Binarne zmienne zależne

- Zmienna zależna nie jest ciągła ale przyjmuje zero lub jeden

- **Przykład:** szukanie determinant bezrobocia (próba przekrojowa)
  - zmienna objaśniana: zerojedynkowa (pracujący, bezrobotny)
  - szukamy związku między charakterystykami ekonomiczno-społecznymi a byciem bezrobotnym

- W modelach dla zmiennych binarnych staramy się wyjaśnić prawdopodobieństwa stanów za pomocą zmiennych niezależnych

- Prawdopodobieństwo pojedynczego zdarzenia dane jest rozkładem dwumianowym

$$\Pr\left(y_i|\, \boldsymbol{x}_i\right) = \begin{cases} 1 - p\left(\boldsymbol{x}_i\right) & \text{dla} \quad y_i = 0 \\ p\left(\boldsymbol{x}_i\right) & \text{dla} \quad y_i = 1 \end{cases}$$

gdzie $p\left(\boldsymbol{x}_i\right) = F\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)$ i $F\left(\cdot\right)$ jest zazwyczaj dystrybuantą jakiegoś rozkładu prawdopodobieństwa.

- Estymator parametru $\boldsymbol{\beta}$ liczymy za pomocą MNW.

Interpretacja ekonomiczna

Podejście indeksowe

- y^* jest użytecznością lub zyskiem netto
- y^* jest nieobserwowalne
- y^* jest losowe
$$y^* = \boldsymbol{x\beta} + \boldsymbol{\varepsilon}$$
gdzie ε ma symetryczną dystrybuantę z $\mathrm{E}\left(\boldsymbol{\varepsilon}\right) = 0$ i $\mathrm{Var}\left(\boldsymbol{\varepsilon}\right) = 1$
- Parametry modelu szacujemy na postawie obserwowalnych wyborów y, na które zdeterminowane są nieobserwowalnym y^*.

- Wybór $y = 1$ jest podejmowany jeśli przyrost związanej z nim użyteczności (zysku) jest dodatni

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

- Prawdopodobieństwo, że $y = 1$ jest równe

$$\begin{aligned} \Pr\left(y = 1 \middle| \boldsymbol{x}\right) &= \Pr\left(y^* > 0 \middle| \boldsymbol{x}\right) = \Pr\left\{\boldsymbol{x\beta} + \boldsymbol{\varepsilon} > 0\right\} \\ &= \Pr\left\{\boldsymbol{\varepsilon} > -\boldsymbol{x\beta}\right\} \end{aligned}$$

- Skoro dystybuanta $\boldsymbol{\varepsilon}$ jest symetryczna, to

$$\Pr\left(y = 1 \middle| \boldsymbol{x}\right) = \Pr\left\{\boldsymbol{\varepsilon} < \boldsymbol{x\beta}\right\} = F\left(\boldsymbol{x\beta}\right)$$

gdzie $F\left(\cdot\right)$ jest dystrybuantą $\boldsymbol{\varepsilon}$

- W tym podejściu, wybór zależy jedynie od charakterystyk podmiotu i alternatywy wybranej

Losowa użyteczność

- Załóżmy, że y_0^* i y_1^* są użytecznościami z działań 0 i 1.

- y_0^* i y_1^* są nieobserwowalne i dane równaniami

$$
\begin{aligned}
y_0^* &= \boldsymbol{x}_0\boldsymbol{\beta} + \boldsymbol{\varepsilon}_0 \\
y_1^* &= \boldsymbol{x}_1\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_1
\end{aligned}
$$

- Załóżmy, że $\varepsilon_0 - \varepsilon_1$ mają symetryczną dystrybuantę

- Obserwujemy y takie że

$$\begin{array}{lll} y=0 & \text{jeżeli} & y_0^* > y_1^* \\ y=1 & \text{jeżeli} & y_0^* \leq y_1^* \end{array}$$

- Podmiot podejmuje działanie 0 jeśli y_0^* jest lepsze niż y_1^* i 1 jeśli y_1^* jest lepsze niż y_0^*.

- Prawdopodobieństwo zaobserwowania 1 jest równe:

$$\begin{aligned} \Pr\left(y=1|\,\boldsymbol{x}\right) &= \Pr\left\{y_0^* \leq y_1^*|\,\boldsymbol{x}\right\} = \Pr\left\{\boldsymbol{x}_0\boldsymbol{\beta} + \boldsymbol{\varepsilon}_0 \leq \boldsymbol{x}_1\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_1\right\} \\ &= \Pr\left\{\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_0 \geq \boldsymbol{x}_0\boldsymbol{\beta} - \boldsymbol{x}_1\boldsymbol{\gamma}\right\} = F\left(\boldsymbol{x}_1\boldsymbol{\gamma} - \boldsymbol{x}_0\boldsymbol{\beta}\right) \end{aligned}$$

- W modelu losowej użyteczności wybór zależy nie tylko od charakterystyk podmiotu, użyteczności alternatywy wybranej, ale też od użyteczności alternatywy niewybranej.

Interpretacja współczynników

- Wartość oczekiwana zmiennej zależnej w modelu dla zmiennej binarnej jest równa:

$$\mathrm{E}\left(y_i \middle| \boldsymbol{x}_i\right) = 0\left[1 - F\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)\right] + 1F\left(\boldsymbol{x}_i\boldsymbol{\beta}\right) = F\left(\boldsymbol{x}_i\boldsymbol{\beta}\right) = p\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)$$

- Ponieważ przyjęta jest konwencja, że 1 oznacza sukces a 0 porażkę więc, wartość oczekiwana zmiennej zależnej w modelu dla zmiennej binarnej równa jest prawdopodobieństwu sukcesu.

- Efekt cząstkowy jest równy:

$$\frac{\partial\,\mathrm{E}\left(y \middle| \boldsymbol{x}\right)}{\partial \boldsymbol{x}} = \frac{\partial p\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)}{\partial \boldsymbol{x}} = f\left(\boldsymbol{x}\boldsymbol{\beta}\right)\boldsymbol{\beta}$$

- Efekt cząstkowy w tym przypadku interpretujemy jako wpływ jednostkowej zmiany zmiennej niezależnej na prawdopodobieństwo sukcesu.

- Ze wzoru widać, że wielkość efektu cząstkowego zależy od tego w jakim punkcie jest on liczony

- Zazwyczaj liczymy go dla średnich wartości zmiennych egzogenicznych.

- Dla zerojedynkowych zmiennych objaśniających efekt krańcowy definiuje się jako $\Delta F\left(\overline{\boldsymbol{x}}\boldsymbol{\beta}\right) = F\left(\overline{\boldsymbol{x}}_1\boldsymbol{\beta}\right) - F\left(\overline{\boldsymbol{x}}_0\boldsymbol{\beta}\right)$ - a więc różnice między prawdopodobieństwem sukcesu dla zmiennej zerojedynkowej równej zeru i równej 1 przy wszystkich pozostałych zmiennych ustalonych na poziomie średnich.

- Ponieważ funkcja gęstości $f\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)$ jest zawsze dodatnia więc znak efektu cząstkowego jest równy znakowi współczynnika przy zmiennej. Wynika z tego, że jakkolwiek same współczynniki nie są zazwyczaj interpretowalne to można interpretować ich znaki.

- Można jednak interpretować proporcje między współczynnikami ponieważ są one równe proporcjom między efektami krańcowymi

$$\frac{\partial \operatorname{E}(y|\boldsymbol{x})}{\partial \boldsymbol{x}_j} \Big/ \frac{\partial \operatorname{E}(y|\boldsymbol{x})}{\partial \boldsymbol{x}_k} = \frac{\boldsymbol{\beta}_j}{\boldsymbol{\beta}_k}$$

Miary dopasowania

- Często stosowaną miarą dopasowania jest

$$pseudo\text{-}R^2 = 1 - \frac{\ell\left(\widetilde{\boldsymbol{\beta}}\right)}{\ell\left(\widetilde{\boldsymbol{\beta}}_R\right)}$$

  gdzie $\widetilde{\boldsymbol{\beta}}_R$ zawiera jedynie stałą

- $pseudo\text{-}R^2$ spełnia

$$0 \leq pseudo\text{-}R^2 \leq 1$$

- $pseudo\text{-}R^2$ nie można jednak w pełni ściśle interpretować jako procent zmienności wyjaśnionej przez model.

- Innym sposobem mierzenia jakości dopasowania jest analiza tablicy częstości:

Prognozowane \ Zaobserwowane	0	1	Razem
0	n_{00}	n_{01}	$n_{00}+n_{01}$
1	n_{10}	n_{11}	$n_{10}+n_{11}$
Razem	$n_{00}+n_{10}$	$n_{01}+n_{11}$	

- Arbitralnie zakładamy, że model przewiduje $y_i = 1$ jeśli $\widehat{p}_i = \boldsymbol{x}_i\boldsymbol{b} > 0.5$

Liniowy model prawdopodobieństwa (LPM)

- W liniowym modelu prawdopodobieństwa LPM (**L**inear **P**robability **M**odel) zakładamy, że dystrybuanta rozkładu ma postać funkcji liniowej:

$$p\left(\boldsymbol{x}_i\right) = F\left(\boldsymbol{x}_i\boldsymbol{\beta}\right) = \boldsymbol{x}_i\boldsymbol{\beta}$$

- W tym przypadku estymacja parametrów $\boldsymbol{\beta}$ sprowadza się do szacowania MNK równania

$$y_i = \boldsymbol{x}_i\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i$$

- Podejście to mimo swojej prostoty związane jest z dwoma poważnymi problemami:

1. Błędy losowe w równaniu regresji będą heteroskedastyczne. Zmienna ma rozkład dwumianowy więc:

$$\mathrm{E}\left(y_i\right) = F\left(\boldsymbol{x}_i\boldsymbol{\beta}\right) = \boldsymbol{x}_i\boldsymbol{\beta}$$

$$\mathrm{E}\left(y_i^2\right) = 0^2\left(1 - \boldsymbol{x}_i\boldsymbol{\beta}\right) + 1^2\boldsymbol{x}_i\boldsymbol{\beta} = \boldsymbol{x}_i\boldsymbol{\beta}$$

$$\begin{aligned} \mathrm{Var}\left(y_i\right) &= \mathrm{E}\left(x_i^2\right) - \mathrm{E}\left(x_i\right)^2 = \boldsymbol{x}_i\,\boldsymbol{\beta} - \left(\boldsymbol{x}_i\boldsymbol{\beta}\right)^2 \\ &= \boldsymbol{x}_i\boldsymbol{\beta}\left(1 - \boldsymbol{x}_i\boldsymbol{\beta}\right) \end{aligned}$$

   Wariancja y_i zależy od $\boldsymbol{x}_i$.
2. Nie ma gwarancji, że $\widehat{y}_i = \boldsymbol{x}_i\ \boldsymbol{b} \in [0,1]$. Ponieważ $\widehat{y}_i$ interpretujemy jako prawdopodobieństwa stanów - ujemne lub większe od 1 wartości dopasowane będą nieinterpretowalne

- Mimo tych wad liniowy model prawdopodobieństwa jest dosyć często używany w praktycznych zastosowaniach ze względu na łatwość estymacji.

- Efekty cząstkowe w tym modelu równe są po prostu $\boldsymbol{\beta}$

- Wpływu heteroskedastyczności na wyniki wnioskowania można wyeliminować stosując odporną macierz wariancji White'a

- Heteroskedastyczność można też wyeliminować za pomocą $UMNK$:
  - estymujemy LPM za pomocą MNK
  - liczymy wariancje $\widehat{\boldsymbol{\sigma}}_i^2 = \boldsymbol{x}_i\boldsymbol{b}\left[1-\boldsymbol{x}_i\boldsymbol{b}\right] = \widehat{y}_i\left(1-\widehat{y}_i\right)$ (wzór ten ma sens jedynie dla $\widehat{y}_i \in [0,1]$)
  - liczymy regresję $y_i/\widehat{\boldsymbol{\sigma}}_i$ na $x_{1i}/\widehat{\boldsymbol{\sigma}}_i, \ldots, x_{Ki}/\widehat{\boldsymbol{\sigma}}_i$

Przykład Zależność aktywności zawodowej od wieku i wykształcenia: dane BAEL - pytanie: czy w ostatnim tygodniu wykonywał Pan(i) jakąkolwiek pracę?

```
      Source |       SS       df       MS              Number of obs =   47860
-------------+------------------------------           F(  9, 47850) =  993.61
       Model |  1827.14253     9  203.015837           Prob > F      =  0.0000
    Residual |  9776.79083 47850  .204321648           R-squared     =  0.1575
-------------+------------------------------           Adj R-squared =  0.1573
       Total |  11603.9334 47859  .242460841           Root MSE      =  .45202

------------------------------------------------------------------------------
   _Iworks_1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wiek |  -.0036288   .0001188   -30.55   0.000    -.0038616   -.0033961
   _Iwyksz_2 |  -.1155233   .0141975    -8.14   0.000    -.1433507    -.087696
   _Iwyksz_3 |  -.1575525   .0082029   -19.21   0.000    -.1736303   -.1414747
   _Iwyksz_4 |  -.3616109   .0098378   -36.76   0.000     -.380893   -.3423287
   _Iwyksz_5 |   -.196475   .0078076   -25.16   0.000    -.2117779    -.181172
  _Iwyksz_60 |  -.7496461   .0163531   -45.84   0.000    -.7816984   -.7175937
  _Iwyksz_61 |  -.4814629   .0077523   -62.11   0.000    -.4966576   -.4662683
  _Iwyksz_70 |  -.5166902   .0162939   -31.71   0.000    -.5486266   -.4847539
```

```
  _Iwyksz_71 |   -.594044   .0302346   -19.65   0.000    -.6533043   -.5347837
       _cons |   .8569807   .0084361   101.59   0.000     .8404459    .8735155
------------------------------------------------------------------------------
```

