PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 176
Taksonomia 18. Klasyfikacja i analiza danych – teoria i zastosowania 2011

Tadeusz Kufel
Uniwersytet Mikołaja Kopernika w Toruniu

Marcin Błażejowski
Wyższa Szkoła Bankowa w Toruniu

Paweł Kufel
Uniwersytet Mikołaja Kopernika w Toruniu

MODELOWANIE ZMIENNYCH JAKOŚCIOWYCH I OGRANICZONYCH Z WYKORZYSTANIEM OPROGRAMOWANIA GRETL

Streszczenie: W artykule przedstawione zostały przykłady modelowania zmiennych ograniczonych, tj. binarnych, dyskretnych i licznikowych, z wykorzystaniem estymacji logitowej dla zmiennej: dwumianowej, wielomianowej uporządkowanej, wielomianowej nieuporządkowanej, estymacji tobitowej i heckitowej (dla zmiennej uciętej), regresji Poissona i regresji ujemnej dwumianowej (dla zmiennej licznikowej). Całość zilustrowana została przykładami i procedurami estymacji z zastosowaniem oprogramowania GRETL.

Słowa kluczowe: zmienne jakościowe, zmienne ograniczone, model logitowy, GRETL.

1. Wstęp

Celem artykułu jest przedstawienie kierunków badawczych ekonometrycznego modelowania zmiennych jakościowych i ograniczonych ze wskazaniem pomocniczych procedur w oprogramowaniu GRETL [Cottrell, Lucchetti 2011, s. 228-247]. Zmienne jakościowe i ograniczone nie posiadają postaci cech zmiennych ciągłych[1]. Zmienne jakościowe są to zmienne, których wartości mają postać niemierzalnych „kategorii" i mogą tworzyć następujące typy zmiennych jakościowych, przedstawionych w formie zmiennych dyskretnych skategoryzowanych:

[1] Zaprezentowane zagadnienia w artykule stanowią rozszerzenie możliwości obliczeniowych oprogramowania GRETL wersji 1.6.5 z maja 2007 roku, które zostały opisane w podręczniku: [Kufel 2007, rozdział 10].

– zmiennej dwumianowej, to jest zmiennej binarnej, typu zero-jedynkowego, np. 0 – kobieta, 1 – mężczyzna,
– zmiennej wielomianowej uporządkowanej, to jest zmiennej dyskretnej utworzonej z uporządkowanej klasyfikacji, np. wykształcenie: 0 – podstawowe, 1 – gimnazjalne, 2 – średnie, 3 – wyższe licencjackie, 4 – wyższe magisterskie,
– zmiennej wielomianowej nieuporządkowanej, to jest zmiennej dyskretnej utworzonej z nieuporządkowanej klasyfikacji, np. marki samochodów: 1 – Fiat, 2 – Opel, 3 – Audi, 4 – Citroen itd.

Zbiory danych, wśród których występują zmienne jakościowe, dotyczą najczęściej mikrodanych[2]. W badaniach ekonomicznych głównym źródłem mikrodanych są badania ankietowe klientów: banków, sklepów, salonów samochodowych, ponadto pracowników, gospodarstw domowych itp.

Zmienne jakościowe maja własności skal pomiaru: nominalnego (dwumianowa oraz wielomianowa nieuporządkowana) lub porządkowego (dwumianowa oraz wielomianowa uporządkowana)[3].

2. Modelowanie zmiennych jakościowych

Modelowanie endogenicznych zmiennych jakościowych w oprogramowaniu GRETL umożliwiają funkcje programu znajdujące się w menu pod nazwą: Model/Nieliniowe modele. Poszczególne kategorie zmiennych jakościowych można modelować za pomocą funkcji:

– Model logitowy/dwumianowy...
– Model logitowy/wielomianowy uporządkowany...
– Model logitowy/wielomianowy nieuporządkowany...

lub

– Model probitowy/dwumianowy...
– Model probitowy/wielomianowy uporządkowany...

