Ekonometria z pakietem Stata - skrypt

Karol Kuhl

25 września 2005

Spis treści

1 Krótkie wprowadzenie

Program Stata jest uniwersalnym pakietem statystycznym wykorzystywanym w statystyce, ekonometrii, socjologii, psychometrii, biometrii i innych dziedzinach. Umożliwia on m.in. obróbkę zbiorów danych, przedstawianie ich zawartości w formie graficznej, estymację modeli statystycznych, prowadzenie obliczeń na macierzach.

Po uruchomieniu programu, na ekranie pojawia się główne okno zatytułowane *Intercooled Stata*, w którym widoczne są następujące okna:

- *Stata Results* - okno, w którym wyświetlane są wyniki obliczeń, komunikaty błędów i inne informacje dotyczące bieżącej sesji,
- *Stata Command* - wiersz poleceń, w którym wpisuje się polecenia wykonywane przez program,
- *Review* - zapis wszystkich (poprawnych i nie poprawnych) poleceń wywołanych z wiersza poleceń podczas bieżącej sesji,
- *Variables* - okno, w którym wyświetlane są wszystkie zmienne z bieżącego zbioru danych.

Najważniejszymi ikonami skrótów są:

- *Do-file Editor* - edytor plików wsadowych (plików z rozszerzeniem *.do*); wywołuje się go za pomocą piątej ikony od lewej strony.
- *Data Editor* - arkusz z danymi; wywołuje się go za pomocą czwartej ikony od prawej,

Praca w pakiecie Stata może odbywać się na dwa sposoby: (1) poprzez polecenia wpisywane do i uruchamiane z wiersza poleceń lub (2) poprzez polecenia wpisane do pliku wsadowego (pliki z rozszerzeniem `.do`) i uruchamiane z edytora plików wsadowych.

2 Pakiet Stata jako kalkulator

Choć podstawawowym zadaniem pakietu Stata są obliczenia prowadzone na zbiorach danych, to można go również używać do podręcznych obliczeń, jako rozbudowany kalkulator. Służy do tego polecenie `display` wpisywane do wiersza poleceń, przykładowo:

```
display (7*(6+5)/4)^(-0.3)
```

Ponadto polecenie `display` pozwala korzystać ze zdefiniowanych funkcji matematycznych, w tym statystycznych. Wartość dystrybuanty rozkładu normalnego standardowego w punkcie 1.96 można obliczyć w następujący sposób:

```
display norm(1.96)
```

3 Działania na macierzach w pakiecie Stata

Pakietu Stata można używać do prowadzenia działań na macierzach. Pierwszym krokiem w tym kierunku jest deklaracja macierzy za pomocą polecenia `matrix`. Aby zdefiniować macierz:

$$\mathbf{X} = \begin{bmatrix} 1 & 4 \\ 1 & -2 \\ 1 & 3 \\ 1 & -5 \end{bmatrix}$$

należy w wierszu poleceń wpisać:

```
matrix X=(1,4\1,-2\1,3\1,-5)
```

i potwierdzić wciśnięciem klawisza `Enter`. W ten sposób zdefiniowana zostanie w pamięci programu macierz oznaczona symbolem `X`. Uwaga: pakiet Stata rozróżnia wielkość znaków w nazwach macierzy. Aby sprawdzić, czy elementy macierzy są prawidłowo wpisane można wyświetlić ją za pomocą polecenia:

```
matrix list X
```

Powyższe polecenie wymaga znajomości nazwy macierzy. Nazwy (i wymiary) wszystkich zdefiniowanych macierzy można uzyskać za pomocą polecenia:

```
matrix dir
```

Działania na macierzach dostępne w pakiecie Stata obejmują m.in. transpozycję, dodawanie mnożenie przez skalar, mnożenie, odwracanie. Dobrą ilustracją może być zadanie obliczenia macierzy $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ i $\mathbf{M} = \mathbf{I} - \mathbf{P}$, dla danej wyżej macierzy $\mathbf{X}$. Pierwszym krokiem może być obliczenie macierzy symetrycznej $\mathbf{X}'\mathbf{X}$, której nadana zostanie nazwa $\mathbf{XX}$:

```
matrix XX=X'*X
```

Po wyświetleniu jej za pomocą polecenia:

```
matrix list XX
```

okazuje się, że macierze symetryczne są w pakiecie Stata przedstawiane w postaci trójkątnej, tzn. elementy znad głównej przekątnej (będące odbiciem elementów spod głównej przekątnej) nie są wyświetlane. Następnym krokiem jest zdefiniowanie macierzy odwrotnej $\mathbf{XX}^{-1}$:

```
matrix IXX=inv(XX)
```

Do tego celu wykorzystana została funkcja `inv`, której argumentem musi być macierz kwadratowa. Kolejnym krokiem może być zdefiniowanie całej macierzy $\mathbf{P}$ w oparciu o policzone wcześniej macierze:

```
matrix P=X*IXX*X'
```

Aby policzyć macierz $\mathbf{M}$ wygodnie jest wykorzystać funkcję `I(n)`, która tworzy macierz jednostkową o wymiarze `n`:

```
matrix I4=I(4)
```

Wtedy macierz $\mathbf{M}$ definiuje się w następujący sposób:

```
matrix M=I4-P
```

Oczywiście wszystkie powyższe polecenia można zebrać w jedno:

```
matrix M=I(4)-X*inv(X'*X)*X'
```

Ciekawy wynik uzyskuje się mnożąc macierze $\mathbf{M}$ i $\mathbf{P}$:

```
matrix MP=M*P
matrix list MP
```

Pierwszym elementem macierzy $\mathbf{MP}$ jest liczba `2.082e-17` (czyli $2.08 \cdot 10^{-17}$) chociaż powinno to być zero. Rozbieżność wynika stąd, że pakiet Stata prowadzi obliczenia numerycznie, a nie analitycznie. Oczywiście liczba $2.08 \cdot 10^{-17}$ jest bardzo bliska zeru.