- Wyniki testu na heteroskedastyczność:

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of _Iworks_1

         chi2(1)      =   961.62
         Prob > chi2  =   0.0000
```

- Niektóre wartości przewidywane:

```
       +-----------+
       |      yhat |
       |-----------|
    1. | -.1974881 |
    2. | -.1793439 |
    3. | -.1720862 |
```

```
4. | -.1466843 |
5. | -.1398648 |
```

- Jednak z 47860 obserwacji tylko $144 \notin (0,1)$

Probit

- Dystrybuanta $F(\boldsymbol{x}_i\boldsymbol{\beta}) = \Phi(\boldsymbol{x}_i\boldsymbol{\beta})$, z rozkładu normalnego.

- Funkcja prawdopodobieństwa zajścia pojedynczego zdarzenia:

$$\Pr(y_i|\, \boldsymbol{x}_i) = \begin{cases} 1 - \Phi(\boldsymbol{x}_i\boldsymbol{\beta}) & \text{dla} \quad y_i = 0 \\ \Phi(\boldsymbol{x}_i\boldsymbol{\beta}) & \text{dla} \quad y_i = 1 \end{cases}$$
$$= \left[1 - \Phi(\boldsymbol{x}_i\boldsymbol{\beta})\right]^{1-y_i} \Phi(\boldsymbol{x}_i\boldsymbol{\beta})^{y_i}$$

- Funkcja wiarygodności ma postać

$$L(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(1 - \Phi(\boldsymbol{x}_i\boldsymbol{\beta})\right)^{1-y_i} \Phi(\boldsymbol{x}_i\boldsymbol{\beta})^{y_i}$$

- Logarytm funkcji wiarygodności będzie ma postać

$$\ell\left(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\beta}\right)=\sum_{i=1}^{n}\left[\left(1-y_i\right)\ln\left(1-\Phi\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)\right)+y_i\ln\Phi\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)\right]$$

- Często podaje się jedynie $\ell_i\left(\boldsymbol{\beta}\right)$

$$\ell_i\left(\boldsymbol{\beta}\right)=\left(1-y_i\right)\ln\left(1-\Phi\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)\right)+y_i\ln\Phi\left(\boldsymbol{x}_i\boldsymbol{\beta}\right).$$

- Efekt cząstkowy

$$\frac{\partial\,\mathrm{E}\left(y|\,\boldsymbol{x}\right)}{\partial\boldsymbol{x}}=\boldsymbol{\phi}\left(\boldsymbol{x\beta}\right)\boldsymbol{\beta}$$

Przykład Zależność aktywności zawodowej od wieku i wykształcenia: kontynuacja

- Współczynniki:

```
Probit estimates                                  Number of obs   =      47860
                                                  LR chi2(9)      =    8274.24
                                                  Prob > chi2     =     0.0000
Log likelihood = -28311.095                       Pseudo R2       =     0.1275

------------------------------------------------------------------------------
   _Iworks_1 |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wiek |  -.0113366   .0003666   -30.93   0.000     -.012055   -.0106181
   _Iwyksz_2 |  -.3254287   .0404874    -8.04   0.000    -.4047825   -.2460749
   _Iwyksz_3 |   -.433603    .023857   -18.18   0.000     -.480362   -.3868441
   _Iwyksz_4 |  -.9675114   .0286116   -33.82   0.000    -1.023589   -.9114338
   _Iwyksz_5 |  -.5330219   .0227775   -23.40   0.000    -.5776649   -.4883789
  _Iwyksz_60 |  -2.513776   .0730264   -34.42   0.000    -2.656905   -2.370647
  _Iwyksz_61 |  -1.343109   .0233078   -57.62   0.000    -1.388792   -1.297427
  _Iwyksz_70 |  -1.678283   .0650819   -25.79   0.000    -1.805842   -1.550725
  _Iwyksz_71 |  -2.321871   .1804895   -12.86   0.000    -2.675624   -1.968118
```

```
       _cons |   1.027194   .0256271    40.08   0.000     .9769656    1.077422
------------------------------------------------------------------------------
```

- Efekty krańcowe:

```
Probit estimates                                  Number of obs =  47860
                                                  LR chi2(9)    =8274.24
                                                  Prob > chi2   = 0.0000
Log likelihood = -28311.095                       Pseudo R2     = 0.1275

------------------------------------------------------------------------------
_Iwork~1 |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
    wiek |  -.0043578   .0001406   -30.93   0.000   43.9481  -.004633 -.004082
_Iwyks~2*|  -.1181235   .0136421    -8.04   0.000   .027267  -.144862 -.091385
_Iwyks~3*|  -.1582424   .0081461   -18.18   0.000    .19135  -.174209 -.142276
_Iwyks~4*|  -.3006814   .0063921   -33.82   0.000   .082804   -.31321 -.288153
_Iwyks~5*|  -.1946251   .0077792   -23.40   0.000    .26728  -.209872 -.179378
_Iwyk~60*|  -.4090495   .0025428   -34.42   0.000   .020079  -.414033 -.404066
_Iwyk~61*|  -.4366777   .0058916   -57.62   0.000   .290639  -.448225  -.42513
_Iwyk~70*|  -.3783463   .0048507   -25.79   0.000   .020664  -.387854 -.368839
_Iwyk~71*|  -.3921597   .0035269   -12.86   0.000   .004952  -.399072 -.385247
```

```
---------+--------------------------------------------------------------------
  obs. P |   .4131425
 pred. P |   .3926327  (at x-bar)
------------------------------------------------------------------------------
(*) dF/dx is for discrete change of dummy variable from 0 to 1
    z and P>|z| are the test of the underlying coefficient being 0
```

Logit

- Model logitowy otrzymujemy dla

$$F(\boldsymbol{x}_i\boldsymbol{\beta}) = \frac{e^{\boldsymbol{x}_i\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i\boldsymbol{\beta}}} = \Lambda(\boldsymbol{x}_i\boldsymbol{\beta})$$

- Funkcja prawdopodobieństwa dla pojedynczej obserwacji

$$\Pr(y_i) = \begin{cases} \frac{1}{1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}} & \text{dla} \quad y_i = 0 \\ \frac{e^{\boldsymbol{x}_i\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}} & \text{dla} \quad y_i = 1 \end{cases}$$

- Logarytm funkcji wiarygodności

$$
\begin{aligned}
&\ell\left(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\beta}\right) \\
&= \sum_{i=1}^{n}\left[\left(1-y_i\right)\ln\left(\frac{1}{1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}}\right)+y_i\ln\left(\frac{e^{\boldsymbol{x}_i\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}}\right)\right] \\
&= \sum_{i=1}^{n} y_i\boldsymbol{x}_i\boldsymbol{\beta}-\sum_{i=1}^{n}\ln\left(1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}\right)
\end{aligned}
$$

- Efekt cząstkowy:

$$
\frac{\partial\,\mathrm{E}\left(y|\,\boldsymbol{x}\right)}{\partial\boldsymbol{x}}=\frac{\partial\Lambda\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)}{\partial\boldsymbol{x}}=\Lambda\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)\left[1-\Lambda\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)\right]\boldsymbol{\beta}
$$

- W przypadku modelu logitowego można także bezpośrednio interpretować same parametry.

- Zdefiniujmy szansę sukcesu jako stosunek prawdopodobieństwa sukcesu do prawdopodobieństwa porażki:

$$Odds\left(\boldsymbol{x}_i\right)=\frac{\Pr\left(y_i=1|\,\boldsymbol{x}_i\right)}{\Pr\left(y_i=0|\,\boldsymbol{x}_i\right)}=\frac{\frac{e^{\boldsymbol{x}_i\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}}}{\frac{1}{1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}}}=e^{\boldsymbol{x}_i\boldsymbol{\beta}}$$

Przykład Jeśli prawdopodobieństwo zdania egzaminu z ekonometrii jeśli się w ogóle nie uczyłem $(x_i=0)$ wynosi $\frac{1}{9}$ a szansa zdania egzaminu jest równa $Odds\left(0\right)=\frac{\frac{1}{9}}{1-\frac{1}{9}}=\frac{1}{8}$

- Policzmy teraz iloraz, szansy dla zmiennej niezależnej równej $\boldsymbol{x}_i$ i szansy dla zmiennej niezależnej równej $\boldsymbol{x}_j$:

$$\frac{Odds\left(\boldsymbol{x}_i\right)}{Odds\left(\boldsymbol{x}_j\right)}=\exp\left[\left(\boldsymbol{x}_i-\boldsymbol{x}_j\right)\boldsymbol{\beta}\right]=\exp\left(\Delta\boldsymbol{x}\boldsymbol{\beta}\right)$$

Przykład Jeśli prawdopodobieństwo zdania egzaminu z ekonometrii jeśli się uczyłem jeden dzień $(x_i = 1)$ wynosi $\frac{1}{3}$ to szansa zdania egzaminu w tym przypadku będzie równa $Odds\,(1) = \frac{\frac{1}{3}}{1-\frac{1}{3}} = \frac{1}{2}$ a iloraz szans będzie równy

$$\frac{Odds\,(1)}{Odds\,(0)} = \frac{\frac{1}{2}}{\frac{1}{8}} = 4$$

Prawdopodobieństwo zdania egzaminu wzrosło więc 3 razy a szansa zdania egzaminu 4 razy.

- Ilorazy szans równe $\exp\left(\boldsymbol{\beta}_k\right)$ interpretuje się w sposób następujący:
  - Dla zmiennej ciągłej jest to współczynnik, przez który należy przemnożyć szansę sukcesu dla x_{ik}, by otrzymać szansę sukcesu dla $x_{ik} + 1$.

– Dla zmiennej zerojedynkowej x_{ik} jest to współczynnik, przez który należy przemnożyć szansę sukcesu dla $x_{ik} = 0$, by otrzymać szansę sukcesu dla $x_{ik} = 1$.

- Dla małych $\boldsymbol{\beta}_k$ wielkości $\boldsymbol{\beta}_k$ można także zinterpretować, korzystając z tego, że $e^{\Delta \boldsymbol{x}\boldsymbol{\beta}} \approx 1 + \Delta \boldsymbol{x}\boldsymbol{\beta}$:
  – dla zmiennych ciągłych - ile procent zmieni się szansa sukcesu jeśli x_k zmieni się o 1
  – dla zmiennych zerojedynkowych - ile procent zmieni się szansa sukcesu jeżeli zmienna x_k przeskoczy z 0 na 1.

Przykład Zależność aktywności zawodowej od wieku i wykształcenia: kontynuacja

- Współczynniki

```
Logit estimates                                   Number of obs   =      47860
                                                  LR chi2(9)      =    8275.22
                                                  Prob > chi2     =     0.0000
Log likelihood = -28310.607                       Pseudo R2       =     0.1275

------------------------------------------------------------------------------
   _Iworks_1 |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wiek |  -.0186039   .0006038   -30.81   0.000    -.0197873   -.0174204
   _Iwyksz_2 |  -.5221753   .0655267    -7.97   0.000    -.6506053   -.3937453
   _Iwyksz_3 |  -.6942371   .0389682   -17.82   0.000    -.7706134   -.6178608
   _Iwyksz_4 |  -1.561855   .0471004   -33.16   0.000     -1.65417    -1.46954
   _Iwyksz_5 |  -.8542947   .0372593   -22.93   0.000    -.9273216   -.7812678
  _Iwyksz_60 |  -4.372065   .1570768   -27.83   0.000     -4.67993     -4.0642
  _Iwyksz_61 |  -2.207483   .0391494   -56.39   0.000    -2.284215   -2.130752
  _Iwyksz_70 |   -2.88925   .1303508   -22.17   0.000    -3.144732   -2.633767
  _Iwyksz_71 |  -4.194055    .416003   -10.08   0.000    -5.009406   -3.378704
```

```
       _cons |   1.666699   .0423531    39.35   0.000     1.583688    1.749709
------------------------------------------------------------------------------
```

- Krańcowe efekty

```
Marginal effects after logistic
      y  = Pr(_Iworks_1) (predict)
         =  .38568338
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
    wiek |  -.0044078      .00014  -30.94   0.000  -.004687 -.004129   43.9481
_Iwyks~2*|  -.1148651      .06553   -1.75   0.080  -.243295  .013565   .027267
_Iwyks~3*|  -.1539078      .03897   -3.95   0.000  -.230284 -.077532    .19135
_Iwyks~4*|  -.2864085       .0471   -6.08   0.000  -.378724 -.194093   .082804
_Iwyks~5*|   -.189642      .03726   -5.09   0.000  -.262669 -.116615    .26728
_Iwyk~60*|   -.398101      .15708   -2.53   0.011  -.705966 -.090236   .020079
_Iwyk~61*|  -.4279628      .03915  -10.93   0.000  -.504694 -.351231   .290639
_Iwyk~70*|  -.3641805      .13035   -2.79   0.005  -.619663 -.108698   .020664
_Iwyk~71*|  -.3810388        .416   -0.92   0.360  -1.19639  .434312   .004952
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

- Interpretacja efektów cząstkowych:

– wzrost wieku o 1 rok zmniejsza prawdopodobieństwo posiadania pracy o 0.4% w stosunku do prawdopodobieństwa dla średnich w próbie
– osoby z wykształceniem policealnym mają o 11% mniejsze prawdopodobieństwo posiadania pracy niż osoby z wykształceniem wyższym (poziom bazowy).