Przykładowe oszacowane modele logitowe dla poszczególnych typów zmiennych jakościowych zostały wykonane dla zbioru danych dotyczących badań ankietowych przeprowadzonych wśród 94 gmin województwa kujawsko-pomorskiego w 2003 roku, dotyczących korzystania z funduszy pomocowych przedakcesyjnych Unii Europejskiej[4]. Rysunek 1 przedstawia oszacowany model logitowy dwumianowej zmiennej jakościowej dotyczącej odpowiedzi na pytanie: Czy gmina korzysta z funduszy pomocy przedakcesyjnej UE? (0 − nie, 1− tak), natomiast rys. 2 przedstawia oszacowany model logitowy wielomianowy uporządkowany dla zmiennej: liczba złożonych wniosków na dofinansowanie (0 − brak − 31%, 1 − wniosek − 54%,

[2] Szeroki opis modelowania mikrodanych zawiera monografia: [Gruszczyński 2010].

[3] Szerszy opis typów skal pomiarowych przedstawiono w pracy: [Walesiak, Gatnar 2009, s. 64-66].

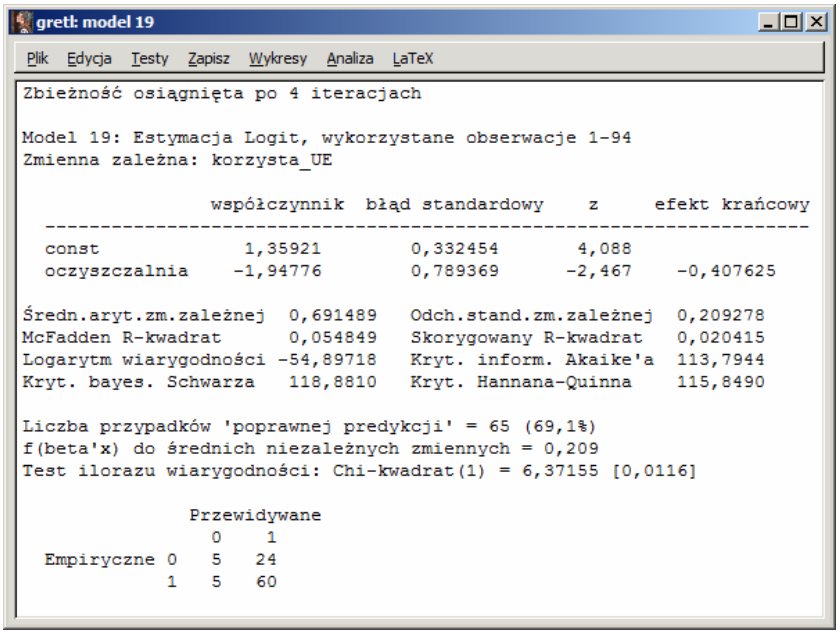
[4] Szerszy opis zbioru danych i analiz na tym zbiorze znajduje się w pracy: [Kufel 2007, s. 139-148]. Plik z danymi dostępny jest na stronie http://www.kufel.torun.pl.

2 − 10%, 3 − 5%). Ze zbioru wielu czynników istotny okazał się tylko czynnik informujący o udziale ludności korzystającej z oczyszczalni ścieków.

Oszacowane wartości *cut1*, *cut2* i *cut3*, zwane punktami odcięcia (*cut points*), wskazują progi przejścia dla kolejnych wartości nieobserwowalnej zmiennej endogenicznej *y**.

Obserwowana zmienna y_i jest zależna od ukrytej zmiennej y_i^* w następujący sposób:

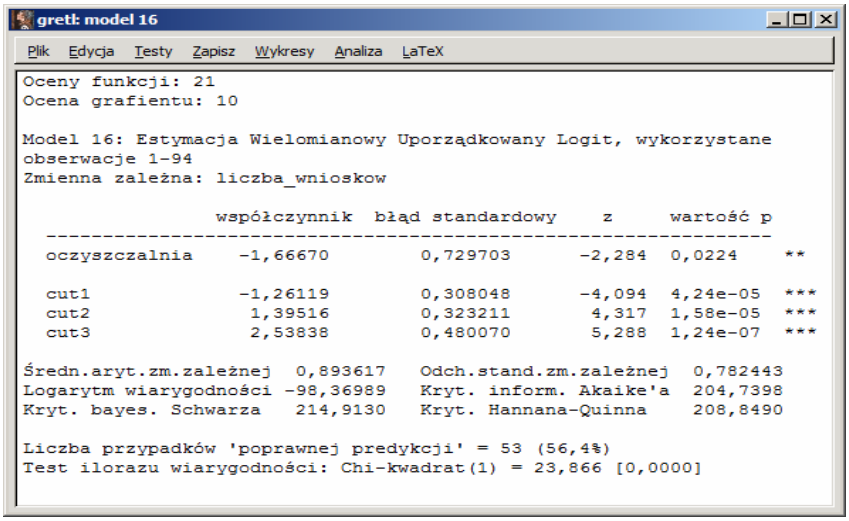
$$y_i^* = \begin{cases} 0, & y_i^* \le \gamma_1 \\ 1, & \gamma_1 < y_i^* \le \gamma_2 \\ 2, & \gamma_2 < y_i^* \le \gamma_3 \\ \dots & \dots \\ M, & \gamma_M < y_i^* \end{cases}.$$