4 Wczytywanie danych do pakietu Stata

Zbiory danych, na których pracę umożliwia pakiet Stata, są tabelami zawierającymi informacje numeryczne lub tekstowe. Kolumny tabeli nazywają się zmiennymi, a wiersze - obserwacjami. Przykładem zbioru danych może być lista obecności, w której występują zmienne: liczba porządkowa (zmienna numeryczna), imię (zmienna tekstowa), nazwisko (zmienna tekstowa), nr indeksu (zmienna numeryczna). Obserwacjami w takim zbiorze są wpisy dotyczące poszczególnych osób.

Zbiory danych można do pakietu Stata wczytywać na kilka sposobów:

1. Bezpośrednio w trybie edycji danych.

2. Wklejając dane z tabeli zdefiniowanej w innym programie (np. Excel).

3. Z zewnętrznego zbioru z danymi (np. pliku .txt).

4. Ze zbioru danych w formacie pakietu Stata.

Uwaga: przed wypróbowaniem każdego z powyższych sposobów należy wyczyścić pamięć pakietu Stata za pomocą polecenia:

```
clear all
```

Metoda 1 polega na otwarciu okna *Stata Editor* i „ręcznym" wpisaniu wartości zmiennych. W tym celu należy kliknąć czwartą ikonę od prawej strony i po pojawieniu się arkusza, wpisywać kolejne wartości. Domyślne nazwy zmiennych (`var1`, `var2`, itd.) pojawią się automatycznie. Podobnie będzie z numerami obserwacji. Braki danych oznaczone są przez kropkę. Zakończenie ręcznego wpisywania danych odbywa się poprzez zamknięcie okna *Stata Editor*. To, że zbiór danych jest już wczytany objawia się tym, że w oknie *Variables* widoczne są nazwy zmiennych.

W metodzie 2 do arkusza *Stata Editor* wkleja się dane skopiowane w innych programach. Można w ten sposób wczytać na przykład tabelę z programu Excel. Tabelę należy zaznaczyć i skopiować (np. skrótem `Control+C`), i wkleić (`Control+V`) w pierwszą komórkę arkusza *Stata Editor*. Nazwy zmiennych i obserwacje pojawią się automatycznie. Po takiej operacji w oknie *Stata Results* pojawi się komunikat o liczbie wklejonych zmiennych i obserwacji.

Metoda 3 wymaga: po pierwsze pliku z danymi w formacie .txt, po wtóre znajomości nazwy i ścieżki dostępu do tego pliku. Pierwszym krokiem do zastosowania tej metody jest zmiana domyślnej ścieżki dostępu na właściwą. Jeżeli plik `dane.txt` znajduje się w folderze `l:\ekonometria`, to właściwą ścieżki ustawia się poprzez polecenie:

```
cd "l:\ekonometria"
```

Możliwe jest wtedy wyświetlenie zawartości bieżącego katalogu za pomocą polecenia `dir`. Jeżeli plik `dane.txt` pojawi się na liście plików w folderze, to można go wczytać za pomocą polecenia:

```
insheet using "dane.txt", names delimiter(" ")
```

Jest to plik tekstowy o strukturze podobnej do struktury zbioru danych (nazwy zmiennych w pierwszym wierszu, zmienne w kolumnach, obserwacje w wierszach), w którym dla każdej obserwacji wartości kolejnych zmiennych oddzielone są spacją (stąd opcja `delimiter`). Oto pierwsze 6 wierszy tego zbioru:

```
y x
946.6 192.4
923.8 157
949 170
750.5 43.8
536.3 8.5
```

Po takiej operacji w oknie *Stata Results* pojawi się komunikat o liczbie wczytanych zmiennych i obserwacji, natomiast w oknie *Variables* pojawią się nazwy zmiennych.

Metoda 4 jest najprostsza - polega na otwarciu odpowiedniego pliku. Jak prawie każdy pakiet statystyczny, Stata ma swój własny format zapisywania zbiorów danych. Pliki zawierające dane w tym formacie mają rozszerzenie `.dta`. Wczytywanie danych w tym formacie odbywa się np. poprzez otwarcie pliku w programie Stata. Zapisanie danych w tym formacie odbywa się analogicznie. Pliki `.dta` można również otwierać z wiersza poleceń. Jeżeli w folderze znajduje się szukany plik `.dta` (co można sprawdzić za pomocą polecenia `dir`), to wczytanie zbioru `inwestycje.dta` odbywa się w następujący sposób:

```
use inwestycje
```

5 Efektywna praca z pakietem Stata

Program Stata pozwala zautomatyzować pracę dzięki wykonywaniu poleceń zawartych w plikach wsadowych (plikach typu-*do* - nazwa związana jest z rozszerzeniem: `.do`). Są to pliki tekstowe, które można otworzyć i uruchomić w oknie *Stata Do-file Editor*. Przykładowy plik `macierze.do` wygląda następująco:

```
matrix X = (1,4\1,-2\1,3\1,-5)
matrix list X
matrix dir
matrix XX = X' * X
matrix list XX
matrix dir
matrix IXX = inv(XX)
matrix list IXX
matrix P = X * IXX * X'
matrix list P
matrix I4 = I(4)
matrix list I4
matrix M = I4 - P
matrix list M
matrix M = I(4) - X*IXX*X'
matrix list M
matrix MP = M * P
matrix list MP
```

Powyższy plik typu-*do* wykonuje opisane wcześniej działania na macierzach. Po otwarciu pliku w oknie *Stata Do-file Editor*, uruchomienie odbywa się poprzez zaznaczenie jego fragmentu i wciśnięcie drugiej ikony od prawej strony *Do current file*. Spowoduje to wykonanie, po kolei poleceń z kolejnych wierszy tak, jakby były one wywoływane z wiersza poleceń.

Praca w programie Stata powinna być prowadzona za pomocą plików typu-*do* wtedy, gdy liczba poleceń wywoływanych z wiersza poleceń przekracza 2. Taki tryb pracy jest początkowo trudny, ale warto jest go stosować ponieważ stosunkowo szybko przynosi korzyści w postaci zwiększonej efektywności.

