

Notatki do wykładu z ekonometrii

Stanisław Galus

10.I.2004

Spis treści

1. Model prostej regresji liniowej.	2
1.1. Założenia.	4
1.2. Metoda najmniejszych kwadratów.	5
1.3. Statystyczne własności estymatorów.	6
1.4. Wnioskowanie statystyczne.	7
2. Model wielorakiej regresji liniowej.	8
2.1. Założenia.	9
2.2. Metoda najmniejszych kwadratów.	9
2.3. Statystyczne własności estymatorów.	10
2.4. Wnioskowanie statystyczne.	10
2.5. Test istotności regresji.	11
3. Sprawdzanie hipotez.	12
3.1. Testowanie ograniczeń liniowych.	12
3.2. Testowanie zmian strukturalnych.	13
3.2.1. Różne wektory parametrów.	13
3.2.2. Różne wariancje składników losowych.	14
4. Autokorelacja składników losowych.	15
4.1. Definicja i konsekwencje autokorelacji składników losowych.	15
4.2. Testowanie autokorelacji składników losowych.	16
4.3. Estymacja.	18
5. Heteroskedastyczność składników losowych.	19
5.1. Definicja i konsekwencje heteroskedastyczności składników losowych.	19
5.2. Testowanie heteroskedastyczności składników losowych.	20

1. Model prostej regresji liniowej.	2
5.3. Estymacja.	21
6. Inne zagadnienia dotyczące regresji liniowej.	22
6.1. Postać analityczna.	22
6.2. Zmienne zero-jedynkowe.	24
6.2.1. Dwie grupy obserwacji.	24
6.2.2. Kilka grup obserwacji.	24
6.2.3. Sezonowość.	25
6.3. Przykład.	25
7. Prognozowanie na podstawie modelu regresji liniowej.	25
7.1. Definicja i własności prognoz.	25
7.2. Miary dokładności prognoz.	27
7.3. Przykład.	27
8. Zadanie przygotowawcze do sprawdzianu.	29
A. Tablice statystyczne.	32

1. Model prostej regresji liniowej.

W podręcznikach ekonomii rozważa się zależność między konsumpcją a dochodami. W skali makroekonomicznej zależność tę można sprowadzić do funkcji $y = f(x)$, wiążącej spożycie y i produkt krajowy brutto x . Keynes [10, s. 87 – 88] stwierdza, że krańcowa skłonność do konsumpcji jest dodatnia i mniejsza od jedności, a udział konsumpcji w dochodzie maleje ze wzrostem dochodu. Oznacza to, że

$$0 < f'(x) < 1 \quad \text{i} \quad \left(\frac{f(x)}{x}\right)' = \frac{f'(x) \cdot x - f(x)}{x^2} < 0.$$

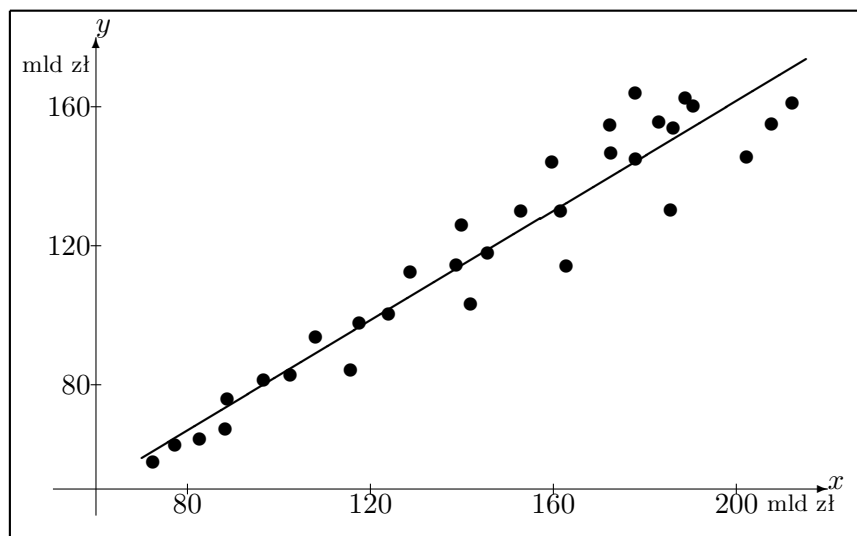
Funkcja liniowa $f(x) = \beta_0 + \beta_1 x$ spełnia te warunki, jeśli $0 < \beta_1 < 1$ i $\beta_0 > 0$; ostatnia nierówność jest równoważna występowaniu autonomicznego spożycia.