Rys. 1. Oszacowany model logitowy dwumianowy

Źródło: opracowanie własne. Okno programu GRETL.

Dwuwartościowy model logitowy (probitowy) jest szczególnym przypadkiem dla *M* = 1. Najlepszym miernikiem oceny jakościowej modelu jest liczba przypadków ‘poprawnej predykcji’, które można wyznaczyć na podstawie tablicy trafności o wymiarach 2 × 2 z liczbą przypadków „trafionych” (główna przekątna) i „nietrafionych”.



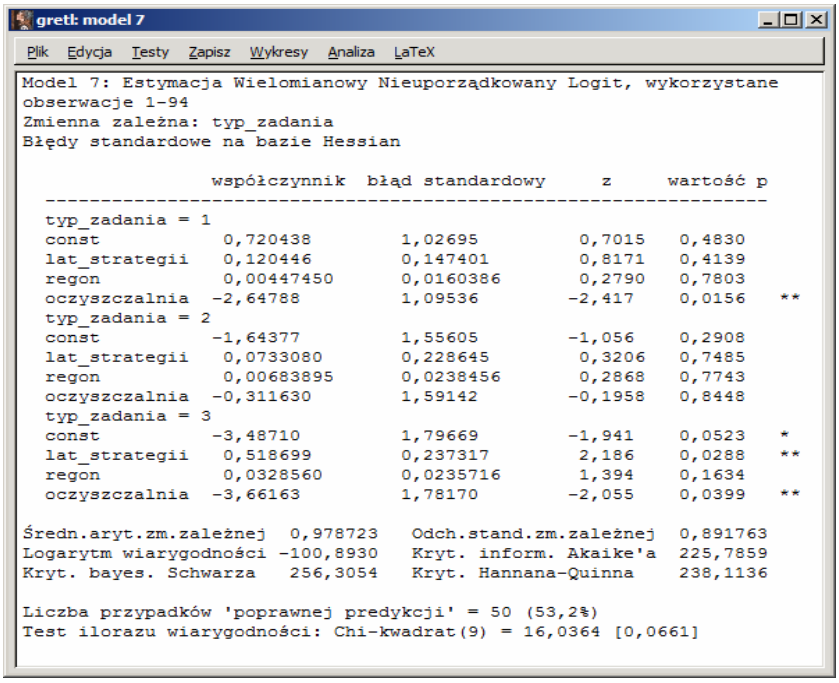
gretl: model 16

Plik Edycja Testy Zapisz Wykresy Analiza LaTeX

```
Oceny funkcji: 21
Ocena grafientu: 10

Model 16: Estymacja Wielomianowy Uporządkowany Logit, wykorzystane
obserwacje 1-94
Zmienna zależna: liczba_wnioskow

                współczynnik  błąd standardowy     z      wartość p
  ------------------------------------------------------------------
  oczyszczalnia    -1,66670        0,729703      -2,284  0,0224    **

  cut1             -1,26119        0,308048      -4,094  4,24e-05  ***
  cut2              1,39516        0,323211       4,317  1,58e-05  ***
  cut3              2,53838        0,480070       5,288  1,24e-07  ***

Średn.aryt.zm.zależnej  0,893617   Odch.stand.zm.zależnej  0,782443
Logarytm wiarygodności -98,36989   Kryt. inform. Akaike'a  204,7398
Kryt. bayes. Schwarza   214,9130   Kryt. Hannana-Quinna    208,8490

Liczba przypadków 'poprawnej predykcji' = 53 (56,4%)
Test ilorazu wiarygodności: Chi-kwadrat(1) = 23,866 [0,0000]
```