6 Charakterystyki zbioru danych i zmiennych

Po wczytaniu danych dobrze jest sprawdzić liczbę obserwacji oraz liczbę zmiennych i ich typ. Iformacje nt. zbioru danych wywoływane są za pomocą polecenia:

```
describe
```

W przypadku danych z pliku `dane.txt` wyswietlone zostaną następujące informacje:

```
Contains data
  obs:           200
 vars:             2
 size:         2,400 (99.9% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
y               float  %9.0g
x               float  %9.0g
-------------------------------------------------------------------------------
Sorted by:
     Note:  dataset has changed since last saved
```

Wynika z nich, że zbiór danych utworzony z pliku `dane.txt` zawiera 200 obserwacji i 2 zmienne, zajmuje 2400Kb pamięci. Zmienne w tym zbiorze to `x` i `y`, obydwie są typu numerycznego (*float*).

Statystyki opisujące zmienne typu numerycznego (liczba ważnych obserwacji, średnia, odchylenie standardowe, minimum, maksimum) można wywołać za pomocą polecenia:

```
summarize x y
```

Otrzymuje się wtedy:

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           x |       200      95.603    57.45135          .8      199.6
           y |       200      702.31    154.6516       321.5       1164
```

Najprostszą miarę współzależności, współczynnik korelacji Pearsona, wywołuje się za pomocą polecenia:

```
correlate x y
```

Wynik przedstawiony jest w postaci tabeli:

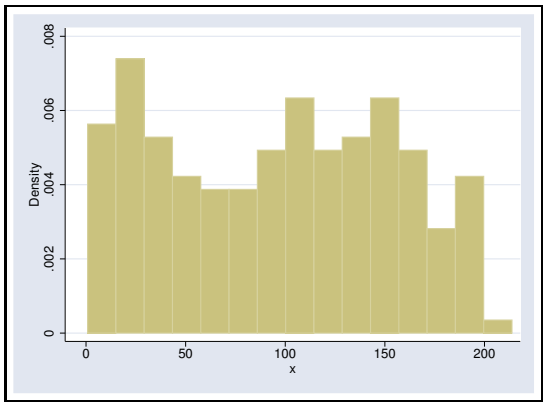
```
(obs=200)
             |        x        y
-------------+------------------
           x |   1.0000
           y |   0.6812   1.0000
```

W celu wyświetlenia histogramu opisującego rozkład zmiennej `x`, należy użyć polecenia:

```
histogram x
```

Realizacja tego polecenia zajmuje trochę czasu, a jego wynik pojawia się w nowym oknie – patrz rysunek 1. Natomiast celem wyświetlenia wykresu rozrzutu (wykresu punktowego) opisującego

Rysunek 1: Przykładowy histogram.

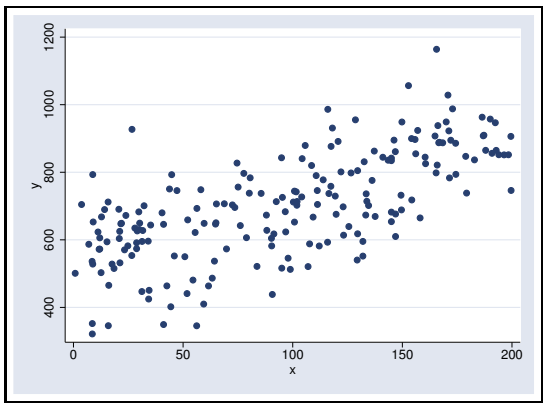


rozkład zmiennych `y` i `x`, należy użyć polecenia:

```
scatter y x
```

Ponownie wyniki realizacji dostępne są w nowym oknie – patrz rysunek 2.

Rysunek 2: Przykładowy wykres rozrzutu.



7 Estymacja MNK w zapisie macierzowym

Obliczenia zostaną przeprowadzone na danych zawartych w pliku `dane.txt`, który należy wczytać za pomocą odpowiedniego polecenia. Wykres rozrzutu (rysunek 2) pokazuje, że pomiędzy zmiennymi występuje dodatni związek. Można zatem podjąć próbę opisania tego związku za pomocą modelu regresji liniowej. Aby skorzystać z zapisu macierzowego należy najpierw zadeklarować odpowiednią ilość pamięci (polecenie `set matsize 800`), a następnie utworzyć potrzebne macierze $\mathbf{y}$ i $\mathbf{X}$. Służy do tego polecenie:

```
mkmat x, matrix(x2)
```

które ze zmiennej `x` tworzy wektor kolumnowy `x2`. Analogicznie tworzony jest wektor kolumnowy `y`. Macierz $\mathbf{X}$ zawiera kolumnę jedynek. Można ją utworzyć jako wektor `x1` za pomocą polecenia:

```
matrix x1=J(200,1,1)
```

w którym funkcja `J` tworzy macierz o wymiarze 200 na 1, której każdym elementem jest 1, czyli dwustuelementowy kolumnowy wektor jedynek. Łączenie wektorów w macierz poprzez zestawienie kolumn odbywa się w następujący sposób:

```
matrix X=x1,x2
```

W ten sposób zdefiniowana została macierz `X`, dzięki czemu można policzyć wektor $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$:

```
matrix b=inv(X'*X)*X'*y
```

Aby wyświetlić elementy wektora `b`, należy użyć polecenia:

```
matrix list b
```

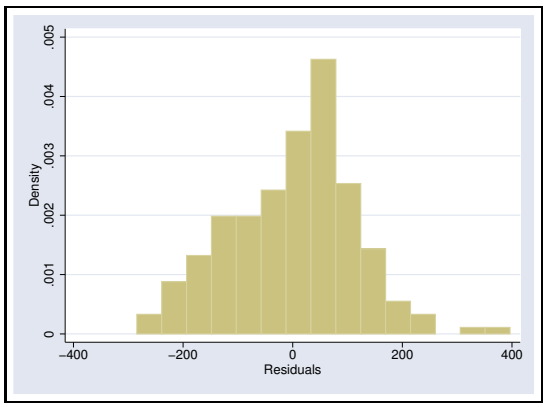
Oceny parametrów regresji liniowej zmiennej `y` na zmienną `x` zapisane są jako elementy macierzy `b` o wymiarach 2 na 1:

```
b[2,1]
            y
c1  526.799756
 x  1.8337546
```

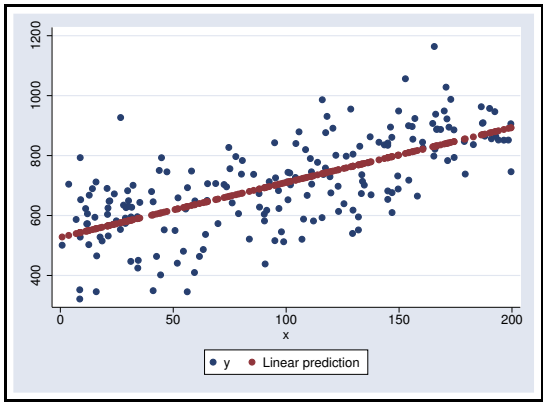
8 Estymacja MNK za pomocą polecenia `regress`

Mając zmienne `y` i `x` można policzyć oceny punktowe estymatorów MNK (oraz wiele innych wielkości) za pomocą polecenia:

Rysunek 3: Histogram reszt.



Rysunek 4: Wartości surowe i dopasowane.