Rysunek 1 przedstawia zależność między spożyciem a produktem krajowym brutto w Polsce w latach 1995 – 2002. Punkty nie układają się dokładnie ani wzdłuż prostej, ani żadnej krzywej o niezbyt skomplikowanym przebiegu. Leżą jednak w pobliżu prostej $y = 3,8834 + 0,78744x$. Jak obliczyć jej parametry? Czy przybliży ona zależność dostatecznie dobrze? Co jeszcze można powiedzieć o badanej zależności? Na te pytania odpowiemy w bieżącym i następnych rozdziałach.

Tabela 1. Spożycie (y) i produkt krajowy brutto (x) w Polsce w cenach bieżących w mld zł.

Rok	Kwartał	y	x	Rok	Kwartał	y	x
1995	I	57,7	72,4	1999	I	125,8	139,9
1995	II	62,5	77,3	1999	II	130,1	152,9
1995	III	64,5	82,6	1999	III	130,1	161,5
1995	IV	67,3	88,2	1999	IV	130,2	185,6
1996	I	75,9	88,7	2000	I	144,1	159,6
1996	II	81,5	96,6	2000	II	146,6	172,5
1996	III	82,7	102,4	2000	III	144,9	177,9
1996	IV	84,1	115,7	2000	IV	145,6	202,2
1997	I	93,6	108,0	2001	I	154,8	172,3
1997	II	97,9	117,5	2001	II	155,6	183,1
1997	III	100,4	124,0	2001	III	154,0	186,2
1997	IV	103,4	141,9	2001	IV	155,2	207,7
1998	I	112,5	128,7	2002	I	164,1	177,9
1998	II	114,4	138,8	2002	II	162,5	188,8
1998	III	117,9	145,7	2002	III	160,1	190,5
1998	IV	114,3	162,8	2002	IV	161,1	212,2

Źródło: Obliczenia własne na podstawie [3, 12]. W związku z wprowadzoną przez GUS zmianą metody obliczania spożycia i produktu krajowego, wielkości sprzed 2000 r. wyrażono w warunkach porównywalnych z późniejszymi, mnożąc je przez przez średnią arytmetyczną stosunków czterech wielkości kwartalnych obliczonych według nowej i starej metody dla 2000 r. Średnia ta wynosi 1,049318 dla spożycia i 1,040336 dla produktu krajowego brutto.



Rysunek 1. Spożycie (y) jako funkcja produktu krajowego brutto (x).
Zaznaczono prostą $y = 3,8834 + 0,78744x$.
Źródło: Obliczenia własne na podstawie danych zawartych w tabeli 1.

1.1. Założenia.

Model prostej regresji liniowej, zwany także modelem regresji liniowej jednej zmiennej, można zapisać jako

$$y_i = \beta_0 + \beta_1 x_i + \xi_i, \quad i = 1, \dots, n, \quad (1)$$

gdzie β_0 , β_1 są nieznanymi parametrami, y_i są obserwacjami dokonanyymi na zmiennej objaśnianej y , x_i są obserwacjami dokonanyymi na zmiennej objaśniającej x , a ξ_i są składnikami losowymi. Wartości przyjmowane przez zmienną objaśniającą traktujemy jako stałe liczby, składniki losowe uważamy za niezależne zmienne losowe o jednakowym rozkładzie normalnym o wartości oczekiwanej 0 i wariancji σ^2 . Obserwacje na zmiennej objaśnianej są więc wartościami przyjmowanymi przez zmienne losowe. Parametrami modelu są β_0 , β_1 i σ^2 . Jeśli przez b_0 , b_1 oznaczmy estymatory parametrów β_0 , β_1 , to wartość $\hat{y}_i = b_0 + b_1 x_i$ nazywamy *wartością teoretyczną* zmiennej objaśnianej odpowiadającą i -tej obserwacji, a $e_i = y_i - \hat{y}_i$ — *resztą* odpowiadającą i -tej obserwacji.