Rys. 2. Oszacowany model logitowy wielomianowy uporządkowany

Źródło: opracowanie własne. Okno programu GRETL.



gretl: model 7

Plik Edycja Testy Zapisz Wykresy Analiza LaTeX

```
Model 7: Estymacja Wielomianowy Nieuporządkowany Logit, wykorzystane
obserwacje 1-94
Zmienna zależna: typ_zadania
Błędy standardowe na bazie Hessian

                współczynnik  błąd standardowy     z      wartość p
  ------------------------------------------------------------------
  typ_zadania = 1
  const            0,720438        1,02695        0,7015   0,4830
  lat_strategii    0,120446        0,147401       0,8171   0,4139
  regon            0,00447450      0,0160386      0,2790   0,7803
  oczyszczalnia   -2,64788         1,09536       -2,417    0,0156   **
  typ_zadania = 2
  const           -1,64377         1,55605       -1,056    0,2908
  lat_strategii    0,0733080       0,228645       0,3206   0,7485
  regon            0,00683895      0,0238456      0,2868   0,7743
  oczyszczalnia   -0,311630        1,59142       -0,1958   0,8448
  typ_zadania = 3
  const           -3,48710         1,79669       -1,941    0,0523   *
  lat_strategii    0,518699        0,237317       2,186    0,0288   **
  regon            0,0328560       0,0235716      1,394    0,1634
  oczyszczalnia   -3,66163         1,78170       -2,055    0,0399   **

Średn.aryt.zm.zależnej  0,978723   Odch.stand.zm.zależnej  0,891763
Logarytm wiarygodności -100,8930   Kryt. inform. Akaike'a  225,7859
Kryt. bayes. Schwarza   256,3054   Kryt. Hannana-Quinna    238,1136

Liczba przypadków 'poprawnej predykcji' = 50 (53,2%)
Test ilorazu wiarygodności: Chi-kwadrat(9) = 16,0364 [0,0661]
```

Rys. 3. Oszacowany model logitowy wielomianowy nieuporządkowany

Źródło: opracowanie własne. Okno programu GRETL.

Rysunek 3 przedstawia oszacowany model logitowy wielomianu nieuporządkowanego dla zmiennej *typ_zadania*, która zawiera następujące warianty cechy nominalnej skategoryzowanej do czterech wariantów wskazujących na typ wniosku o dofinansowanie zadań gminy z funduszy UE (0 − brak – 30%; 1 − ochrona środowiska – 50%; 2 − komunikacyjne – 10%, 3 − inne lub kilka wniosków – 10%).

Oszacowany model wskazuje poziom istotności czynników dla każdego wariantu – kategorii osobno. Powyższe przykłady można oszacować za pomocą modelu probitowego w oprogramowaniu GRETL – menu: Model/Nieliniowe modele/Model Probitowy.

3. Modelowanie zmiennych ograniczonych

Przykładami zmiennej ograniczonej (szerzej w: [Maddala 2006, s. 349-392; Gruszczyński 2002, s. 34-39; Gruszczyński 2010, s. 193-216]) są zmienne: cenzurowana i ucięta, które stanowią kombinację zmiennej jakościowo-ilościowej.

Obserwowana zmienna y_i jest zależna od ukrytej zmiennej y_i* w następujący sposób:

$$y_i = \begin{cases} y_i^* = \beta\, x_i + u, & y_i^* > 0 \\ 0, & y_i^* \le 0 \end{cases},$$

gdzie u_i to niezależny składnik resztowy o rozkładzie normalnym. Powyższy model jest nazywany modelem tobitowym.

Przykład modelu tobitowego[5] przedstawia rys. 4. Został on oszacowany dla zmiennej opisującej udział dofinansowania zadań przez UE za pomocą funkcji: Model/Nieliniowe modele/Model Tobitowy... Wśród badanych 94 gmin tylko 39 korzysta z dofinansowania, 55 gmin jest bez dofinansowania − zmienna ta przyjmuje wartość zero.

Oszacowany model tobitowy posiada składnik resztowy o rozkładzie, który nie ma cech rozkładu normalnego.