```
regress y x
```

Bezpośrednio po zastosowaniu polecenia `reg` można wygenerować dwie interesujące zmienne: `e` (reszty modelu e) i `yy` (wartości teoretyczne $\hat{y}$) za pomocą poleceń:

```
predict e, r
predict yy, xb
```

Zmienna `e` powinna, zgodnie z założeniami modelu powinna mieć rozkład zbliżony do normalnego, co można wizualnie sprawdzić za pomocą polecenia:

```
histogram e
```

Natomiast zmienna `yy` powinna być liniową funkcją zmienne `x`, co można wizualnie sprawdzić za pomocą polecenia:

```
scatter yy x
```

Wyniki przedstawiają rysunki 3 i 4.

9 Zapisywanie wyników obliczeń do pliku

Pierwszym krokiem pracy w programie Stata jest wyczyszczenie pamięci i ustwienie ścieżki dostępu do folderu zawierającego plik z danymi:

```
clear all
cd "l:\ ekonometria"
```

Następnie należy zadeklarować nazwę zewnętrznego pliku, do którego mają być zapisywane wyniki, które ukazują się w oknie *Stata Results*:

```
log using dziennik, replace
```

Po czym należy tę opcję włączyć:

```
log on
```

Wszystko co od tego momentu ukaże się w oknie *Stata Results* zostanie zapisane w pliku `dziennik.smcl`, aż do momentu, w którym opcja zapisu zostanie wyłączona poprzez polecenie:

```
log off
```

Na koniec sesji należy zamknąć zapisywanie poprzez:

```
log close
```

Pliku `dziennik.smcl` jest plikiem, którego prawidłowo nie otworzy się poza programem Stata. W celu „przetłumaczenia" tego pliku na plik tekstowy używa się polecenia:

```
translate dziennik.smcl dziennik.txt, replace
```

Plik `dziennik.txt` można otworzyć i edytować w edytorze tekstowym, np. programie Word.

10 Omówienie wyników estymacji KMRL

Poniżej znajduje się 20 obserwacji zmiennych `x2`, `x3`, `y`, które zostały wykorzystane do oszacowania parametrów KMRL:

x2	x3	y
89	248	445
48	209	394
22	214	324
93	211	424
89	230	471
86	203	476
19	227	294
62	235	395
73	212	427
54	210	419
8	223	297
38	231	351
4	209	280
80	243	392
41	235	354
50	218	382
36	218	376
29	213	295
41	241	340
19	248	310

Na podstawie powyższych danych estymowany jest następujący model:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$
$$\epsilon_i \underset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$$

Parametry powyższego KMRL szacuje się za pomocą polecenia:

```
reg y x2 x3
```

Pakiet Stata zwraca następujący zestaw wyników:

```
    Source |       SS       df       MS              Number of obs =      20
-----------+------------------------------           F(  2,    17) =   64.78
     Model |  61394.8437     2  30697.4219           Prob > F      =  0.0000
  Residual |   8055.3563    17  473.844488           R-squared     =  0.8840
-----------+------------------------------           Adj R-squared =  0.8704
     Total |    69450.20    19  3655.27368           Root MSE      =  21.768

------------------------------------------------------------------------------
         y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
        x2 |   1.997733   .1762852    11.33   0.000     1.625804    2.369663
        x3 |  -.5884121   .3521532    -1.67   0.113     -1.33139    .1545662
     _cons |   406.0566   79.01142     5.14   0.000     239.3571    572.7561
------------------------------------------------------------------------------
```

W tabeli z wynikami podane są: liczba obserwacji i liczba zmiennych (oraz stała):

```
    Source |       SS       df       MS              Number of obs =      20
-----------+------------------------------           F(   ,      ) =
     Model |                                         Prob > F      =
  Residual |                                         R-squared     =
-----------+------------------------------           Adj R-squared =
     Total |                                         Root MSE      =

------------------------------------------------------------------------------
         y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
        x2 |
        x3 |
     _cons |
------------------------------------------------------------------------------
```

Zgodnie z oczekiwaniami $n = 20$, a liczba zmiennych jest równa liczbie wierszy w dolnej tabeli, czyli k=3.

Ocena wektora $\boldsymbol{\beta}$ dana jest wzorem:

$$\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad .$$

Wielkości b_1, b_2, b_3 odczytuje się z dolnej częsci tabeli. Należy pamiętać, że program Stata wyświetla je w innej kolejności, tzn. stała modelu (oznaczona słowem `_cons`) znajduje sie na ostatnim miejscu.

```
------------------------------------------------------------------------------
         y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
        x2 |   1.997733
        x3 |  -.5884121
     _cons |   406.0566
------------------------------------------------------------------------------
```

Zatem:

$$\mathbf{b} = \begin{bmatrix} 406.0566 \\ 1.997733 \\ -0.5884121 \end{bmatrix}$$

Poza wektorem $\boldsymbol{\beta}$ szacowana jest również wariancja składnika losowego σ^2. Zgodnie ze wzorem estymatorem tego parametru jest:

$$\hat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{n-k} = \frac{\mathbf{y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y}}{n-k}$$

Program Stata zgłasza pierwiastek tej wartości (*MSE* to skrót od *mean square error*):

