	.0248109	-23.05	0.000	-.6204723	-.5232154
_Ipgssy~1993		.1094373	.0512778	2.13	0.033	.0089346	.2099399
_Ipgssy~1994		.2147374	.0525084	4.09	0.000	.1118228	.317652
_Ipgssy~1995		.4566194	.0559933	8.15	0.000	.3468746	.5663643
_Ipgssy~1997		.6723417	.0580917	11.57	0.000	.558484	.7861994
_cons		1.719688	.2771778	6.20	0.000	1.17643	2.262947

-----+-----	
mills	
lambda	.2405734 .0387668 6.21 0.000 .1645918 .316555
-----+-----	
rho	0.44030
sigma	.54638986
lambda	.24057336 .0387668

- Wyniki regresji *MNW*

Heckman selection model
(regression model with sample selection)

Number of obs = 7895
Censored obs = 4490
Uncensored obs = 3405

Log likelihood = -6984.195

Wald chi2(14) = 3803.36
Prob > chi2 = 0.0000

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
lrrincome						
_Isex_2	-.4034674	.0195752	-20.61	0.000	-.4418341	-.3651007
_Idegree_1	-.1548126	.2491579	-0.62	0.534	-.6431531	.3335278
_Idegree_2	.3813667	.2321498	1.64	0.100	-.0736385	.8363719
_Idegree_3	.6560963	.2328991	2.82	0.005	.1996224	1.11257
_Idegree_4	.700648	.2388085	2.93	0.003	.232592	1.168704
_Idegree_5	.9247913	.2346873	3.94	0.000	.4648126	1.38477
_Idegree_6	.932306	.2332397	4.00	0.000	.4751646	1.389447
_Idegree_7	1.072007	.2360107	4.54	0.000	.6094347	1.53458
_Idegree_8	.8073619	.2392929	3.37	0.001	.3383564	1.276367
_Idegree_9	1.327589	.2347856	5.65	0.000	.8674171	1.78776

_Ipgssy~1993		.266303	.0291733	9.13	0.000	.2091243	.3234816
_Ipgssy~1994		.5195537	.0293472	17.70	0.000	.4620343	.5770731
_Ipgssy~1995		.7725663	.0294722	26.21	0.000	.7148018	.8303308
_Ipgssy~1997		1.295903	.0272155	47.62	0.000	1.242562	1.349244
_cons		4.394026	.2363644	18.59	0.000	3.93076	4.857292
-----+-----							
select							
_Isex_2		-.2284008	.031874	-7.17	0.000	-.2908726	-.165929
_Idegree_1		-.194801	.2648981	-0.74	0.462	-.7139916	.3243897
_Idegree_2		.6475847	.2499664	2.59	0.010	.1576596	1.13751
_Idegree_3		1.493884	.24996	5.98	0.000	1.003972	1.983797
_Idegree_4		.9605035	.2625117	3.66	0.000	.44599	1.475017
_Idegree_5		1.352345	.2551876	5.30	0.000	.8521863	1.852504
_Idegree_6		1.720029	.2506042	6.86	0.000	1.228854	2.211204
_Idegree_7		2.188793	.2594437	8.44	0.000	1.680293	2.697294
_Idegree_8		1.385735	.265691	5.22	0.000	.8649903	1.90648
_Idegree_9		2.313172	.2549965	9.07	0.000	1.813388	2.812956
lextrincome		-.5774507	.0246827	-23.39	0.000	-.6258278	-.5290735
_Ipgssy~1993		.1083887	.0512495	2.11	0.034	.0079415	.2088359
_Ipgssy~1994		.2182079	.0524451	4.16	0.000	.1154174	.3209985
_Ipgssy~1995		.4611189	.055938	8.24	0.000	.3514825	.5707554
_Ipgssy~1997		.6751271	.0579792	11.64	0.000	.56149	.7887642
_cons		1.753689	.2769203	6.33	0.000	1.210936	2.296443

```

-----+-----
      /athrho |      .3486491      .0611634      5.70      0.000      .2287711      .4685271
      /lnsigma |     -.6277223      .0163238     -38.45      0.000     -.6597164     -.5957282
-----+-----
      rho |      .335177      .054292
      sigma |      .5338063      .0087138
      lambda |      .1789196      .0310179
-----+-----
LR test of indep. eqns. (rho = 0):   chi2(1) =      31.12   Prob > chi2 = 0.0000
-----

```

- odrzucamy $H_0 : \rho = 0$, ponieważ
 - dla metody dwustopniowej statystyki z dla oszacowania parametru $\rho\sigma$ (przy zmiennej λ) jest istotnie różna od zera
 - dla MNW na podstawie statystyki LR
- Oszacowany współczynnik korelacji ρ między błędami losowymi z obu równań jest zgodnie z teorią dodatni