Odmiennym sposobem szacowania modelu dla zmiennej uciętej jest budowa modelu selekcji próby [Heckman 1979] o następującej postaci:

– równanie główne: $y_{1i} = \beta x_i + \sigma u_1$,

– równanie selekcji: $y_{2i}^* = \gamma z_i + u_2$, gdzie $y_{2i}^* = \begin{cases} 1, & y_{2i}^* \ge C \\ 0, & y_{2i}^* < C \end{cases}$.

Rysunek 5 przedstawia model selekcji próby szacowany metodą największej wiarygodności, tzw. model heckitowy, dla którego zmienną selekcji y_{2i} jest zmienna zero-jedynkowa *zamierza_UE*. Powyższy model oszacowano za pomocą funkcji: Model/Nieliniowe modele/Model z selekcją próby (hecki)...

[5] Pełny opis przykładu w: [Kufel 2007, s. 135-137].

gretl: model 1

Plik Edycja Testy Zapisz Wykresy Analiza LaTeX

```
Oceny funkcji: 27
Ocena grafientu: 10

Model 1: Estymacja Tobit, wykorzystane obserwacje 1-94
Zmienna zależna: dofinans_udzial

                współczynnik  błąd standardowy    z     wartość p
  ------------------------------------------------------------------
  const            0,115772       0,0967550      1,197   0,2315
  oczyszczalnia   -0,752370       0,393488      -1,912   0,0559   *

Średn.aryt.zm.zależnej  0,223308   Odch.stand.zm.zależnej  0,294005
Cenzurowane obserwacje        55   Sigma (Se)              0,570930
Logarytm wiarygodności -67,01215   Kryt. inform. Akaike'a  140,0243
Kryt. bayes. Schwarza   147,6542   Kryt. Hannana-Quinna    143,1062

Test na normalność rozkładu reszt -
  Hipoteza zerowa: składnik losowy ma rozkład normalny
  Statystyka testu: Chi-kwadrat(2) = 37,5848
  z wartością p = 6,89547e-009
```

Rys. 4. Oszacowany model tobitowy

Źródło: opracowanie własne. Okno programu GRETL.

gretl: model 11

Plik Edycja Testy Zapisz Wykresy Analiza LaTeX

```
Oceny funkcji: 53
Ocena grafientu: 23

Model 11: Estymacja ML Heckit, wykorzystane obserwacje 1-94
Zmienna zależna: dofinans_udzial
Zmienna selekcji: zamierza_UE

                współczynnik  błąd standardowy     z     wartość p
  ------------------------------------------------------------------
  const            0,407821       0,100126       4,073   4,64e-05  ***
  oczyszczalnia   -0,264700       0,139450      -1,898   0,0577    *
  lambda          -0,170237       0,126209      -1,349   0,1774

                    Równanie selekcji próby

  const           -0,0283015      0,168524      -0,1679  0,8666
  lat_strategii    0,138494       0,0707131      1,959   0,0502    *

Średn.aryt.zm.zależnej  0,235006   Odch.stand.zm.zależnej  0,300034
Sigma (Se)              0,314831   Autokorel.reszt - rho1 -0,540732
Logarytm wiarygodności -71,73860   Kryt. inform. Akaike'a  149,4772
Kryt. bayes. Schwarza   155,4442   Kryt. Hannana-Quinna    151,7784

Liczba obserwacji: 94
Cenzurowane obserwacje: 40 (42,6%)
```

Rys. 5. Oszacowany model selekcji próby

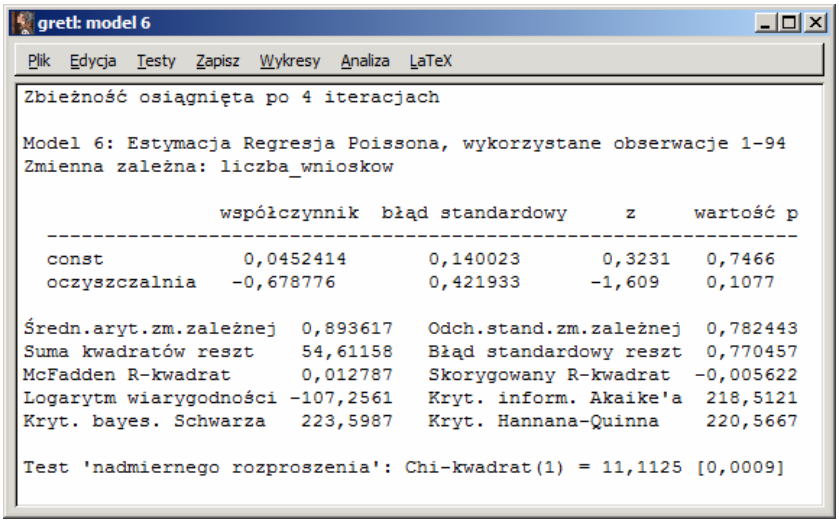
Źródło: opracowanie własne. Okno programu GRETL.