```
    Source |       SS       df       MS              Number of obs =
-----------+------------------------------           F(  2,    17) =
     Model |                                         Prob > F      =
  Residual |                                         R-squared     =
-----------+------------------------------           Adj R-squared =
     Total |                                         Root MSE      =  21.768
```

Czyli: $\hat{\sigma}^2 = 21.768^2 = 473.845$.

Estymator macierzy wariancji-kowariancji dany jest wzorem:

$$\widehat{\mathrm{Var}}[\mathbf{b}] = \begin{bmatrix} \widehat{\mathrm{Var}}[b_1] & \widehat{\mathrm{Cov}}[b_1, b_2] & \widehat{\mathrm{Cov}}[b_1, b_3] \\ \widehat{\mathrm{Cov}}[b_1, b_2] & \widehat{\mathrm{Var}}[b_2] & \widehat{\mathrm{Cov}}[b_2, b_3] \\ \widehat{\mathrm{Cov}}[b_1, b_3] & \widehat{\mathrm{Cov}}[b_2, b_3] & \widehat{\mathrm{Var}}[b_3] \end{bmatrix} = \frac{\mathbf{y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y}}{n-k}(\mathbf{X}'\mathbf{X})^{-1}$$

Pakiet Stata „sam z siebie" nie wyświetla tej macierzy w całości, ale w tabeli z wynikami znajdują się błędy standardowe ocen parametrów, będące pierwiastkami elementów znajdujących się na głównej przekątnej tej macierzy:

```
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x2 |               .1762852
          x3 |               .3521532
       _cons |               79.01142
------------------------------------------------------------------------------
```

Zatem (ponownie uwaga na kolejność):

$$\widehat{\mathrm{Var}}[b_1] = 79.01142^2 = 6242.804$$
$$\widehat{\mathrm{Var}}[b_2] = 0.1762852^2 = 0.310765$$
$$\widehat{\mathrm{Var}}[b_3] = 0.3521532^2 = 0.124012$$

Do oceny jakości dopasowania modelu używa się dekompozycji sumy kwadratów odchyleń zmiennej objaśnianej od jej średniej (*TSS - total sum of squares*) na sumę kwadratów wyjaśnioną modelem (*ESS - explained sum of squares*) oraz sumę kwadratów nie wyjaśnioną modelem (*RSS - residual sum of squares*). Pamiętając od macierzy idempotentnej $\mathbf{M_0} = \mathbf{I} - \frac{1}{n}\mathbf{i}\mathbf{i}'$, która przemienia macierz w macierz odchyleń od średnich, można dekompozycję całkowitej zmienności zapisać następująco:

$$TSS = ESS + RSS$$
$$\mathbf{y}'\mathbf{M_0}\mathbf{y} = \mathbf{b}'\mathbf{X}'\mathbf{M_0}\mathbf{X}\mathbf{b} + \mathbf{e}'\mathbf{e}$$

Program Stata oblicza każdą z powyższych wielkości:

```
    Source |       SS       df       MS              Number of obs =
-----------+------------------------------           F(   ,      ) =
     Model |  61394.8437                             Prob > F      =
  Residual |   8055.3563                             R-squared     =
-----------+------------------------------           Adj R-squared =
     Total |    69450.20                             Root MSE      =
```

I stąd:

$$TSS = 69450.20$$
$$ESS = 61394.8437$$
$$RSS = 8055.3563 \quad .$$

Współczynnik determinacji liniowej jest obliczany zgodnie ze wzorem:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Zawsze istnieje możliwość polepszenia jakości dopasowania poprzez zwiększenie liczby regresorów w modelu. Dlatego inną miarą dopasowania jest skorygowany wpółczynnik determinacji liniowej, który bierze pod uwagę liczbę regresorów. Opisany jest on wzorem:

$$R_*^2 = 1 - (1 - R^2)\frac{n-1}{n-k}$$

Program Stata zgłasza obydwie te wielkości.

```
    Source |       SS       df       MS              Number of obs =
-----------+------------------------------           F(  2,    17) =
     Model |                                         Prob > F      =
  Residual |                                         R-squared     =  0.8840
-----------+------------------------------           Adj R-squared =  0.8704
     Total |                                         Root MSE      =
```

W oszacowanym modelu współczynniki te wynoszą:

$$R^2 = 0.8840 = 88.4\%$$
$$R_*^2 = 0.8704 = 87.04\%$$

Zgodnie z założeniami KMRL, wektor $\mathbf{b}$ jest wektorem losowym o rozkładzie normalnym z następującymi parametrami (wektorem wartości oczekiwanych i macierzą wariancji-kowariancji):

$$\mathrm{E}\mathbf{b} = [\beta_1, \beta_2, \ldots, \beta_k]' \quad \mathrm{Var}\mathbf{b} = \sigma^2(X'X)^{-1}$$

Każdy z elementów wektora $\mathbf{b}$ można poddać standaryzacji w celu otrzymania zmiennej losowej o rozkładzie normalnym standardowym:

$$b_j \sim \mathrm{N}(\beta_j, \sigma_j^2) \Rightarrow \frac{b_j - \beta_j}{\sigma_j} \sim \mathrm{N}(0,1) \quad j = 1 \cdots k$$

Niestety macierz wariancji-kowariancji nie jest znana i musi być szacowana za pomocą $\hat{\mathrm{Var}}\mathbf{b} = \frac{\mathbf{e}'\mathbf{e}}{n-k}(X'X)^{-1}$. W związku z tym, powyższa standaryzacja odbywa się poprzez podzielenie przez $\hat{\sigma}_j$, czego konsekwenjcą jest inny rozkład takiej zmiennej:

$$\frac{b_j - \beta_j}{\hat{\sigma}_j} \sim \mathrm{t}_{n-k} \quad j = 1 \cdots k$$

Dzięki temu można testować hipotezy statystyczne mówiące o tym, że $\beta_j = 0$. Jeżeli taka hipoteza byłaby prawdziwa, to statystyka $t_j = \frac{b_j}{\hat{\sigma}_j}$ ma rozkład t-Studenta o $n-k$ stopniach swobody. Jeżeli zatem wartość tej statystyki „wpadnie" do obszaru krytycznego (czyli stanie się coś mało prawdopodobnego), to hipotezę zerową należy odrzucić i tym samym regresor j staje się istotny statystycznie.

Aby sprawdzić, czy wartość statystyki zawiera się w przedziale krytycznym, ocenia się prawdopodobieństwo dla tej statystyki (oznaczane w pakietach statystycznych: `p-value`, `prob`, `P>|t|`). Wylicza się je w oparciu o wzór:

$$\texttt{p} - \texttt{value}_j = 2(1 - F_{t_{n-k}}(\frac{b_j}{\hat{\sigma}_j}))$$

w którym $F_{t_{n-k}}$ to dystrybuanta rozkładu t-Studenta o $n-k$ stopniach swobody. Jeżeli prawdopodobieństwo jest większe od założonego *a priori* poziomu istotności, to wartość statystyki nie zawiera się w zbiorze krytycznym i nie ma podstaw do odrzucenia hipotezy zerowej.