1.2. Metoda najmniejszych kwadratów.

Metodę estymacji parametrów β_0, β_1 polegającą na minimalizacji ze względu na b_0, b_1 sumy kwadratów reszt

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$

nazywamy metodą najmniejszych kwadratów. Estymatory otrzymujemy rozwiązując układ *równań normalnych* powstały z przyrównania do zera pochodnych cząstkowych sumy kwadratów reszt ze względu na b_0, b_1 :

$$-2 \sum_{i=1}^n e_i = -2 \sum_{i=1}^n (x_i e_i) = 0, \quad (2)$$

czyli

$$nb_0 + \sum_{i=1}^n (x_i) b_1 = \sum_{i=1}^n y_i, \quad (3)$$

$$\sum_{i=1}^n (x_i) b_0 + \sum_{i=1}^n (x_i^2) b_1 = \sum_{i=1}^n (x_i y_i). \quad (4)$$

Oznaczając $\Delta = n \sum_{i=1}^n (x_i^2) - (\sum_{i=1}^n x_i)^2$, estymatory możemy wyrazić jako

$$b_0 = \left(\sum_{i=1}^n (y_i) \cdot \sum_{i=1}^n (x_i^2) - \sum_{i=1}^n (x_i) \cdot \sum_{i=1}^n (x_i y_i) \right) / \Delta, \quad (5)$$

$$b_1 = \left(n \cdot \sum_{i=1}^n (x_i y_i) - \sum_{i=1}^n (x_i) \cdot \sum_{i=1}^n (y_i) \right) / \Delta. \quad (6)$$

Ponadto, sumując względem i obie strony tożsamości $y_i - \bar{y} = \hat{y}_i - \bar{y} + e_i$ i korzystając z (2), otrzymujemy rozkład sumy kwadratów odchyłeń zmiennej objaśnianej od średniej na sumę kwadratów odchyłeń wartości teoretycznych od średniej i sumę kwadratów reszt:

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{\text{zmiennosc wartosci empirycznych}} = \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{\text{zmiennosc wartosci teoretycznych}} + \underbrace{\sum_{i=1}^n e_i^2}_{\text{zmiennosc reszt}}.$$

Wielkość

$$0 \leq R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \leq 1$$

nazywa się *współczynnikiem determinacji*. Przyjęcie przez współczynnik determinacji wartości 1 oznacza, że wszystkie reszty są równe 0, a więc wszystkie wartości teoretyczne równe są wartościom rzeczywistym; oznacza to stu-procentowe dopasowanie modelu. Przyjęcie przez współczynnik determinacji wartości 0 oznacza, że wszystkie wartości teoretyczne są równe, a więc zmienna objaśniająca w ogóle nie wpływa na zmienną objaśnianą. Na ogół pożądane są wysokie wartości współczynnika determinacji.

Przykład 1. Układem równań normalnych (3) – (4) w przypadku danych z tabeli 1 jest

$$\begin{aligned} 32 \cdot b_0 + 4662,1 \cdot b_1 &= 3795,4, \\ 4662,1 \cdot b_0 + 732311,1 \cdot b_1 &= 594757,0, \end{aligned}$$

a jego rozwiązaniem są estymatory

$$\begin{aligned} b_0 &= 3,8834, \\ b_1 &= 0,78744. \end{aligned}$$

Autonomiczne spożycie zostało oszacowane na poziomie 3,88 mld zł, a krańcowa skłonność do konsumpcji – na poziomie 0,78744. Wzrost produktu krajowego brutto o 1 mld zł powoduje wzrost spożycia przeciętnie o około 787 mln zł.

Współczynnik determinacji modelu wynosi 0,91725, a więc zmienność spożycia została wyjaśniona w około 92 procentach. Pozostała część zmienności nie została wyjaśniona przez model.