- Ilorazy szans

```
Logistic regression                               Number of obs   =      47860
                                                  LR chi2(9)      =    8275.22
                                                  Prob > chi2     =     0.0000
Log likelihood = -28310.607                       Pseudo R2       =     0.1275

------------------------------------------------------------------------------
   _Iworks_1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wiek |   .9815681   .0005927   -30.81   0.000     .9804071    .9827304
   _Iwyksz_2 |   .5932287   .0388723    -7.97   0.000     .5217299    .6745259
   _Iwyksz_3 |   .4994553   .0194629   -17.82   0.000     .4627292    .5390964
   _Iwyksz_4 |   .2097466   .0098792   -33.16   0.000     .1912507    .2300313
   _Iwyksz_5 |   .4255833   .0158569   -22.93   0.000     .3956119    .4578252
  _Iwyksz_60 |   .0126251   .0019831   -27.83   0.000     .0092797    .0171767
  _Iwyksz_61 |   .1099771   .0043055   -56.39   0.000      .101854     .118748
  _Iwyksz_70 |   .0556179   .0072498   -22.17   0.000     .0430784    .0718075
  _Iwyksz_71 |    .015085   .0062754   -10.08   0.000     .0066749    .0340916
------------------------------------------------------------------------------
```

- Interpretacja współczynników:

- szansę posiadania pracy trzeba pomnożyć przez 0.98 aby uzyskać szansę posiadania pracy przy wieku większym o 1 - spadek o 2%
- szanse posiadania pracy dla wykształcenia wyższego (poziom bazowy) należy pomnożyć przez 0.58 aby uzyskać szansę posiadania pracy przy wykształceniu policealnym (poziom 1) - spadek o 42%

- Sukcesy i porażki przewidywane i zaobserwowane

```
  Logistic model for _Iworks_1

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |     12568          7211  |      19779
     -     |      7205         20876  |      28081
-----------+--------------------------+-----------
   Total   |     19773         28087  |      47860

Classified + if predicted Pr(D) >= .5
True D defined as _Iworks_1 != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   63.56%
Specificity                     Pr( -|~D)   74.33%
Positive predictive value       Pr( D| +)   63.54%
Negative predictive value       Pr(~D| -)   74.34%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   25.67%
False - rate for true D         Pr( -| D)   36.44%
False + rate for classified +   Pr(~D| +)   36.46%
```

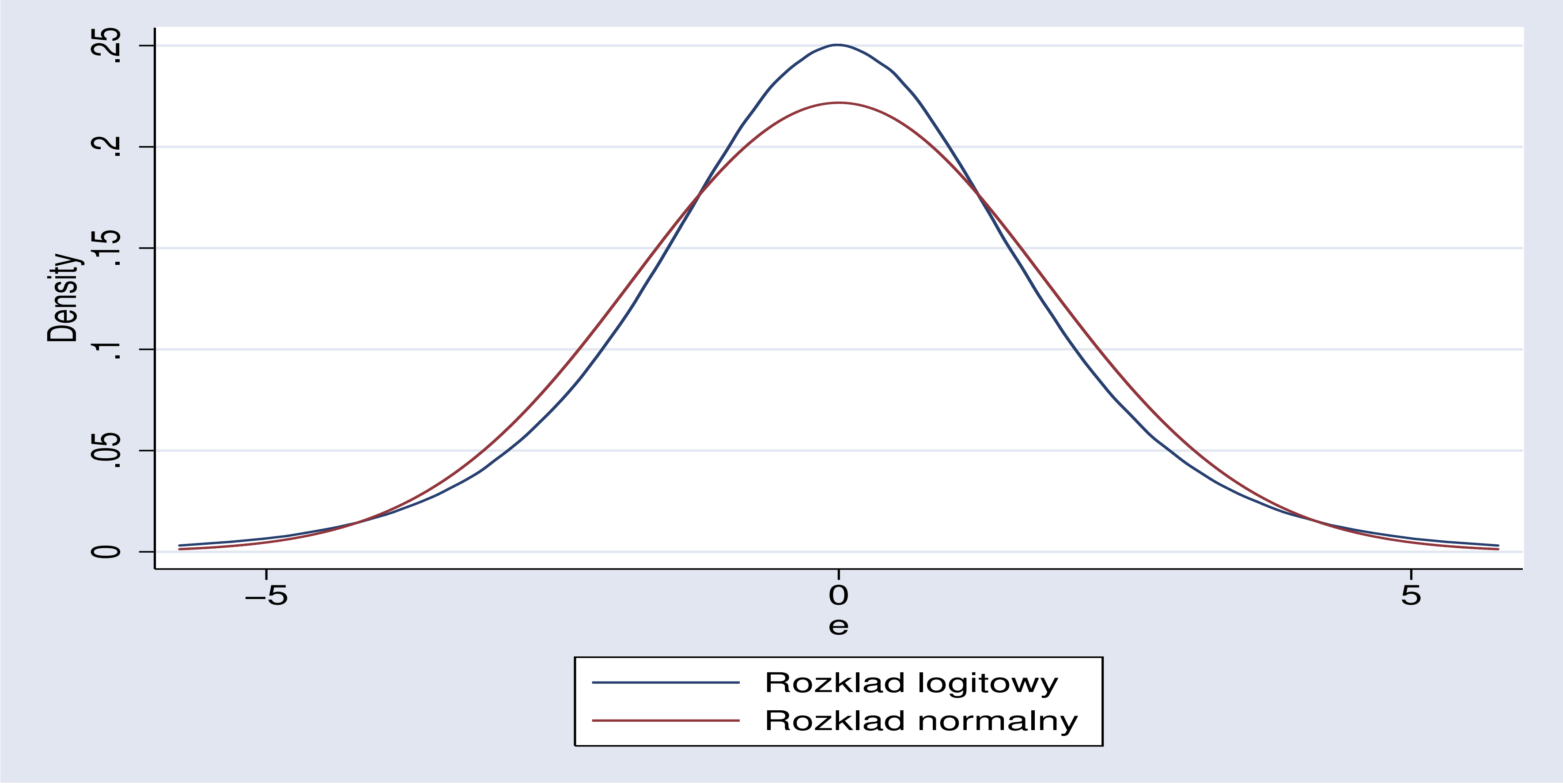
```
False - rate for classified -   Pr( D| -)   25.66%
--------------------------------------------------
Correctly classified                        69.88%
--------------------------------------------------
```

- To na którą z tych policzonych częstości powinniśmy patrzeć zależy od zastosowania:
  - czy szczególnie ważne jest wyeliminowanie potencjalnych porażek (np. kontrola jakości)
  - czy szczególnie ważne jest uniknięcie fałszywej klasyfikacji (np. identyfikacja przestępców)

Różnice między probitem a logitem

- Inna interpretacja ale interpretacja efektów cząstkowych taka sama

- Oba rozkłady prawdopodobieństwa są symetryczne jednak rozkład logistyczny ma nieco grubsze ogony.



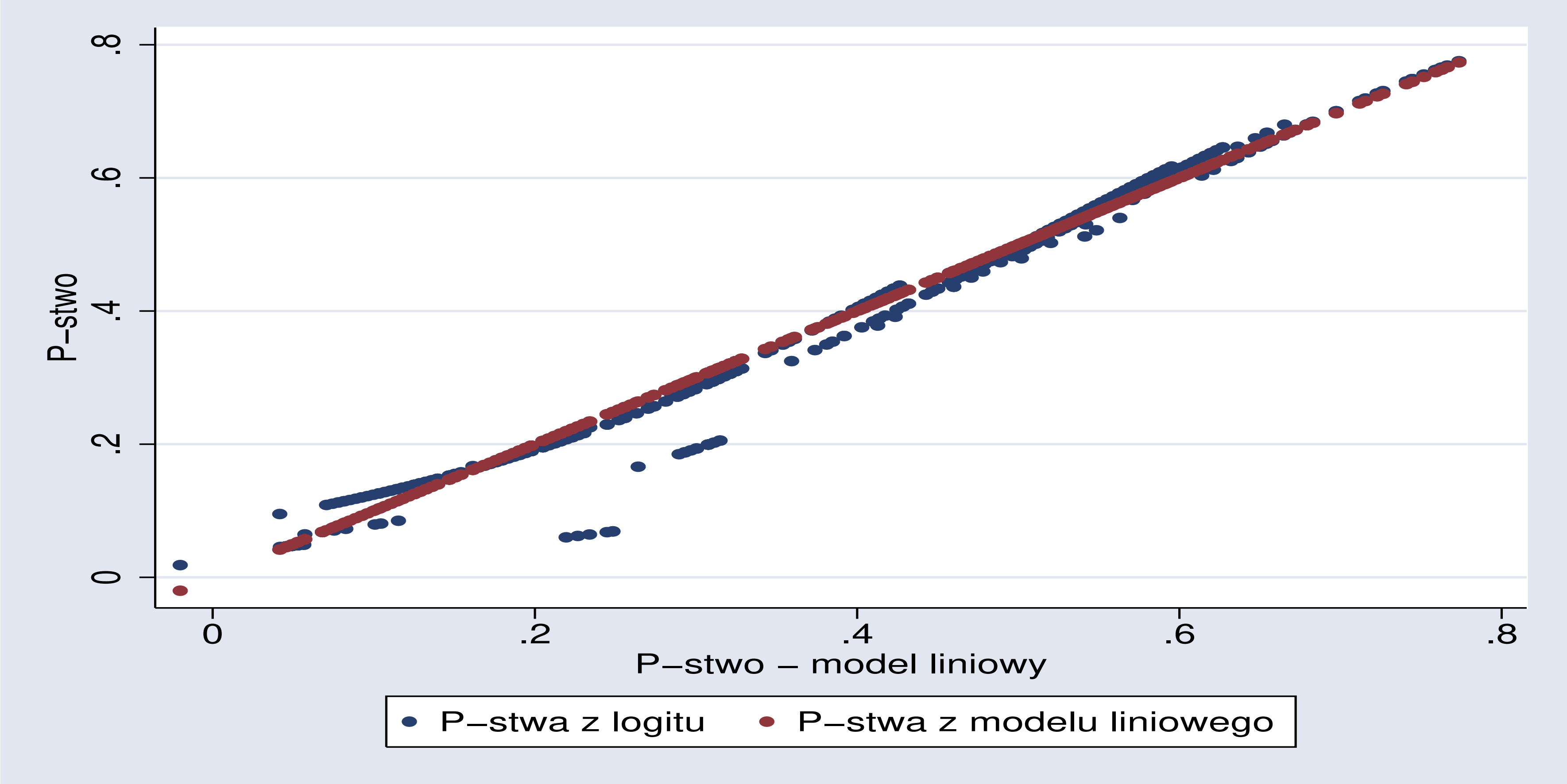
- W związku z tym istotne różnice między modelami będą powstawać dla prób, o nikłym odsetku odpowiedzi 0 albo odpowiedzi 1 i bardzo zróżnicowanych zmiennych niezależnych.

- Dla $\overline{\boldsymbol{x}}\boldsymbol{\beta}$ bliskiego 0 funkcja gęstości $f(0) \approx .4$ dla logitu, $f(0) \approx .25$ dla probitu, i $g(0) = 1$ dla LPM. Z tego powodu często zdarza się, że $\boldsymbol{\beta}_{\text{logit}} \approx 1.6\boldsymbol{\beta}_{\text{probit}}$ i $\boldsymbol{\beta}_{\text{logit}} \approx 4\boldsymbol{\beta}_{\text{LPM}}$.