```
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x2 |    1.997733   .1762852    11.33   0.000
          x3 |   -.5884121   .3521532    -1.67   0.113
       _cons |    406.0566   79.01142     5.14   0.000
------------------------------------------------------------------------------
```

Hipoteza mówiąca o tym, że $\beta_2 = 0$ jest odrzucana ponieważ:

$$t_2 = \frac{1.997733}{0.1762852} \approx 11.33$$

co jest wartością na tyle dużą (co do wartości bezwzględnej), że zawiera się w obszarze odrzuceń, o czym świadczy bardzo mała wartość prawdopodobieństwa. Również hipoteza mówiąca o tym, że $\beta_1 = 0$ (stała) jest odrzucana. Natomiast analogiczna hipoteza dotycząca β_3 nie jest odrzucana ponieważ $0.113 > 5\%$ (5% to standardowy poziom istotności testów). Mówi się wtedy, że zmienna nie jest istotna w tak wyspecyfikowanym modelu. W oparciu o fakt, że:

$$\frac{b_j - \beta_j}{\hat{\sigma}_j} \sim \mathrm{t}_{n-k} \quad j = 1 \cdots k$$

można konstruować przedziały ufności dla parametrów b_j:

$$\Pr(-t^*_{\alpha/2} < \frac{b_j - \beta_j}{\hat{\sigma}_j} < t^*_{\alpha/2}) = 1 - \alpha$$

gdzie $t^*_{\alpha/2}$ to wartość krytyczna taka, że dla zmiennej losowej o rozkładzie t:

$$\Pr(t > -t^*_{\alpha/2}) = \alpha/2$$

Ostatecznie przedziały ufności o poziomie $1 - \alpha/2$ są następujące:

$$(b_j - \hat{\sigma}_j t^*_{\alpha/2}; b_j + \hat{\sigma}_j t^*_{\alpha/2})$$

Wartość krytyczna dla rozkładu t o $n - k = 17$ stopniach swobody i $\alpha = 5\%$ wynosi $t^*_{0.025} \approx 2.11$.

```
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x2 |    1.997733   .1762852                      1.625804    2.369663
          x3 |   -.5884121   .3521532                     -1.33139     .1545662
       _cons |    406.0566   79.01142                      239.3571    572.7561
------------------------------------------------------------------------------
```

95%-przedział ufności dla stałej w modelu wynosi:

$$b_1 \pm \hat{\sigma}_j t^*_{\alpha/2} = 406.0566 \pm 79.01142 \cdot 2.11 \approx (239.3571; 572.7561)$$

Istnieje również możliwość testowania hipotezy nt. istotności całej regresji. Hipoteza zerowa ma w tym przypadku postać – $H_0 : \beta_2 = \beta_2 = \cdots = \beta_k = 0$, natomiast hipoteza alternatywna mówi

o tym, że przynajmniej jeden z parametrów β jest różny od zera. Jeżeli taka hipoteza byłaby prawdziwa, to dostępne byłyby dwa estymatory wariancji składnika losowego σ^2:

$$\hat{\sigma}^2_{ESS} = \mathbf{b}'\mathbf{X}'\mathbf{M_0}\mathbf{X}\mathbf{b}$$
$$\hat{\sigma}^2_{RSS} = \mathbf{e}'\mathbf{e}$$

Powyższe estymatory są zmiennymi losowym o rozkładach χ^2 i liczbie stopni swobody równej odpowiednio $n-k$ i $n-1$. W związku z tym statystyka:

$$F = \frac{\hat{\sigma}^2_{ESS}/(k-1)}{\hat{\sigma}^2_{RSS}/(n-k)}$$

miałaby rozkład F-Snedecora o $k-1$ i $n-k$ stopniach swobody. Powyższy wzór można uprościć:

$$\begin{aligned} F &= \frac{\hat{\sigma}^2_{ESS}/(k-1)}{\hat{\sigma}^2_{RSS}/(n-k)} = \frac{\mathbf{b}'\mathbf{X}'\mathbf{M_0}\mathbf{X}\mathbf{b}/(k-1)}{\mathbf{e}'\mathbf{e}/(n-k)} \\ &= \frac{ESS/(k-1)}{RSS/(n-k)} = \frac{\frac{ESS}{TSS}/(k-1)}{\frac{RSS}{TSS}/(n-k)} \\ &= \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \sim \mathrm{F}_{k-1,n-k} \end{aligned}$$

Hipoteza zerowa ($H_0 : \beta_2 = \beta_2 = \cdots = \beta_k = 0$) jest odrzucana, jeżeli wartość tej statystyki przekroczy wartość krytyczną F_*. Świadczy o tym odpowiednia wartość prawdopodobieństwa dla tej statystyki: jeżeli prawdopodobieństwo jest mniejsze od założonego poziomu, to hipoteza jest odrzucana, a regresja statystycznie istotna.