1.3. Statystyczne własności estymatorów.

Używając (1) we wzorach (5) – (6) dostajemy

$$\begin{aligned} b_0 &= \beta_0 - (n/\Delta) \sum_{i=1}^n \left[(\bar{x}x_i - (1/n) \sum_{i=1}^n x_i^2) \xi_i \right], \\ b_1 &= \beta_1 + (n/\Delta) \sum_{i=1}^n [(x_i - \bar{x}) \xi_i]. \end{aligned}$$

Łatwo teraz widzieć, że estymatory b_0 , b_1 mają rozkłady normalne odpowiednio $N(\beta_0, \sigma^2 \sum_{i=1}^n (x_i^2)/\Delta)$ i $N(\beta_1, n\sigma^2/\Delta)$, a kowariancja między nimi wynosi $-\sigma^2 \sum_{i=1}^n (x_i)/\Delta$. Żmudniejsze rachunki pozwalają stwierdzić, że

wartością oczekiwaną sumy kwadratów reszt jest $(n - 2)\sigma^2$, zatem

$$s^2 = \frac{1}{n - 2} \sum_{i=1}^n e_i^2$$

jest nieobciążonym estymatorem wariancji σ^2 , a $s_0^2 = s^2 \sum_{i=1}^n (x_i^2) / \Delta$ i $s_1^2 = ns^2 / \Delta$ są nieobciążonymi estymatorami wariancji estymatorów b_0 , b_1 .

Można pokazać, że zmienna $(n - 2)s^2 / \sigma^2$ ma rozkład χ^2 o $n - 2$ stopniach swobody. Stąd i z definicji rozkładu t Studenta wynika, że zmienne $(b_0 - \beta_0) / s_0$ i $(b_1 - \beta_1) / s_1$ mają rozkład t Studenta o $n - 2$ stopniach swobody.

Przykład 2. Oszacowaniem wariancji składników losowych modelu z poprzedniego przykładu jest $s^2 = 98,9866 \cdot 10^{18}$ zł². Przydatniejsze jest oszacowanie odchylenia standardowego składnika losowego $s = \sqrt{s^2}$, zwane *standardowym błędem regresji*, wynoszące 9,9492 mld zł, informujące, o ile przeciętnie wartości teoretyczne konsumpcji odchylają się od wartości rzeczywistych.

Oszacowanie odchylenia standardowego $s_1 = \sqrt{s_1^2}$ estymatora b_1 , zwane *standardowym błędem szacunku* parametru β_1 , wynosi 0,043181. Oznacza to, że wzrost produktu krajowego brutto o 1 mld zł powoduje wzrost spożycia przeciętnie o około 787 mln zł z błędem wynoszącym około 43 mln zł.

1.4. Wnioskowanie statystyczne.

Tablice wartości krytycznych rozkładu t Studenta podają dla różnych ilości stopni swobody n wartości $t(\alpha, n)$ takie, że $P(|t_n| > t(\alpha, n)) = \alpha$, gdzie t_n jest zmienną losową o rozkładzie t Studenta o n stopniach swobody. Pozwala to na budowę przedziałów ufności dla parametrów β_0 , β_1 :

$$P(b_i - s_i t(\alpha, n - 2) \leq \beta_i \leq b_i + s_i t(\alpha, n - 2)) = 1 - \alpha, \quad i = 0, 1. \quad (7)$$

Umożliwia także testowanie hipotezy $H_0: \beta_i = \beta_i^0$ wobec hipotezy alternatywnej $H_1: \beta_i \neq \beta_i^0$. Mianowicie, jeśli

$$\frac{|b_i - \beta_i^0|}{s_i} \leq t(\alpha, n - 2), \quad (8)$$

to nie ma podstaw do odrzucenia hipotezy zerowej przy poziomie istotności α , w przeciwnym razie hipotezę tę należy odrzucić na rzecz alternatywnej. Odrzucenie hipotezy zerowej dla $\beta_1^0 = 0$ powoduje uznanie wpływu zmiennej objaśniającej na zmienną objaśnianą za *statystycznie istotny*.