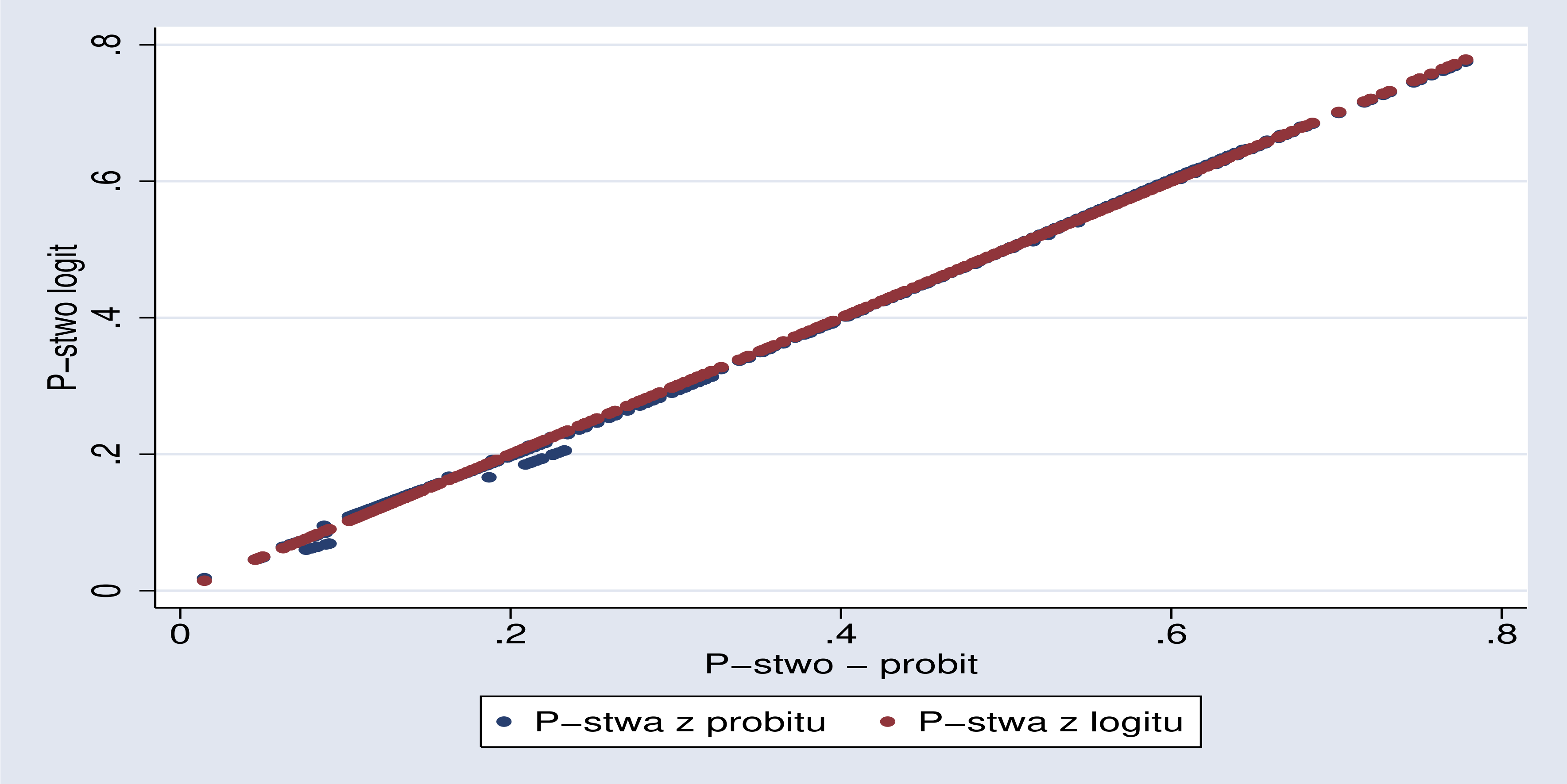
- Nie ma dobrej statystyki, która mogłaby posłużyć do wyboru między tymi modelami.

- W praktyce wybieramy ten, który jest analitycznie bardziej wygodny.

- Porównanie przewidywanych prawdopodobieństw z modelu liniowego i logitowego



- Porównanie przewidywanych prawdopodobieństw z modelu probitowego i logitowego



Modele dla zmiennych uporządkowanych

- Zmienna zależna przyjmuje wartości ze zbioru $\{0, 1, \ldots, J\}$ i wartości te można uporządkować (n.p. oceny z egzaminu, poziomy wykształcenia itp.).

- Uporządkowany probit

$$y^* = x\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

i dystrybuanta $\boldsymbol{\varepsilon}$ jest dana przez $F(\boldsymbol{\varepsilon})$. Zakładamy, że istnieje J *nieznanych* punktów przecięć $\boldsymbol{\alpha}_1 < \boldsymbol{\alpha}_2 < \ldots < \boldsymbol{\alpha}_J$

$$\begin{array}{lll} y = 0 & \text{jeśli} & y^* \leq \boldsymbol{\alpha}_1 \\ y = 1 & \text{jeśli} & \boldsymbol{\alpha}_1 < y^* \leq \boldsymbol{\alpha}_2 \\ \vdots & & \vdots \\ y = J & \text{jeśli} & y^* > \boldsymbol{\alpha}_J \end{array}$$

Prawdopodobieństwa poszczególnych wyborów są równe

$$
\begin{aligned}
\Pr\left(y=0|\,\boldsymbol{x}\right) &= \Pr\left(y^{*}\leq\boldsymbol{\alpha}_{1}|\,\boldsymbol{x}\right)=F\left(\boldsymbol{\alpha}_{1}-\boldsymbol{x}\boldsymbol{\beta}\right)\\
\Pr\left(y=1|\,\boldsymbol{x}\right) &= \Pr\left(\boldsymbol{\alpha}_{1}<y^{*}\leq\boldsymbol{\alpha}_{2}|\,\boldsymbol{x}\right)=F\left(\boldsymbol{\alpha}_{2}-\boldsymbol{x}\boldsymbol{\beta}\right)-F\left(\boldsymbol{\alpha}_{1}-\boldsymbol{x}\boldsymbol{\beta}\right)\\
&\vdots\\
\Pr\left(y=J|\,\boldsymbol{x}\right) &= \Pr\left(y^{*}>\boldsymbol{\alpha}_{J}|\,\boldsymbol{x}\right)=1-F\left(\boldsymbol{\alpha}_{J}-\boldsymbol{x}\boldsymbol{\beta}\right)
\end{aligned}
$$

- Te prawdopodobieństwa możemy użyć do zdefiniowania funkcji wiarygodności

- Dla uporządkowanego probitu $F\left(\boldsymbol{\varepsilon}\right)=\Phi\left(\boldsymbol{\varepsilon}\right)$ a dla uporządkowanego logitu $F\left(\boldsymbol{\varepsilon}\right)=\Lambda\left(\boldsymbol{\varepsilon}\right)$

- Dla modelu dla zmiennych uporządkowanych efekt cząstkowy jest równy

$$\begin{aligned}\frac{\partial p_0(\boldsymbol{x})}{\partial \boldsymbol{x}} &= -\boldsymbol{\beta}_k f(\boldsymbol{\alpha}_1 - \boldsymbol{x}\boldsymbol{\beta}), \ \frac{\partial p_J(\boldsymbol{x})}{\partial \boldsymbol{x}} = \boldsymbol{\beta}_k f(\boldsymbol{\alpha}_J - \boldsymbol{x}\boldsymbol{\beta}) \\ \frac{\partial p_j(\boldsymbol{x})}{\partial \boldsymbol{x}} &= \boldsymbol{\beta}_k [f(\boldsymbol{\alpha}_{j-1} - \boldsymbol{x}\boldsymbol{\beta}) - f(\boldsymbol{\alpha}_j - \boldsymbol{x}\boldsymbol{\beta})], \quad 1 < j < J\end{aligned}$$

  Znaki efektów cząstkowych są zdeterminowane przez znaki $\boldsymbol{\beta}$ tylko dla $j = 0$ i $j = J$. Dla pośrednich wyborów znak zależy także od znaku różnicy $\boldsymbol{\phi}(\boldsymbol{\alpha}_{j-1} - \boldsymbol{x}\boldsymbol{\beta}) - \boldsymbol{\phi}(\boldsymbol{\alpha}_j - \boldsymbol{x}\boldsymbol{\beta})$

- Dla tych modeli możemy także policzyć procenty prawidłowych przewidywań (przewidywany wybór byłby tu wyborem najbardziej prawdopodobnym według modelu)

Przykład Czy pani minister Środa ma rację? (przemoc w rodzinie a religijność) Zmienna objaśniana: odpowiedź na pytanie Proszę powiedzieć, czy się Pan(i) zdecydowanie zgadza, czy też zdecydowanie się Pan(i) nie zgadza ze stwierdzeniem, że czasami jest konieczne, by porządnym laniem przywołać dziecko do porządku. Możliwe odpowiedzi: zdecydowanie się zgadzam, zgadzam się, nie zgadzam się, zdecydowanie się nie zgadzam, nie wiem. Zmienne objaśniające: jak często chodzi do kościoła, wykształcenie, region zamieszkania, liczba dzieci, czy respondent nadużywa alkoholu, czy respondent jest szczęśliwy (odpowiedzi nie wiem zostały pominięte)

```
Ordered probit estimates                          Number of obs   =       7141
                                                  LR chi2(19)     =     145.53
                                                  Prob > chi2     =     0.0000
Log likelihood = -9160.8296                       Pseudo R2       =     0.0079

------------------------------------------------------------------------------
    spanking |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      attend |  -.0212552   .0062126    -3.42   0.001    -.0334317   -.0090787
        educ |   .0186807   .0044628     4.19   0.000     .0099338    .0274276
 _Iregion8_2 |   -.015627   .0458017    -0.34   0.733    -.1053966    .0741426
 _Iregion8_3 |  -.0547608    .042694    -1.28   0.200    -.1384396    .0289179
 _Iregion8_4 |   -.022339   .0475385    -0.47   0.638    -.1155127    .0708348
```

```
 _Iregion8_5 |   -.106774   .0498113    -2.14   0.032    -.2044024   -.0091457
 _Iregion8_6 |  -.1000209   .0585965    -1.71   0.088    -.2148679    .0148261
 _Iregion8_7 |  -.1936478   .0560845    -3.45   0.001    -.3035713   -.0837243
 _Iregion8_8 |  -.1391899   .0460914    -3.02   0.003    -.2295274   -.0488525
      childs |   -.039004   .0087318    -4.47   0.000    -.0561181   -.0218899
  _Idrink1_1 |   .1151006   .0363647     3.17   0.002     .0438271    .1863741
  _Idrink1_2 |   .2352853    .039625     5.94   0.000     .1576217     .312949
_Ihapunhap_2 |  -.0617953   .0363861    -1.70   0.089    -.1331108    .0095201
_Ihapunhap_3 |  -.0669545   .0452502    -1.48   0.139    -.1556433    .0217344
_Ihapunhap_4 |   -.236623   .0808379    -2.93   0.003    -.3950625   -.0781835
_Ipgssy~1993 |    .101719   .0386854     2.63   0.009     .0258971    .1775409
_Ipgssy~1994 |  -.0125805   .0388896    -0.32   0.746    -.0888027    .0636417
_Ipgssy~1995 |   .0798399   .0390542     2.04   0.041     .0032951    .1563847
_Ipgssy~1997 |   .1065954   .0428122     2.49   0.013     .0226851    .1905057
-------------+----------------------------------------------------------------
       _cut1 |  -.8671028   .0866442          (Ancillary parameters)
       _cut2 |   .2583954   .0864078
       _cut3 |   1.243044   .0871357
------------------------------------------------------------------------------
```

- Rzeczywiście istnieje istotny statystycznie dodatni związek między prawdopodobieństwem odpowiedzi twierdzącej na pytanie, że zdecydowanie się zgadzam z pytaniem i ujemny, że zdecydowanie się nie zgadzam a tym jak często respondent chodzi do kościoła.

- Uwaga 1: Związek jest statystycznie istotny ale nie trudno odczytać bezpośrednio z wyników na ile jest względnie ważny.

```
Ordered probit estimates                          Number of obs   =       7792
                                                  LR chi2(18)     =    3671.76
                                                  Prob > chi2     =     0.0000
Log likelihood =  -13018.18                       Pseudo R2       =     0.1236

------------------------------------------------------------------------------
      degree |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0609912   .0042303    14.42   0.000        .0527    .0692824
        age2 |  -.0008239   .0000429   -19.19   0.000     -.000908   -.0007397
    _Isize_2 |   .4141903   .0532129     7.78   0.000      .309895    .5184856
    _Isize_3 |   .5087877   .0438097    11.61   0.000     .4229222    .5946532
    _Isize_4 |   .5838797   .0461496    12.65   0.000     .4934282    .6743312
    _Isize_5 |   .5355975   .0441367    12.13   0.000      .449091    .6221039
    _Isize_6 |    .571307   .0415982    13.73   0.000     .4897759    .6528381
```

```
    _Isize_7 |   .7573527   .0488191    15.51   0.000       .661669    .8530364
    _Isize_8 |   .8325512   .0432759    19.24   0.000       .747732    .9173703
   _Ipadeg_1 |   .4402944   .0753946     5.84   0.000      .2925236    .5880651
   _Ipadeg_2 |   .9115508   .0750879    12.14   0.000      .7643813     1.05872
   _Ipadeg_3 |   1.175989    .080232    14.66   0.000      1.018737     1.33324
   _Ipadeg_4 |   1.729912   .1116625    15.49   0.000      1.511057    1.948767
   _Ipadeg_5 |   1.676097    .102199    16.40   0.000      1.475791    1.876403
   _Ipadeg_6 |   1.612315   .0869192    18.55   0.000      1.441957    1.782674
   _Ipadeg_7 |   2.077365   .1787112    11.62   0.000      1.727097    2.427632
   _Ipadeg_8 |   2.153149   .1821429    11.82   0.000      1.796156    2.510143
   _Ipadeg_9 |    2.18002   .1007946    21.63   0.000      1.982466    2.377573
-------------+----------------------------------------------------------------
       _cut1 |  -1.382214   .1321335          (Ancillary parameters)
       _cut2 |   .0751018    .120416
       _cut3 |   1.595465   .1210239
       _cut4 |   2.457993   .1213689
       _cut5 |   2.552495     .12147
       _cut6 |   2.753064   .1217385
       _cut7 |    3.52315   .1230554
       _cut8 |   3.762048   .1235789
       _cut9 |   3.943192   .1241567
------------------------------------------------------------------------------
```

- Wnioski:
  - Istnieje nieliniowa zależność między wiekiem a poziomem wykształcenia
  - Istnieje dodatnia zależność między wielkością miasta i wykształceniem rodziców a wykształceniem respondenta

Modele dla wyborów nieuporządkowanych (wielomianowy logit)

- Zmienna zależne przyjmuje wartości ze zbioru $\{0, 1, \ldots, J\}$. Warttości tych nie da się uporządkować (np. wybór zawodu, wybór miejsca zamieszkania itd.)