```
    Source |       SS       df       MS              Number of obs =
-----------+------------------------------           F(  2,    17) =   64.78
     Model |  61394.8437     2  30697.4219           Prob > F      =  0.0000
  Residual |   8055.3563    17  473.844488           R-squared     =
-----------+------------------------------           Adj R-squared =
     Total |    69450.20                             Root MSE      =
```

Program Stata oblicza kolejne wartości:

$$F = \frac{61394.8437/2}{8055.3563/17} = \frac{30697.4219}{473.844488} \approx 64.78$$

Ostatecznie hipoteza o łącznej nieistotności współczynników regresji jest odrzucana ponieważ $0.0000 \lll 5\%$.

11 Liniowa kombinacja współczynników regresji

Pakiet Stata umożliwia również oddzielne testowanie hipotez nt. liniowych kombinacji elementów wektora współczynników. Kombinacja liniowa elementów zmiennej losowej o wielowymiarowym rozkładzie normalnym jest zmienną losową o rozkładzie normalnym:

$$\mathbf{v}'\mathbf{b} = [v_1, v_2, \cdots, v_k] \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_k \end{bmatrix} = \sum_{i=1}^{k} v_i b_i \sim \mathrm{N}$$

Wartość oczekiwana zmiennej losowej będącej kombinacją liniową elementów zmiennej losowej o wielowymiarowym rozkładzie normalnym jest taką samą kombinacją liniową wartości oczekiwanych elementów tejże wielowymiarowej zmiennej losowej:

$$\mathrm{E}(\mathbf{v}'\mathbf{b}) = \mathbf{v}'\mathrm{E}(\mathbf{b}) = \sum_{i=1}^{k} v_i \mathrm{E}(b_i) = \sum_{i=1}^{k} v_i \beta_i$$

Wariancja zmiennej losowej będącej kombinacją liniową elementów zmiennej losowej o wielowymiarowym rozkładzie normalnym jest następującą formą kwadratową:

$$\mathrm{Var}(\mathbf{v}'\mathbf{b}) = \mathbf{v}'\mathrm{Var}(\mathbf{b})\mathbf{v}$$

Do tego typu obliczeń służy w pakiecie Stata polecenie `lincom` użyte bezpośrednio po oszacowaniu modelu (!!!). W celu ponownego sprawdzenia istotności współczynnika przy zmiennej `x3` należy wpisać:

```
lincom 0*x2+1*x3+0*_cons
```

co odpowiada przemnożeniu wektora współczynników przez wektor $\mathbf{v} = [0, 1, 0]'$. Należy pamiętać o tym, że Stata umiejscawia współczynnik przy wyrazie wolnym na końcu listy współczynników. Wynikiem takiego polecenia jest:

```
 ( 1)  x3 = 0

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.5884121   .3521532    -1.67   0.113     -1.33139    .1545662
------------------------------------------------------------------------------
```

co odpowiada wierszowi wcześniejszej tabeli.

Zagadnieniem zbliżonym jest wyznaczanie prognoz na podstawie oszacowanego modelu. Mając wektor $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, można szacować wartość zmiennej objaśnianej przy zadanych wartościach zmiennych objaśniających:

$$\widehat{y}^* = \mathbf{x}^{*\prime}\mathbf{b} = \sum_{i=1}^{k} x_i^* b_i$$

W tej sytuacji, rozważany na początku, wektor $\mathbf{v}$ przyjmuje wartości $\mathbf{v} = [x_1^*, x_2^*, \cdots, x_k^*]'$. Na zasadzie analogii, zmienna losowa $\widehat{y}^*$ ma rozkład normalny z wartością oczekiwaną równą:

$$\mathrm{E}(\widehat{y}^*) = \mathrm{E}(\mathbf{x}^{*\prime}\mathbf{b}) = \mathbf{x}^{*\prime}\mathrm{E}(\mathbf{b}) = \mathbf{x}^{*\prime}\boldsymbol{\beta} = [x_1^*, x_2^*, \cdots, x_k^*] \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_k \end{bmatrix} = \sum_{i=1}^{k} x_i^* \beta_i$$

i wariancją równą:

$$\begin{aligned} \mathrm{Var}(\widehat{y}^*) &= \mathrm{Var}(\mathbf{x}^{*\prime}\mathbf{b}) = \mathbf{x}^{*\prime}\mathrm{Var}(\mathbf{b})\mathbf{x}^* \\ &= [x_1^*, x_2^*, \cdots, x_k^*] \begin{bmatrix} \sigma_1^2 & \mathrm{Cov}_{12} & \cdots & \mathrm{Cov}_{1k} \\ \mathrm{Cov}_{12} & \sigma_2^2 & \cdots & \mathrm{Cov}_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ \mathrm{Cov}_{1k} & \mathrm{Cov}_{2k} & \cdots & \sigma_k^2 \end{bmatrix} \begin{bmatrix} x_1^* \\ x_2^* \\ \cdots \\ x_k^* \end{bmatrix} = \sigma^{*2} \end{aligned}$$

W tym przypadku wariancja jest sumą zawierającą poszczególne wariancje i kowariancje, ponieważ elementy wektora $\mathbf{b}$ są pomiędzy sobą skorelowane.

Wnioskowanie statystyczne nt. $\widehat{y}^*$ również bazuje na standaryzacji:

$$\frac{\widehat{y}^* - y^*}{\sigma^*} \sim \mathrm{N}(0, 1)$$

Znów nieznaną macierz $\mathbf{V}$ przybliża się jej oszacowaniem i znów wnioskowanie statystyczne nt. $\widehat{y}^*$ prowadzone jest w oparciu o:

$$\frac{\widehat{y}^* - y^*}{\widehat{\sigma}^*} \sim \mathrm{t}_{(n-k)}$$

W przypadku prognozy najbardziej interesujące są: jej ocena punktowa i przedział ufności dla niej. Ocenę punktową oblicza się podstawiając do oszacowanego równania regresji odpowiednie wartości z wektora $\mathbf{x}^*$:

$$\widehat{y}^* = \mathbf{x}^{*\prime}\mathbf{b}$$

Uwaga: element tego wektora odpowiadający wyrazowi wolnemu musi być równy 1. Natomiast przedział ufności wyznaczają liczby:

$$\mathbf{x}^{*\prime}\mathbf{b} \pm t_{\frac{\alpha}{2},n-k}\widehat{\sigma}^*$$

gdzie $t_{\frac{\alpha}{2},n-k}$ jest kwantylem rzędu $\frac{\alpha}{2}$ rozkładu t z $n-k$ stopniami swobody.

Do przeprowadzania tych obliczeń w pakiecie Stata również służy polecenie `lincom`, użyte bezpośrednio po oszacowaniu modelu. Prognoza modelu oszacowanego wcześniej dla `x2`=0.5 i `x3`=0.7 odbywa się w następujący sposób:

```
lincom 0.5*x1+.7*x2+1*_cons
```

co odpowiada przemnożeniu wektora współczynników przez wektor $[0.5, 0.7, 1]'$ (pod warunkiem, że w wektorze $\mathbf{b}$ wyraz wolny jest elementem ostatnim). Wynikiem jest:

```
 ( 1)  .5 x2 + .7 x3 + _cons = 0

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   406.6436   78.76191     5.16   0.000     240.4705    572.8167
------------------------------------------------------------------------------
```

Zatem prognoza zmiennej `y` dla `x2`=0.5 i `x3`=0.7 wynosi 406.64, a 95%. przedział ufności dla niej to przedział (240.47,572.82).

12 Liniowe ograniczenia współczynników regresji

Innym przykładem ograniczenia może być warunek: $\beta_2 + \beta_3 = 1$, który można przedstawić w zapisie macierzowym za pomocą odpowiedniego wektora $\mathbf{v}$:

$$\mathbf{v}'\boldsymbol{\beta} = [0, 1, 1, \cdots, 0] \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \cdots \\ \beta_k \end{bmatrix} = \beta_2 + \beta_3 = 1$$

Testowanie takiego warunku opiera się o następującą statystykę:

$$\frac{(b_2 + b_3) - (\beta_2 + \beta_3)}{\sqrt{\widehat{\sigma}_2^2 + \widehat{\sigma}_3^2 + 2\widehat{\mathrm{Cov}}_{23}}} = \frac{(b_2 + b_3) - 1}{\sqrt{\widehat{\sigma}_2^2 + \widehat{\sigma}_3^2 + 2\widehat{\mathrm{Cov}}_{23}}} \sim \mathrm{t}_{(n-k)}$$

Test takiego warunku można przeprowadzić za pomocą polecenia `test` (wykonywanego po oszacowaniu modelu):

```
test (x2+x3=1)
```

Polecenie `test` opiera się o inną statystykę testującą - F, która co do wartości jest równa kwadratowi statystyki t. Przykładowym wynikiem jest:

```
 ( 1)  x2 + x3 = 1

       F(  1,    17) =    1.13
            Prob > F =    0.3030
```

W tym przypadku niewielka wartość statystyki F (duża wartość prawdopodobieństwa) nie pozwalają odrzucić hipotezy zerowej mówiącej o tym, że $\beta_2 + \beta_3 = 1$.

13 Opis zbioru danych

Wczytanie zbioru danych `inwestycje.dta` odbywa się poprzez polecenie:

```
use "l:\ekonometria\inwestycje"
```

Uwaga: pakiet Stata domyśla się, jakiego formatu są te dane. Po wczytaniu można wyświetlić ich opis:

```
describe
```

z którego wynika, że zmienne nie mają etykiet (kolumna *variable label* jest pusta). Przed nadaniem etykiet dobrze jest wiedzieć, czego dotyczą dane. Zbiór `inwestycje.dta` pochodzi z materiałów szkoleniowych dołączonych do podręcznika *Econometric Analysis* (W.H.Greene [2000]) i zawiera amerykańskie dane makroekonomiczne z lat 1968-1982. Kolejne zmienne to:

- `year` - rok,
- `gnp` - nominalny PNB (w mld USD),
- `invest` - nominalne inwestycje (w mld USD),
- `cpi` - wskaźnik zmian cen,
- `interest` - stopa procentowa.

Etykiety nadaje się w następujący sposób (nie mogą przekraczać 80 znaków):

```
label variable year "Rok"
label variable gnp "Nominalny PNB (mld USD)"
label variable invest "Nominalne inwestycje (mld USD)
label variable cpi "Wskaznik zmian cen"
label variable interest "Stopa procentowa"
```

Nadawanie etykiet zmiennym w zbiorze danych nie jest konieczne do oszacowania modelu, ale należy do dobrej praktyki statystycznej.

14 Prosty model ekonometryczny

W oparciu o dane ze zbioru `inwestycje.dta` można zbadać funkcję inwestycji:

$$I = f(t, PNB, r, \pi)$$

której argumentami są odpowiednio: trend liniowy (t), PNB, stopa procentowa (r), stopa inflacji (π). Funkcja f może być dowolną funkcją, ale aby zastować KMRL należy przyjąć, że f jest liniową funkcją swoich argumentów:

$$I = \beta_1 + \beta_2 t + \beta_3 PNB + \beta_4 r + \beta_5 \pi$$

Pozornie zbiór danych zawiera wszystkie zmienne z powyższego równania. W rzeczywistości postać tych zmiennych nie jest właściwa i przed ich wykorzystaniem należy dokonać pewnych transformacji. Po pierwsze nie ma zmiennej `t`. Możnaby zamiast niej użyć zmiennej `year`, ale jej duże wartości komplikują obliczenia. Dlatego dobrze jest, za pomocą polecenia `generate`, zdefiniować nową zmienna, przyjmującą wartości od 1 do 15:

```
generate t=year-1967
```

Po drugie zarówno `gnp`, jak i `invest` wyrażone są w wartościach nominalnych, w miliardach dolarów. Należy te wielkości urealnić poprzez podzielenie przez wskaźnik zmian cen (ale nie wyrażony w procentach!) i podzielić przez 1000 w celu zamiany na biliony:

```
generate rgnp=(gnp/1000)/(cpi/100)
```

To samo dotyczy zmiennej `invest`:

```
generate rinvest=(invest/1000)/(cpi/100)}
```

Ostatnią kwestią jest zamiana jednopodstawowego wskaźnika zmian cen na roczną stopę inflacji wyrażoną w procentach (podobnie do stopy procentowej). W notacji matematycznej sprowadza się to do policzenia

$$\pi_t = \frac{CPI_t - CPI_{t-1}}{CPI_{t-1}}$$

co w programie Stata osiąga się poprzez:

```
generate pi=100*(cpi[n]-cpi[n-1])/cpi[n-1]
```

Szacowany model ekonometryczny ma postać:

$$\begin{aligned} I_i &= \beta_1 + \beta_2 t_i + \beta_3 PNB_i + \beta_4 r_i + \beta_5 \pi_i + \epsilon_i \\ \epsilon_i &\underset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2) \end{aligned}$$

a szacuje się go w pakiecie Stata, po zdefiniowaniu potrzebnych zmiennych, za pomocą polecenia:

```
regress rinvest t rgnp interest pi
```