4. Modelowanie zmiennej licznikowej

Przykładem zmiennej licznikowej jest zmienna dyskretna dodatnia informująca o liczbie zdarzeń, na przykład liczba złożonych wniosków przez gminę o dofinansowanie ze środków UE [Kufel 2007, s. 129-130]. Rozkład tej zmiennej (*liczba wniosków*) dla 94 obserwacji jest następujący: 0 (brak wniosków) dla 29 gmin, 1 wniosek dla 51 gmin, 2 wnioski dla 9 gmin oraz 3 dla 5 gmin.

Rysunek 6 przedstawia oszacowany model zmiennej licznikowej (*count data*) za pomocą regresji Poissona, dla której opisywaną zmienną jest liczba wniosków. Powyższy model oszacowano za pomocą funkcji: Model/Nieliniowe modele/Model zmiennej licznikowej (Count data)...

Test „nadmiernego rozproszenia" (*overdispersion*) wskazuje na zjawisko dużego rozproszenia, co oznacza, że model nie potrafi objaśnić zmienność y_i. W przypadku gdyby wariancja (0,7824) była większa od wartości oczekiwanej (0,8936), należałoby zastosować uogólniony model regresji Poissona, to jest model regresji ujemnej dwumianowej (NegBin 1 lub NegBin 2).



Rys. 6. Oszacowany model zmiennej licznikowej

Źródło: opracowanie własne. Okno programu GRETL.

Prezentowany model ma wyznaczone wskaźniki dobroci: współczynnik R-kwadrat McFaddena, logarytm wiarygodności oraz kryteria informacyjne.

5. Podsumowanie

Oprogramowanie GRETL pozwala na oszacowanie modeli dla specjalnych zmiennych, np. binarnych, dyskretnych oraz licznikowych. Zaprezentowane przykłady pozwalają stwierdzić, że analiza zmiennych jakościowych i ograniczonych, jako mikrodanych pochodzących z badań ankietowych, za pomocą modeli dla logitowej zmiennej binarnej, logitowej uporządkowanej, logitowej wielomianowej (nieuporządkowanej), probitowej, tobitowej, heckitowej, Poissona, ujemnej dwumianowej (dla zmiennej licznikowej) umożliwia sformułowanie nowych wniosków badawczych.

Literatura

Cottrell A., Lucchetti R. „Jack", *Gretl User's Guide. Gnu Regression, Econometrics and Time-series Library*, February 2011, http://gretl.sf.net.

Gruszczyński M. (red.), *Mikroekonometria*, Wydawnictwo Wolters Kluwer, Warszawa 2010.

Gruszczyński M., *Modele i prognozy zmiennych jakościowych w finansach i bankowości*, Oficyna Wydawnicza Szkoły Głównej Handlowej, Warszawa 2002.

Heckman J.J., *Sample Selection Bias as a Specification Error*, „Econometrica" 1979, vol. 47, 1, s. 153-162.

Kufel T., *Ekonometria. Rozwiązywanie problemów z wykorzystaniem oprogramowania GRETL*, Wydawnictwo Naukowe PWN, wyd. 2, Warszawa 2007.

Maddala G.S., *Ekonometria*, Wydawnictwo Naukowe PWN, Warszawa 2006.

Walesiak M., Gatnar E. (red.) *Statystyczna analiza danych z wykorzystaniem programu R*, Wydawnictwo Naukowe PWN, Warszawa 2009.

MODELLING QUALITATIVE DATA AND LIMITED DATA USING GRETL PACKAGE

Summary: The paper presents examples of modelling limited data for binary, discrete, count variables using logit estimation for variables of binominal, multinomial, ordered multinominal, tobit and heckit estimation, Poisson regression and regression for binomial negative count variable. All examples are presented using estimation procedures available in GRETL package.