- Najbardziej popularnym modelem w tym kontekście jest wielomianowy logit:

$$
\begin{aligned}
p_j(\boldsymbol{x}) &= \Pr(y = j|\,\boldsymbol{x}) = \frac{\exp\left(\boldsymbol{x}\boldsymbol{\beta}_j\right)}{1 + \sum_{h=1}^{J} \exp\left(\boldsymbol{x}\boldsymbol{\beta}_h\right)} \text{ dla } j = 1, \ldots, J \\
p_0(\boldsymbol{x}) &= \Pr(y = 0|\,\boldsymbol{x}) = \frac{1}{\left[1 + \sum_{h=1}^{J} \exp\left(\boldsymbol{x}\boldsymbol{\beta}_h\right)\right]} \text{ dla } j = 0
\end{aligned}
$$

- Poziom zero jest interpretowany jako poziom bazowy. Współczynniki interpretujemy w odniesieniu do poziomu bazowego.

- Najprościej intepretować współcznniki przy użyciu ilorazów szans (relatywnego szansy zdarzenia $y = j$ dla $\boldsymbol{x}_1$ względem szansy zdarzenia zdarzenia dla $\boldsymbol{x}_0$)

$$
\begin{aligned}
Odds_j\left(\boldsymbol{x}_i\right) &= \frac{p_r\left(\boldsymbol{x}_i\right)}{p_0\left(\boldsymbol{x}_i\right)} = \exp\left(\boldsymbol{x}\boldsymbol{\beta}_j\right) \\
\frac{Odds_j\left(\boldsymbol{x}_1\right)}{Odds_j\left(\boldsymbol{x}_0\right)} &= \exp\left[\left(\boldsymbol{x}_1 - \boldsymbol{x}_0\right)\boldsymbol{\beta}_j\right] = \exp\left(\Delta\boldsymbol{x}\boldsymbol{\beta}_j\right)
\end{aligned}
$$

- Dodanie $\boldsymbol{\beta}_j$ oznacza więc, że ilaraz szans rośnie.

- Możliwe jest także policzenie efektów cząstkowych dla prawdopodobieństw poszczególnych alternatyw (wzory są dosyć skomplikowane)

- Liczba parametrów w modelu wielomianowym jest w takim modelu często bardzo duża: równa $J \times K$.

Przykład Stan cywilny a wiek, płeć i poziom wykształcenia.

```
Multinomial logistic regression                   Number of obs   =       8894
                                                  LR chi2(16)     =    3779.69
                                                  Prob > chi2     =     0.0000
Log likelihood = -6972.1839                       Pseudo R2       =     0.2133

------------------------------------------------------------------------------
     marital |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
malz / konk  |
         age |   .0873834   .0030204    28.93   0.000     .0814636    .0933032
        educ |   .0998039   .0190436     5.24   0.000     .0624791    .1371287
     _Isex_2 |   2.015474   .2663872     7.57   0.000     1.493365    2.537583
_IsexXeduc_2 |  -.1583433   .0236057    -6.71   0.000    -.2046096   -.1120769
       _cons |  -3.083293   .2413295   -12.78   0.000     -3.55629   -2.610296
-------------+----------------------------------------------------------------
wdowi        |
         age |   .1903022   .0045226    42.08   0.000     .1814382    .1991663
```

```
        educ |   .0798383   .0305318     2.61   0.009      .019997    .1396796
     _Isex_2 |   3.665258   .3614685    10.14   0.000     2.956793    4.373723
_IsexXeduc_2 |  -.1435851    .034277    -4.19   0.000    -.2107668   -.0764034
       _cons |  -11.67067   .4211294   -27.71   0.000    -12.49607   -10.84527
-------------+----------------------------------------------------------------
rozwiedz     |
         age |   .1039853   .0046831    22.20   0.000     .0948066    .1131641
        educ |   .1544554   .0348377     4.43   0.000     .0861747    .2227361
     _Isex_2 |   2.917523   .4353736     6.70   0.000     2.064207     3.77084
_IsexXeduc_2 |  -.1719227   .0378221    -4.55   0.000    -.2460526   -.0977928
       _cons |  -7.721986   .4648842   -16.61   0.000    -8.633142   -6.810829
-------------+----------------------------------------------------------------
separacja    |
         age |   .0950069   .0078496    12.10   0.000     .0796219    .1103919
        educ |   .0836903   .0636173     1.32   0.188    -.0409973     .208378
     _Isex_2 |   2.245084   .7776824     2.89   0.004     .7208547    3.769313
_IsexXeduc_2 |  -.1169839   .0704979    -1.66   0.097    -.2551573    .0211894
       _cons |  -7.842432   .8270322    -9.48   0.000    -9.463385   -6.221479
------------------------------------------------------------------------------
(marital==kaw / panna is the base outcome)
```

marital	RRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
malz / konk						
age	1.091315	.0032962	28.93	0.000	1.084874	1.097795
educ	1.104954	.0210423	5.24	0.000	1.064472	1.146976
_Isex_2	7.504284	1.999045	7.57	0.000	4.452051	12.64907
_IsexXeduc_2	.8535567	.0201488	-6.71	0.000	.8149654	.8939755
wdowi						
age	1.209615	.0054706	42.08	0.000	1.19894	1.220385
educ	1.083112	.0330694	2.61	0.009	1.020198	1.149905
_Isex_2	39.06621	14.1212	10.14	0.000	19.23618	79.33847
_IsexXeduc_2	.8662471	.0296924	-4.19	0.000	.809963	.9264424
rozwiedz						
age	1.109584	.0051963	22.20	0.000	1.099446	1.119816
educ	1.167022	.0406564	4.43	0.000	1.089997	1.249491
_Isex_2	18.49542	8.052418	6.70	0.000	7.879044	43.41651
_IsexXeduc_2	.8420443	.0318479	-4.55	0.000	.7818811	.9068368
separacja						

```
         age |   1.099666    .008632    12.10   0.000     1.082878    1.116716
        educ |   1.087292   .0691706     1.32   0.188     .9598317    1.231679
     _Isex_2 |   9.441209   7.342262     2.89   0.004      2.05619    43.35029
_IsexXeduc_2 |   .8895995   .0627149    -1.66   0.097     .7747946    1.021415
------------------------------------------------------------------------------
(marital==kaw / panna is the base outcome)
```

Probabilistyczne modele wyboru optymalnego

- Jest możliwe zbudowanie modeli dla wyborów opartych na maksymalizacji użyteczności

- Załóżmy, że wybór, że użyteczność uzyskana dzięki wyborowi j dla osoby i jest równa

$$y_{ij}^* = \boldsymbol{x}_{ij}\boldsymbol{\beta} + a_{ij}$$

  gdzie a_{it} są nieobserwowalnymi czynnikami wpływającymi na wybór.

- Zakładamy, że $\boldsymbol{x}_{ij}$ różnią się między osobami ale także między alternatywami

- Osoba maksymalizuje swoją użyteczność dokonując wyboru alternatywy y_i, która maksymalizuje jej użytecznośc

$$y_i = \arg\max\left(y_{i1}^*, y_{i2}^*, \ldots, y_{iJ}^*\right)$$

- Można pokazać, że jeśli a_{ij} są niezależne dla $j = 1, \ldots, J$ i mają dystrybuantę daną $F\left(a_{ij}\right) = \exp\left[-\exp\left(-a_{ij}\right)\right]$ (dystrybuanta wartości ekstremalnej I rodzaju), to

$$\Pr\left(y_i = j \middle| \boldsymbol{x}_{ij}\right) = \frac{\exp\left(\boldsymbol{x}_{ij}\boldsymbol{\beta}\right)}{\sum_{h=1}^{J}\exp\left(\boldsymbol{x}_{ih}\boldsymbol{\beta}\right)}, \quad j = 1, \ldots, J$$

- Taki model nazywany jest warunkowym logitem.

- Różnice między warunkowym logitem a wielomianowym logitem:

– obserwujemy charakterystyki osób ale także charakterystyki wyborów
– wektor parametrów jest taki sam dla wszystkich alternatyw

Przykład Modelujemy wybory dotyczące sposobu dojazdu do pracy (samochód, autobus). Możliwe wybory charakteryzują się różnymi kosztami (bilet, benzyna+opłata parkingowa) oraz czasem dojazdu. Osoby badane różnią się także dochodami. Dane magą wyglądać następująco:

osoba	wybór	decyzja	koszt	czas	dochód
1	bus	1	2	25	2000
1	car	0	3	15	2000
2	bus	0	2	25	3000
2	car	1	3	15	3000
3	bus	1	2	25	1000
3	car	0	3	15	1000

- Zauważmy, że iloraz prawdopodobieństw dla alternatyw j i h jest rowny:

$$\frac{p\left(\boldsymbol{x}_{ij}\right)}{p\left(\boldsymbol{x}_{ih}\right)}=\frac{\exp\left(\boldsymbol{x}_{ij}\boldsymbol{\beta}\right)}{\exp\left(\boldsymbol{x}_{ih}\boldsymbol{\beta}\right)}=\exp\left[\left(\boldsymbol{x}_{ij}-\boldsymbol{x}_{ih}\right)\boldsymbol{\beta}\right]$$

  względne prawdopodobieństwo alternatyw zależy więc wyłącznie od charakterystyk tych alternatyw

- Oznacza to, że warunkowy logit oparty jest na założeniu o niezależności od zewnętrznych alternatyw (IIA **I**ndependence from **I**rrelevant **A**lternatives): dodanie lub zmiana charakterystyk trzeciej alternatywy nie zmienia relacji prawdopodobieństw.

- Tego typu założenie jest często nierealistyczne dla przypadku podobnych alternatyw.

Przykład (McFadden 1974) Załóżmy, że pasażerwie mogą wybierać między samochodem i czerwonym autobusem. Załóżmy, że prawdopodbieństwo wyboru tych dwóch alternatyw jest równe $p_{car} = p_{red\,bus} = \frac{1}{2}$ tak, że iloraz szans dla tych dwóch alternatyw jest równy $\frac{p_{car}}{p_{red\,bus}} = 1$. Dodajmy teraz trzecią alternatywę: niebieski autobus. Załóżmy, że dla psażerów nie ma znaczenia kolor autobusu i w związku z tym $p_{red\,bus} = p_{blue\,bus}$. Jednak by utrzymać założenie o niezależności ilorazu szans między samochodem i czerwonym autobusem musimy teraz przyjąć, że $p_{car} = p_{red\,bus} = \frac{1}{3}$ - wprowadzenie dodatkowego koloru autobusu prowadzi do znaczącego spadku prawdobododobieństwa podróży samochodem!

Modele dla liczebności

- Liczebność jest to zmienna, która przyjmuje nieujemne wartości calkowite (np. liczba dzieci, liczba strajków ciągu roku w danej firmie, liczba papierosów wypalanych dziennie itd.)

- Czasami liczebność może mieć jakąś (zależną od obserwacji) logiczną górną granice (np. liczba dzieci danej w rodzinie, które ukończyły wyższe studia nie może być wyższa niż ogólna liczba dzieci w tej rodzinie)

- W tych przypadkach zastosowanie modelu dla zmiennych uporządkowanych jest nieuprawnione, ponieważ zbiór możliwych wartości y nie jest z góry określony (naogół wyniki estymacji uzyskane tą metodą nie będą przy tym satysfakcjonujące)

- Najprostsza metoda estymacji w przypadku zmiennych to założyć, że $\mathrm{E}(y|\,\boldsymbol{x}) = \boldsymbol{x\beta}$ i estymować model przy użyciu OLS.

- Wadą tej metody jest fakt, że niektóre z dopasowanych $\widehat{y} = \boldsymbol{xb}$ mogą się okazać ujemne, co nie ma sensu, ponieważ liczebność musi być zawsze dodatnia.

- Inna możliwość to dokonanie na zmiennych przekształcenia logarytmicznego i założenie, że $\log \mathrm{E}[y|\,\boldsymbol{x}] = \boldsymbol{x\beta}$. Wartości dopasowane są w tym modelu równe $\widehat{y} = \exp(\boldsymbol{xb}) > 0$ i większe od zera. Te podejście jest jednak niemożliwe do zastosowanie jeśli dla istotnej części obserwacji $y = 0$ (np. liczba dzieci)

Model Poissona

- Najpopularniejszym modelem dla liczebności jest model Poissona

- Zakładamy, że warunkowa wartość oczekiwana y jest dana przez $\mathrm{E}\left(y|\,\boldsymbol{x}\right) = \boldsymbol{\mu}\left(\boldsymbol{x}\right)$ (najczęściej $\boldsymbol{\mu}\left(\boldsymbol{x}\right) = \exp\left(\boldsymbol{x}\boldsymbol{\beta}\right)$)

- Warunkowa dystrybuanta y jest dane przez rozkład Poissona dla $\boldsymbol{\lambda} = \boldsymbol{\mu}\left(\boldsymbol{x}\right)$

$$f\left(y|\,\boldsymbol{x}\right) = \frac{e^{-\boldsymbol{\lambda}}\boldsymbol{\lambda}^{y}}{y!} = \frac{\exp\left[-\boldsymbol{\mu}\left(\boldsymbol{x}\right)\right]\left[\boldsymbol{\mu}\left(\boldsymbol{x}\right)\right]^{y}}{y!}$$

- Rzeczywiście dla rozkładu Poissona $\mathrm{E}\left(y|\,\boldsymbol{x}\right) = \boldsymbol{\lambda} = \boldsymbol{\mu}\left(\boldsymbol{x}\right)$

- Efekty cząstkowe w rozkładzie Poissona z $\boldsymbol{\mu}(\boldsymbol{x}) = \exp(\boldsymbol{x}\boldsymbol{\beta})$ intepretujemy tak samo jak w modelu, w którym zmienna zależna jest jest zlogarytmowana a niezależne nie są zlogarytmowane (semielastyczności wartości oczekiwanej)

$$\frac{\partial \ln \mathrm{E}(y|\,\boldsymbol{x})}{\partial \boldsymbol{x}} = \frac{\partial \ln \boldsymbol{\mu}(\boldsymbol{x})}{\partial \boldsymbol{x}} = \boldsymbol{\beta}$$

a więc $\frac{\Delta \boldsymbol{\mu}(\boldsymbol{x})}{\boldsymbol{\mu}(\boldsymbol{x})} \Big/ \Delta x_j = \boldsymbol{\beta}_j$.