- Współczynniki przy zmiennych w równaniu regresji interpretujemy identycznie jak w zwykłym modelu liniowym
 - Poza oszacowaniami współczynników w równaniu regresji często interesują nas wielkości efektów cząstkowych.
 - Najciekawsza jest
 - wielkość efektów cząstkowych związanych z prawdopodobieństwem selekcji
- $$\frac{\partial \Pr(y_2 = 1)}{\partial x} = \phi(x\beta) \beta$$
- wielkość efektów cząstkowych dla y_1^+ gdzie

$$y_1^+ = \begin{cases} y_1 & \text{dla } y_1 \text{ obserwowalnego} \\ 0 & \text{dla } y_1 \text{ nieobserwowalnego} \end{cases}$$

Przykład Kontynuacja

- Efekty czątkowe dla prawdopodobieństw

Marginal effects after heckman

y = Pr(select) (predict, psel)
= .41197215

variable	dy/dx	Std. Err.	z	P> z	[95% C.I.]	X
-----+-----						
_Isex_2*	-.0890236	.01242	-7.17	0.000	-.113359 -.064688	.572894
_Idegr~1*	-.0740154	.09772	-0.76	0.449	-.26554 .117509	.062571
_Idegr~2*	.2533293	.09542	2.65	0.008	.066313 .440346	.262318
_Idegr~3*	.5392117	.07054	7.64	0.000	.400951 .677473	.244459
_Idegr~4*	.3601846	.08217	4.38	0.000	.199128 .521242	.031032
_Idegr~5*	.4725578	.06144	7.69	0.000	.352141 .592975	.062065
_Idegr~6*	.5866782	.05685	10.32	0.000	.47525 .698106	.185687
_Idegr~7*	.592645	.02194	27.01	0.000	.549638 .635652	.041545
_Idegr~8*	.471792	.05734	8.23	0.000	.359405 .584179	.024446
_Idegr~9*	.6294342	.02332	26.99	0.000	.58373 .675139	.079417
_Ip~1993*	.0424598	.02019	2.10	0.035	.002894 .082025	.186067
_Ip~1994*	.0858635	.02078	4.13	0.000	.045144 .126583	.184927

_Ip~1995*	.1818316	.02191	8.30	0.000	.138881	.224782	.181507
_Ip~1997*	.2638418	.02203	11.98	0.000	.220663	.307021	.260165
lextin~e	-.2247384	.00964	-23.32	0.000	-.243623	-.205854	5.93937

(*) dy/dx is for discrete change of dummy variable from 0 to 1

- Te efekty cząstkowe interpretujemy tak jak w przypadku probitu i logitu

- Efekty cząskowe dla y_1^+

Marginal effects after heckman

$$y = E(\ln \text{income}^* | \text{Pr}(\text{select})) \text{ (predict, yexpected)}$$

$$= 2.3184213$$

variable	dy/dx	Std. Err.	z	P> z	[	95% C.I.	]	X
-----+-----								
_Isex_2*	-.6608731	.0709	-9.32	0.000	-.799843	-.521903		.572894
_Idegr~1*	-.4618541	.54719	-0.84	0.399	-1.53432	.610615		.062571
_Idegr~2*	1.589577	.56954	2.79	0.005	.473302	2.70585		.262318
_Idegr~3*	3.380553	.45301	7.46	0.000	2.49266	4.26844		.244459
_Idegr~4*	2.477582	.54925	4.51	0.000	1.40106	3.5541		.031032
_Idegr~5*	3.313689	.44516	7.44	0.000	2.4412	4.18618		.062065
_Idegr~6*	3.895918	.39973	9.75	0.000	3.11246	4.67937		.185687
_Idegr~7*	4.192403	.27253	15.38	0.000	3.65826	4.72655		.041545
_Idegr~8*	3.237574	.42774	7.57	0.000	2.39922	4.07592		.024446
_Idegr~9*	4.605388	.27225	16.92	0.000	4.07178	5.13899		.079417
_Ip~1993*	.350028	.11715	2.99	0.003	.120413	.579643		.186067
_Ip~1994*	.7133279	.12388	5.76	0.000	.470528	.956128		.184927
_Ip~1995*	1.403312	.13421	10.46	0.000	1.14026	1.66636		.181507
_Ip~1997*	2.143354	.13806	15.52	0.000	1.87276	2.41395		.260165
lextin~e	-1.235701	.05394	-22.91	0.000	-1.34142	-1.12998		5.93937

(*) dy/dx is for discrete change of dummy variable from 0 to 1

- Zwróćmy uwagę, że całkowity efekt płci i wykształcenia na dochody osiąmane z pracy jest znacznie silniejszy niż by to wynikało z oszacowań uzyskanych z równia płac (np. dla degree9 mamy odpowiednio o 133% lepszą ofertę ale o 461% wyższy oczekiwany dochód z pracy)
- Na oczekiwany dochód osiągany z pracy (y_1^+) wpływa
 - decyzja o podjęci pracy: jeśli nie pracujemy do dochód z pracy równy jest zero
 - wielkość płacy jaką oferuje nam rynek