- Jeśli zmienne niezależne są także zlogarytmowane to parametry będziemy intepretować jako elastyczności wartości oczekiwanej.

- Parametry tego modelu mogą być zgodnie i efektywnie oszacowane za pomocą MNW

- **Najważniejsze ograniczenie modelu Poissona**. Dla dystrybuanty rozkładu Poissona:

$$\mathrm{E}\left(y|\,\boldsymbol{x}\right) = \mathrm{Var}\left(y|\,\boldsymbol{x}\right) = \boldsymbol{\lambda} = \boldsymbol{\mu}\left(\boldsymbol{x}\right)$$

- Założenie to bardzo często nie jest spełnione w praktyce (rzadko kiedy można na podstawie rozważań teoretycznych uzasadnić rowność wartości oczekiwanej i wariancji)

- Założenie o równości wartości oczekiwanej i wariancji można przetestować.

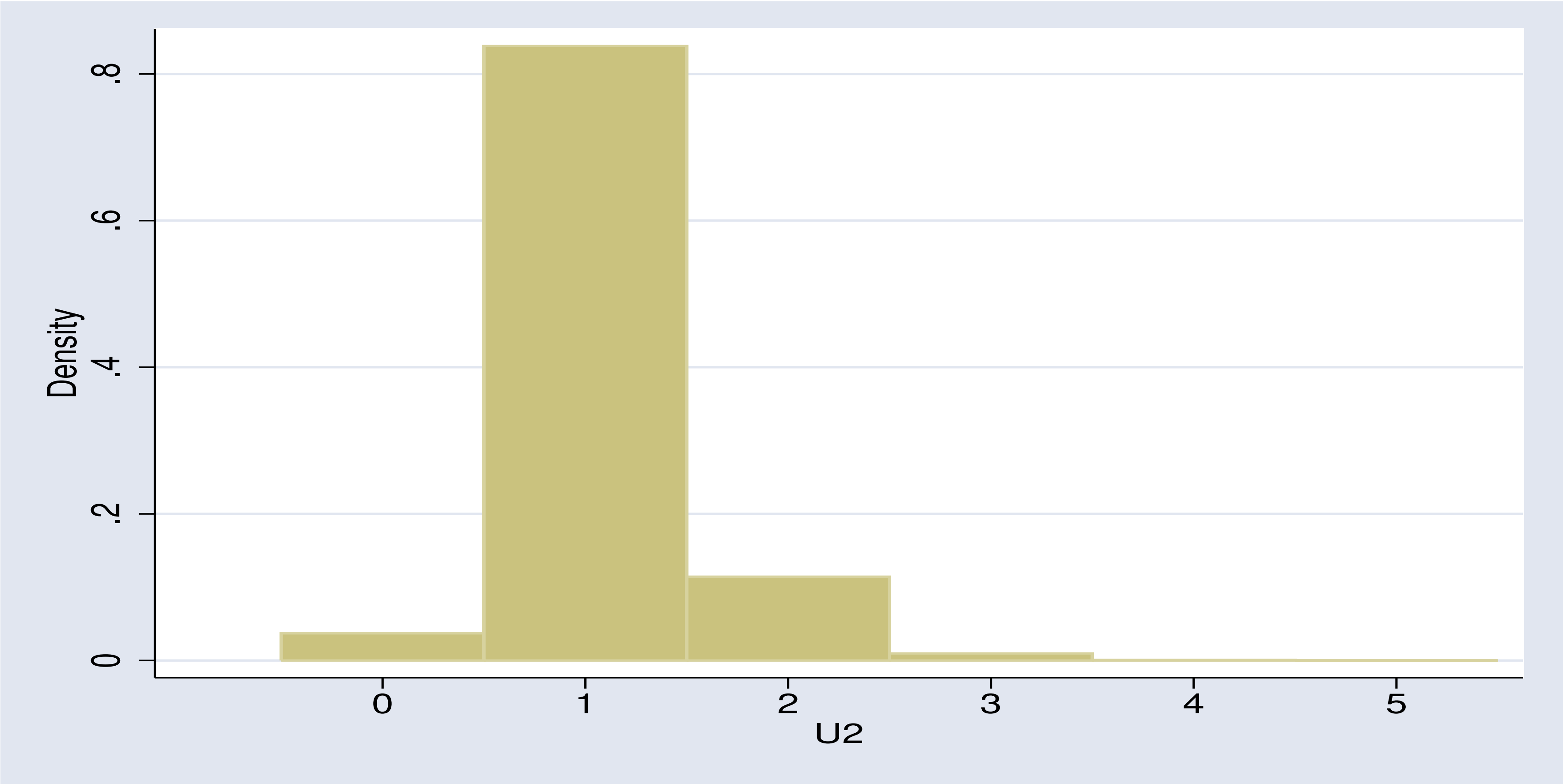
- Można pokazać, jeśli tylko dobrze jest wyspecyfikowany model dla warunkowej wartości oczekiwanej ($\mathrm{E}\left(y|\,\boldsymbol{x}\right) = \boldsymbol{\mu}\left(\boldsymbol{x}\right)$), to nawet dla $\mathrm{Var}\left(y|\,\boldsymbol{x}\right) \neq \boldsymbol{\mu}\left(\boldsymbol{x}\right)$, model Poissona będzie dawał zgodne oszacowania parametrów.

- W takim przypadku oszacowanie macierzy wariancji kowariancji będzie jednak nieprawidłowe. Można jednak w tym przypadku zastosować macierz odporną wariancji kowariancji ($Pseudo - MNW$)

- Czasami zdarza się, że obserwujemy zbyt wysoką liczbę zer w danych (np. przy liczbie wypalanych papierosów wiąże się ro z tym, że ludzie napierw wybierają palenie bądź nie a później, to ile będą palić). W takim przypadku powinno się stosować model ZIP (**Z**ero **I**nflated **P**oisson - rozdęty w zerze model Poissona)

Przykład Liczba telewizorów w gospodarstwie domowym a logarytm dochodu i liczba osób w rodzinie. Założono, że $\boldsymbol{\mu}(\boldsymbol{x}) = \exp(\boldsymbol{x}\boldsymbol{\beta})$.

- Liczba telewizorów w gospodarstwie domowym (dene z budżetów gospodarstw domowych)



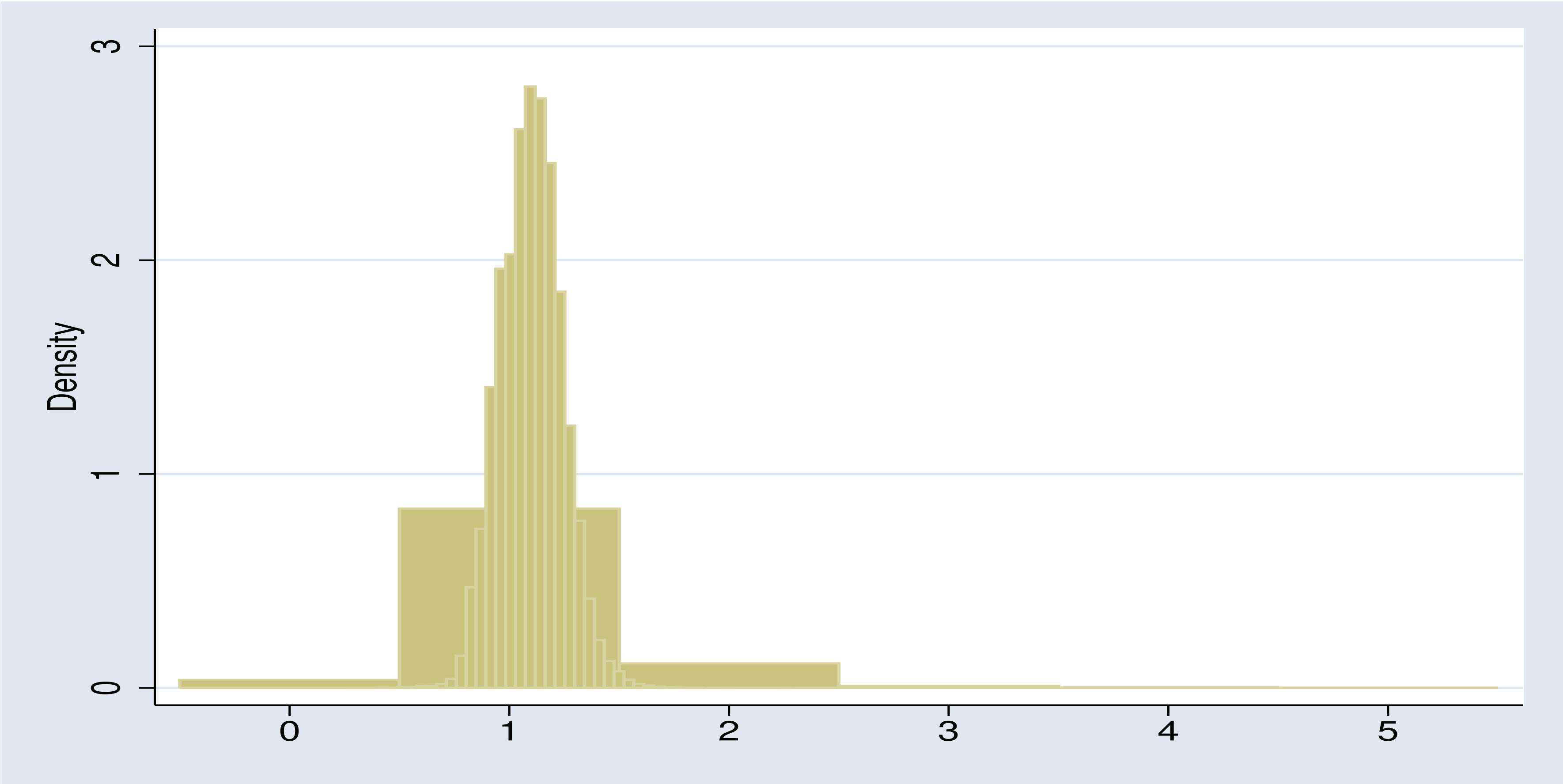
```
Poisson regression                                Number of obs   =      35952
                                                  LR chi2(7)      =     655.67
                                                  Prob > chi2     =     0.0000
Log likelihood = -39036.892                       Pseudo R2       =     0.0083
```

tv	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ldochg	.15787	.00923	17.10	0.000	.1397795	.1759604
los	.0289983	.0035735	8.11	0.000	.0219944	.0360021
_Iklm_2	-.0091403	.0206934	-0.44	0.659	-.0496986	.0314181
_Iklm_3	-.0225221	.02266	-0.99	0.320	-.0669348	.0218906
_Iklm_4	.001201	.0183397	0.07	0.948	-.0347441	.0371461
_Iklm_5	.0097115	.0199079	0.49	0.626	-.0293072	.0487301
_Iklm_6	-.0873394	.0178009	-4.91	0.000	-.1222285	-.0524503
_cons	-1.137438	.0686759	-16.56	0.000	-1.272041	-1.002836

- Intepretacja parametrów:
  - zwiększenie się dochodu o 1% powoduje zwiększenie wartości oczekiwanej liczby telewizorów w gospodarstwie o 0.15%

– zwiększenie się liczby osób o 1 zwiększa liczbę telewizorów o 2.9%
– zamieszkanie na wsi zmniejsza oczekiwaną liczbę telewizorów o 8.7%

- Liczba telewizorów w gospodarstwie domowym (liczebności rzeczywiste i dopasowane)



Model tobitowy

- Modele tobitowy używany jest w dwóch różnych kontekstach:
  - cenzurowania danych (*censored data*)
  - rozwiązań brzegowych (*corner solutions*)

- Dane ocenzurowane to 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 wszytskich pracowników pewnej firmy. Jednak dla płac otrzymaliśmy dane dotyczące jedynie szeregowych pracowników a w stosunku do osób na kierowniczych stanowiskach wiemy jedynie, że zarabiają więcej niż XXX.

- W wielu przypadkach wynikiem decyzji podmiotów ekonomicznych jest albo zero albo jakaś liczba dodatnia (np. wydatki na dobro trwałego użytku, wydatki na ubezpieczenie, wydatki na badania i rozwój)

- Z tego powodu w próbie możemy zaobserwować dodatnie wartości zmiennej oraz pewną liczbę zer, które obserwujemy dla tych podmiotów, dla których wybór $y = 0$ był optymalny

- Tego rodzaju rozwiązania problemu optymalizacji funkcji celu nazywamy rozwiązaniami brzegowymi

- W tego rodzaju przypadku zastosowanie liniowej bądź nieliniowej regresji da nieprawidłowe wyniki, ponieważ:
  - w regresji liniowej $y = \boldsymbol{x}\boldsymbol{\beta} + \varepsilon$ nie ma sposobu by zagwarantować, by wszystkie wartości dopasowane były dodatnie
  - za pomocą regresji $\ln y = \boldsymbol{x}\boldsymbol{\beta} + \varepsilon$ nie da się modelować obserwacji $y = 0$

- w regresji nieliniowej $y = \exp(\boldsymbol{x\beta}) + \varepsilon$, zazwyczaj wystąpi heteroskedastyczność
- w żadnej z tych regresji nie da się oszacować dwóch ważnych własności y:
  * $P(y > 0|\, \boldsymbol{x})$ - prawdopodobieństwa, że nie zajdzie rozwiązanie brzegowe
  * $\mathrm{E}(y|\, \boldsymbol{x}, y > 0)$ - warunkowej wartości oczekiwanej y jeśli $y > 0$ (nie jest rozwiązaniem brzegowym)

- Ważne jest zauważenie, że choć dane otrzymane w wyniku ocenzurowania i dane związane z rozwiązaniami brzegowymi wyglądają identycznie, to jednak są różnią się sposobem powstawania. W pierwszym przypadku obserwujemy prawdziwe wartości jedynie dla części obserwacji, w drugim przypadku obserwujemy prawdziwe wyniki decyzji dla wszystkich obserwacji.

- Model tobitowy definije się w sposób następujący:

$$y^* = \boldsymbol{x\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\varepsilon} \sim N(0,1)$$

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

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

- W zależności od przypadku różne interesować nas będą różne własności tego modelu:
  - dla zmiennych ocenzurowanych będzej nas interesować $\mathrm{E}\left(y^*|\,\boldsymbol{x}\right) = \boldsymbol{x\beta}$ (model dla zmiennej ukrytej)
  - dla zmiennej związanej z rowiązaniami brzegowymi, interesuje nas zazwyczaj $P\left(y > 0|\,\boldsymbol{x}\right)$ lub $\mathrm{E}\left(y|\,\boldsymbol{x},y > 0\right)$, 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:

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

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

- Efekty cząskowe dla poszczególnych tych wartości są równe:

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

- Znak $\boldsymbol{\beta}_k$ jest taki sam jak znak wszystkich rodzajów efektów cząstkowych - można pokazać, że $1-\boldsymbol{\lambda}\left(\boldsymbol{x\beta}/\,\boldsymbol{\sigma}\right)\left[\boldsymbol{x\beta}/\,\boldsymbol{\sigma}+\boldsymbol{\lambda}\left(\boldsymbol{x\beta}/\,\boldsymbol{\sigma}\right)\right]>0$

- Ostatni wynik ma następującą intepretację: jeśli prawdopodobieństwo uzyskania $y=0$ jest bliskie zeru $P\left(y^*>0|\,\boldsymbol{x}\right)=\Phi\left(\boldsymbol{x\beta}/\,\boldsymbol{\sigma}\right)\approx 1$, 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 $\boldsymbol{\beta}_{\text{MNK}} \approx \Phi\left(\overline{\boldsymbol{x}}\boldsymbol{\beta}/\,\boldsymbol{\sigma}\right)\boldsymbol{\beta} \approx \frac{\text{liczba obserwacji } y_i>0}{\text{liczba obserwcji}} \times \boldsymbol{\beta}$.

- Oszacowania uzyskane z modelu liniowego zastosowanego do danych ocenzurowanych daje obciążony estymator

  prawdziwego parametru $\boldsymbol{\beta}$ przy czym obciążenie to rośnie wraz ze wzrostem udziału zer w próbie.

- Estymator MNK nie jest zgodnym estymatorem $\boldsymbol{\beta}$ zarowno dla przypadku, kiedy używamy wszystkich obserwacji jak i dla przypadku, kiedy szacujemy model używając jedynie $y_i > 0$.

- Model tobitowy szacujemy za pomocą MNW. Funkcja gęstości dla poszczególnych obserwacji ma postać:

$$
\begin{aligned}
f\left(y_i \mid \boldsymbol{x}_i\right) &= \begin{cases} 1-\Phi\left(\boldsymbol{x}_i\boldsymbol{\beta}/\,\boldsymbol{\sigma}\right) & \text{dla} \quad y_i=0 \\ \frac{1}{\boldsymbol{\sigma}}\boldsymbol{\phi}\left(\frac{y_i-\boldsymbol{x}_i\boldsymbol{\beta}}{\boldsymbol{\sigma}}\right) & \text{dla} \quad y_i>0 \end{cases} \\
&= \left\{1-\Phi\left(\boldsymbol{x}_i\boldsymbol{\beta}/\,\boldsymbol{\sigma}\right)\right\}^{1[y_i=0]}\left[\frac{1}{\boldsymbol{\sigma}}\boldsymbol{\phi}\left(\frac{y_i-\boldsymbol{x}_i\boldsymbol{\beta}}{\boldsymbol{\sigma}}\right)\right]^{1[y_i>0]}
\end{aligned}
$$

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
Pseudo R2	= 0.0309

Log likelihood = -53588.396

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 intepretacji.

- 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 $\frac{\partial\,\mathrm{E}(y|\boldsymbol{x})}{\partial \boldsymbol{x}}$
     – oczekiwane wydatki gospodarstwa na obuwie wzrosną o 0.58% jeśli dochód wzrośnie o 1%
  3. efekty cząstkowe dla warunkowej wartości oczekiwanej $\frac{\partial\,\mathrm{E}(y|\boldsymbol{x},y>0)}{\partial \boldsymbol{x}}$
     – dla gospodarsw, które poniosły w danym miesiącu jakikolwiek wydatki na obuwie, oczekiwane wydatki gospodarstwa na obuwie wzrosną o 0.44% jeśli dochód wzrośnie o 1%
  4. prawdopodobieństwo, że gospodarstwo poniesie jakiekolwiek wydatki na obuwie wzrośnie o 0.13% jeśli dochód wzrośnie o 1%

- Uwagi: zauważmy że całkowity wpływ wzrostu dochodu na wydatki gospodarstw na obuwie jest w przybliżeni równy sumie wzrostu wydatków tych gospodarstw, które wydają coś na obuwie i wzrostu proporcji ych gospodarstw w ogóle populacji.

Selekcja próby

- W wielu przypadkach próba, którą dysponujemy nielosowa.
- Możliwe są trzy przypadki:
  - selekcja próby oparta na zmiennej niezależnej
  - selekcja próby oparta na zmiennej zależnej
  - selekcja próby oparta na osobnym mechanizmie (samoselekcja)

Przykład (selekcja oparta na zmiennej niezależnej) W próbie BAEL obserwujemy jedynie zarobki osób w wieku >16 lat

Przykład (selekcja oparta na zmiennej zależnej) W próbie Z12 obserwujemy płace jedynie osób, których produktywność jest wyższa niż płaca minimalna - pozostałe nie są przyjmowane do pracy (zakładamy, że wszyscy bezrobotni chcą pracować - nie mogą tego jednak legalnie robić ponieważ są zbyt mało produktywni).

Przykład Oferowana płaca. Oferowana płaca jest to funkcja, która opisuje wysokość płac dla wszytkich osób, łącznie z tymi, ktore nie pracują. Dane dotyczące płac posiadamy jednak jedynie w odniesieniu do osób o które pracują, przy czym wybór dotyczący tego czy pracować, czy nie związany jest z innym rówaniem - równaniem uczestnictwa w rynku pracy. Zakładamy, że osoby, ktore nie pracują robią to dobrowolnie - ich oczekiwania dotyczące płacy są wyższe niż to co oferują im pracodawcy.

- **Selekcja oparta na zmiennej niezależnej**
  – jeśli zmienna niezależna, która determinuje selekcje do próby jest niezależna od błędu losowego, wtedy estymator β uzyskany z regresji liniowej przeprowadzonej dla nielosowej próby *jest zgodny*

Selekcja oparta na zmiennej zależnej

- Różnica w stosunku do tobita: w przypadku tobita cenzurowanie dotyczy jedynie wartości zmiennej zależnej, zmienne niezależne obserwowane są dla wszystkich obserwacji.

- W przypadku selekcji próby obserwujemy *jedynie* obserwacje wyselekcjonowane - takie próby nazywa się niekiedy próbami uciętymi (*truncated samples*)

- Załóżmy, że do próby wybrane mogły być jedynie obserwacje, dla których $a_1 < y < a_2$

- Stosując twierdzenie Bayesa uzyskujemy, że dystrybyanta y jest dana przez

$$\begin{aligned}\Pr\left(y \leq Y \middle| \boldsymbol{x}, a_1 < y < a_2\right) &= \frac{\Pr\left(y \leq Y, a_1 < y_i < a_2 \middle| \boldsymbol{x}\right)}{\Pr\left(a_1 < y < a_2 \middle| \boldsymbol{x}\right)} \\ &= \frac{F\left(Y \middle| \boldsymbol{x}_i\right) - F\left(a_1 \middle| \boldsymbol{x}\right)}{F\left(a_2 \middle| \boldsymbol{x}\right) - F\left(a_1 \middle| \boldsymbol{x}\right)}\end{aligned}$$

ponieważ

$$\Pr\left(a_1 < y < a_2 \middle| \boldsymbol{x}\right) = \Pr\left(a_1 < y < a_2\right) = F\left(a_2 \middle| \boldsymbol{x}\right) - F\left(a_1 \middle| \boldsymbol{x}\right)$$

oraz

$$\Pr\left(y \leq Y, a_1 < y < a_2 \middle| \boldsymbol{x}\right) = \Pr\left(a_1 < y \leq Y \middle| \boldsymbol{x}\right) = F\left(Y \middle| \boldsymbol{x}_i\right) - F\left(a_1 \middle| \boldsymbol{x}\right)$$

- Wynika z tego, że funkcja gęstości y jest równa

$$\frac{\partial \Pr\left(y \leq Y \mid \boldsymbol{x}, a_1 < y < a_2\right)}{\partial Y} = \frac{f\left(Y \mid \boldsymbol{x}\right)}{F\left(a_2 \mid \boldsymbol{x}\right) - F\left(a_1 \mid \boldsymbol{x}\right)}$$

- Jeśli założymy, że $\Pr\left(y \leq Y \mid \boldsymbol{x}\right) = \Phi\left(\boldsymbol{x}\boldsymbol{\beta}\right)$ a więc jest dystrybuntą rozkładu normalnego, to otrzymamy tak zwany regresję uciętą z funkcją gęstości dla poszczególnej obserwacji równą

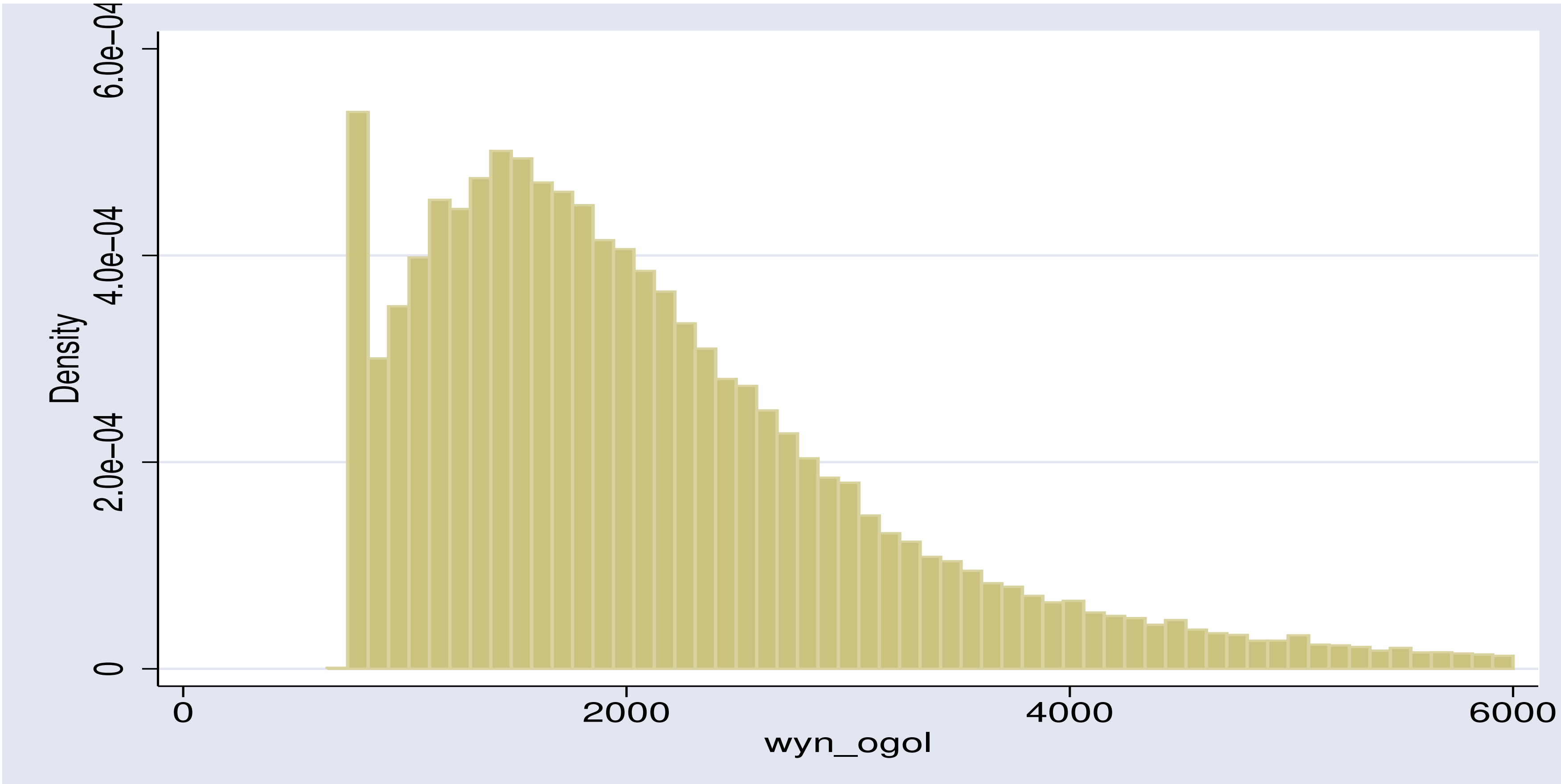
$$f\left(y_i \mid \boldsymbol{x}_i\right) = \frac{\frac{1}{\boldsymbol{\sigma}}\boldsymbol{\phi}\left(\frac{y_i - x_i\boldsymbol{\beta}}{\boldsymbol{\sigma}}\right)}{\Phi\left(a_2 \mid \boldsymbol{x}_i\right) - \Phi\left(a_1 \mid \boldsymbol{x}_i\right)}$$

- W zależności od zastosowania będą nas interesować efekty cząskowe dla wszystkich obserwacji $\frac{\partial \,\mathrm{E}(y|\boldsymbol{x})}{\partial x} = \boldsymbol{\beta}$, bądź efekty cząskowe dla populacji zaobserwowanej

$$\frac{\partial \,\mathrm{E}\left(y|\, \boldsymbol{x}, a_1 < y < a_2\right)}{\partial \boldsymbol{x}}$$

Przykład Funcja oferty płacy (funkcja popytu na pracę) liczona na podstawie danych z poziomu przedsiębiorstw. Dane polskie 2002 roku z badania Z12.

- Płaca brutto na podstawie danych z przedsiębiorstw



6.0e−04
4.0e−04
2.0e−04
0
Density
0
2000
4000
6000
wyn_ogol

- Na poziomie przedsiębiorstw obserwujemy jedynie płace powyżej płacy minimalnej

- Nie jest obserwowalna pełna populacja zatrudnionych a jedynie ci, który chcą pracować i ich produktywność jest wyższa od płacy minimalnej

```
Truncated regression
Limit:   lower =  6.6333184                            Number of obs =  63233
         upper =       +inf                            Wald chi2(9)  =20960.00
Log likelihood =   -35064.9                            Prob > chi2   = 0.0000

------------------------------------------------------------------------------
   lwyn_ogol |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
eq1          |
        wiek |   .0642988   .0016065    40.03   0.000     .0611502    .0674474
       wiek2 |  -.0006349   .0000196   -32.46   0.000    -.0006732   -.0005966
    _Iplec_2 |  -.3271812   .0044484   -73.55   0.000    -.3358998   -.3184626
_Ipoz_wyks_2 |  -.4024077   .0094504   -42.58   0.000    -.4209302   -.3838852
_Ipoz_wyks_3 |   -.436441   .0056581   -77.14   0.000    -.4475306   -.4253514
_Ipoz_wyks_4 |  -.3749871   .0085319   -43.95   0.000    -.3917094   -.3582648
```

```
_Ipoz_wyks_5 |  -.6870413   .0060402  -113.75   0.000    -.6988798   -.6752028
_Ipoz_wyks_6 |  -.8139249   .0085966   -94.68   0.000    -.8307739   -.7970759
_Ipoz_wyks_7 |  -.9548006   .0641728   -14.88   0.000    -1.080577   -.8290242
       _cons |   6.645369   .0314885   211.04   0.000     6.583653    6.707085
-------------+----------------------------------------------------------------
sigma        |
       _cons |   .4766114   .0016778   284.07   0.000     .4733231    .4798998
------------------------------------------------------------------------------
```

- Intepretacja parametrów w tym modelu jest identyczna jak w modelu liniowym i dotyczy ofert płacy w całej populacji

- Można także policzyć efekty cząskowe dla podpopulacji zatrudnionych

```
Marginal effects after truncreg
      y  = E(lwyn_ogol|lwyn_ogol>6.633318433280377) (predict, e(6.633318433280377,.))
         =  7.5877517
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
    wiek |   .0562668       .0014   40.23   0.000   .053525  .059008   40.1953
```

```
   wiek2 |  -.0005556      .00002  -32.57   0.000  -.000589 -.000522   1713.87
_Iplec_2*|   -.284661      .00379  -75.05   0.000  -.292096 -.277227   .489128
_Ipoz_~2*|  -.3164657      .00648  -48.83   0.000  -.329168 -.303763   .059431
_Ipoz_~3*|  -.3587534      .00429  -83.55   0.000   -.36717 -.350337   .258789
_Ipoz_~4*|  -.2984899      .00603  -49.51   0.000  -.310305 -.286675   .073395
_Ipoz_~5*|  -.5410371      .00409 -132.29   0.000  -.549053 -.533021   .265763
_Ipoz_~6*|  -.5650978      .00428 -132.00   0.000  -.573488 -.556707   .092562
_Ipoz_~7*|  -.5855983      .02254  -25.98   0.000  -.629781 -.541415   .001423
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

- Efekty cząskowe dla podpopulacji są bardzo bliskie oszacowaniom z MNK

```
      Source |       SS       df       MS              Number of obs =   65110
-------------+------------------------------           F(  9, 65100) = 2829.63
       Model |  5196.35059     9  577.372288           Prob > F      =  0.0000
    Residual |  13283.3214 65100  .204044875           R-squared     =  0.2812
-------------+------------------------------           Adj R-squared =  0.2811
       Total |   18479.672 65109   .28382669           Root MSE      =  .45171

------------------------------------------------------------------------------
   lwyn_ogol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wiek |    .056535   .0013053    43.31   0.000     .0539766    .0590935
       wiek2 |    -.00056   .0000161   -34.73   0.000    -.0005916   -.0005284
    _Iplec_2 |  -.2751897   .0037418   -73.55   0.000    -.2825236   -.2678558
_Ipoz_wyks_2 |   -.375408    .008194   -45.82   0.000    -.3914682   -.3593478
_Ipoz_wyks_3 |  -.4123654   .0050178   -82.18   0.000    -.4222003   -.4025305
_Ipoz_wyks_4 |  -.3625898    .007469   -48.55   0.000     -.377229   -.3479505
_Ipoz_wyks_5 |  -.6298677   .0050945  -123.64   0.000    -.6398529   -.6198824
_Ipoz_wyks_6 |  -.7247809   .0068707  -105.49   0.000    -.7382475   -.7113144
_Ipoz_wyks_7 |   -.808392   .0469905   -17.20   0.000    -.9004933   -.7162907
       _cons |   6.798214   .0252975   268.73   0.000     6.748631    6.847797
```

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 charakterystki wpływające na zmienną, którą modelujemy są skorelowane są z nieobserwowalnymi charakterystykami wpływającymi na selekcje do próby, to otzymane wyniki estymacji mogą być obciążone

Przykład (klasyczny) Decyzje o podjęciu pracy o oferowana płaca. Chcemy oszacować równanie wyjaśniające zależność oferowanej płacy od chrakterystyk 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 respondanta danych przez zmienne $\boldsymbol{x}_{1i}$ (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, kóre otrzymują rodzinę są silniej motywowane do podjęcia pracy nawet przy niskich stawkach).

1. założmy, że równanie oferowanej płacy ma postać

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

2. respondent podejmie pracę jeśli

$$w_i^o > w_i^r$$

założmy, że:

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

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

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

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

Uwagi:

- ponieważ $\boldsymbol{x}_i = (\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i})$, więc wszystkie zmienne wpływajace na równanie płac wpływają też równanie selekcji do póby

- 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 $\log w_i^o - \log w_i^r = \boldsymbol{x}_i \boldsymbol{\delta}_2 + v_{2i}$ są mieszaniną parametrów równania płac i równania selekcji więc nie można ich intepretować jako parametrów równania podaży pracy.

- Podstawowy model:

$$
\begin{aligned}
y_1 &= \boldsymbol{x}_1\boldsymbol{\beta} + u_1 && \text{(Rwnanie regresji)} \\
y^* &= \boldsymbol{x}\boldsymbol{\delta} + u_2 && \text{(Rwnanie selekcji)} \\
y_2 &= \begin{cases} 0 & \text{dla} \quad y^* \leq 0 \\ 1 & \text{dla} \quad y^* > 0 \end{cases}
\end{aligned}
$$

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

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

- Warunkowa wartość oczekiwana dla obserwacji wyselekcjonowanych

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

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

- Zwykłą regresja liniową regresję liniowa y_1 na $\boldsymbol{x}_1$, da niezgodne estymatory parametru $\boldsymbol{\beta}$ ponieważ pominięty został element $\boldsymbol{\sigma}\boldsymbol{\rho}\boldsymbol{\lambda}(\boldsymbol{x}\boldsymbol{\delta})$ (problem zmiennej pominiętej: $\boldsymbol{\sigma}\boldsymbol{\rho}\boldsymbol{\lambda}(\boldsymbol{x}\boldsymbol{\delta})$ jest zazwyczaj skorelowane z $\boldsymbol{x}_1$)

- Uwaga:
  - w szczególnym przypadku, gdy błędy losowe (charakterystyki nieobserwowalne) w obu równaniach są nieskorelowane ($\boldsymbol{\rho} = 0$) estymator MNK jest zgodny

- Zauważmy, że:
  - równanie selekcji opisuje zwykły model probitowy - z modelu probitowego można uzyskać oszacowania $\boldsymbol{\delta}$
  - w równaniu $y_1 = \boldsymbol{x}_1\boldsymbol{\beta} + \boldsymbol{\sigma}\boldsymbol{\rho}\boldsymbol{\lambda}(\boldsymbol{x}\boldsymbol{\delta}) + \boldsymbol{\varepsilon}$ błąd losowy $\boldsymbol{\varepsilon}$ jest nieskorelowany z $\boldsymbol{x}_1$ i $\boldsymbol{\lambda}(\boldsymbol{x}\boldsymbol{\delta})$ - parametr $\boldsymbol{\beta}$ można zgodnie oszacować MNK

- Dwustopniowa procedura Heckmana:

  1. estymujemy najpierw regresję probitową y_{2i} na $\boldsymbol{x}_i$, liczymy $\widehat{\boldsymbol{\lambda}}_{2i} = \boldsymbol{\lambda}\left(\boldsymbol{x}\widehat{\boldsymbol{\delta}}\right)$
  2. przeprowadzamy zwykłą regresję y_{2i} na $\boldsymbol{x}_{1i}$ i $\widehat{\boldsymbol{\lambda}}_{2i}$

- Uzyskane w ten sposób estymatory parametru $\boldsymbol{\beta}$ są zgodne, trudniej uzyskać jest prawidłowe oszacowania jego macierzy wariancji

- Istnienie problemu selekcji próby można przetestować testując istotność parametru $\gamma = \boldsymbol{\sigma\rho}$

- Estymator parametrów w modelu Heckmana można także uzyskać za pomocą MNW

- Uwaga:

  – formalnie biorąc model Heckmana można wyestymować nawet wtedy, gdy $\boldsymbol{x} = \boldsymbol{x}_1$ (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 rspondenta pojawiają się dochody pozostałych członków gospodastwa. Zakładamy, że wysokie dochody pozostałych członków gospodarstwa wpływają negatywnie na motywację do podjęcia pracy ale nie mają wpływu na poziom płacy. Do obu rownań dodano zmienne zerojedynkowe związane z rokiem badania.

- Wyniki regresji metodą dwustopniową

```
Heckman selection model -- two-step estimates   Number of obs      =      7895
(regression model with sample selection)        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
  lextincome |  -.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                           Number of obs      =       7895
(regression model with sample selection)          Censored obs       =       4490
                                                  Uncensored obs     =       3405

                                                  Wald chi2(14)      =    3803.36
Log likelihood = -6984.195                        Prob > chi2        =     0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lrincome     |
     _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
  lextincome |  -.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                      .2248619    .4370084
       sigma |   .5338063   .0087138                       .516998    .5511611
      lambda |   .1789196   .0310179                      .1181257    .2397134
------------------------------------------------------------------------------
LR test of indep. eqns. (rho = 0):   chi2(1) =    31.12   Prob > chi2 = 0.0000
------------------------------------------------------------------------------
```

- odrzucamy $H_0 : \boldsymbol{\rho} = 0$, ponieważ
  - dla metody dwustopniowej statystyki z dla oszacowania parametru $\boldsymbol{\rho\sigma}$ (przy zmiennej $\boldsymbol{\lambda}$) jest istotnie różna od zera
  - dla MNW na postawie statystyki LR

- Oszacowany współczynnik korelacji $\boldsymbol{\rho}$ między błędami losowymi z obu równań jest zgodnie z teorią dodatni

- Współczynniki przy zmiennych w równaniu regresji intepretujemy 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 prawdopodbieństwem selekcji
$$\frac{\partial \Pr\left(y_2 = 1\right)}{\partial \boldsymbol{x}} = \boldsymbol{\phi}\left(\boldsymbol{x}\boldsymbol{\beta}\right)\boldsymbol{\beta}$$

  – wielkość efektów cząstkowych dla y_1^+ gdzie

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

Przykład Kontynuacja

- Efekty cząątkowe 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(lrincome*|Pr(select)) (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 osiągane 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: jesli nie pracujemy do dochód z pracy równy jest zero
  - wielkość płacy jaką oferuje nam rynek