Tablice wartości krytycznych rozkładu χ^2 podają dla różnych ilości stopni swobody n wartości $\chi^2(\alpha, n)$ takie, że $P(\chi_n^2 > \chi^2(\alpha, n)) = \alpha$, gdzie χ_n^2 jest zmienną losową o rozkładzie χ^2 o n stopniach swobody. Możemy to wykorzystać do budowy przedziału ufności dla wariancji składnika losowego:

$$P\left(\frac{(n-2)s^2}{\chi^2(\alpha_1, n-2)} \leq \sigma^2 \leq \frac{(n-2)s^2}{\chi^2(\alpha_2, n-2)}\right) = \alpha_2 - \alpha_1, \quad (9)$$

gdzie $\alpha_1 < \alpha_2$.

Przykład 3. Wartością krytyczną rozkładu t Studenta o 30 stopniach swobody przy poziomie istotności 0,05 jest $t(0,05, 30) = 2,042$. 95-procentowym przedziałem ufności parametru β_1 jest $[0,6993, 0,8756]$. Pozwala to uściślić interpretację oszacowania parametru β_1 : wzrost produktu krajowego brutto o 1 mld zł powoduje wzrost spożycia o liczbę z prawdopodobieństwem 95% leżącą między 699,3 a 875,6 mln zł. Ponieważ statystyka $|b_1|/s_1 \approx 18,2$ jest znacznie większa od 2,042, uznajemy wpływ produktu krajowego brutto na spożycie za *istotny*.

Wartościami krytycznymi rozkładu χ^2 o 30 stopniach swobody są

$$\chi^2(0,025, 30) = 46,979, \chi^2(0,975, 30) = 16,791.$$

95-procentowym przedziałem ufności dla odchylenia standardowego składnika losowego jest przedział

$$\sqrt{\frac{30 \cdot 9,9492^2}{46,979}} \leq \sigma \leq \sqrt{\frac{30 \cdot 9,9492^2}{16,791}},$$

czyli $[7,9505, 13,2987]$. Przeciętne odchylenie wartości teoretycznych konsumpcji od wartości rzeczywistych zawiera się z prawdopodobieństwem 95% między około 7,95 a 13,30 mld zł.

2. Model wielorakiej regresji liniowej.

Rozszerzymy metody opisane w poprzednim rozdziale na sytuację, gdy przypuszczamy, że na zmienną objaśnianą wpływ ma więcej niż jeden czynnik. Na przykład możemy przypuszczać, że spożycie w każdym kwartale zależy od produktu narodowego brutto nie tylko w tym, ale i poprzednich kwartałach.

2.1. Założenia.

Model prostej regresji liniowej (1) jest szczególnym przypadkiem modelu wielorakiej regresji liniowej, zwanego także modelem regresji liniowej wielu zmiennych

$$y_i = \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \xi_i, \quad i = 1, \dots, n, \quad (10)$$

gdzie β_1, \dots, β_k są parametrami, x_{i1}, \dots, x_{ik} są obserwacjami dokonanymi na zmiennych objaśniających x_1, \dots, x_k , a znaczenie wielkości y_i i ξ_i jest takie, jak w (1). Parametrami modelu są β_1, \dots, β_k i σ^2 . Model (1) uzyskujemy przyjmując $k = 2$ i $x_{i1} \equiv 1$. Parametr przy zmiennej tożsamościowo równej 1 nazywa się *wyrazem wolnym*.

Wprowadzając oznaczenia macierzowe $\mathbf{y} = [y_i]_{n \times 1}$, $\mathbf{X} = [x_{ij}]_{n \times k}$, $\boldsymbol{\beta} = [\beta_j]_{k \times 1}$, $\boldsymbol{\xi} = [\xi_i]_{n \times 1}$, można (10) wyrazić zwięźle jako

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi}. \quad (11)$$

Założenie o niezależności i jednakowej normalności składników losowych oznacza, że $\boldsymbol{\xi}$ ma n -wymiarowy rozkład normalny o wartości oczekiwanej $\mathbf{0} = [0]_{n \times 1}$ i macierzy kowariancji $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_n$, gdzie \mathbf{I}_n jest macierzą jednostkową stopnia n . Jeśli $\mathbf{b} = [b_j]_{k \times 1}$ jest estymatorem wektora $\boldsymbol{\beta}$, to $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$ jest wektorem wartości teoretycznych, a $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ – wektorem reszt.

2.2. Metoda najmniejszych kwadratów.

Minimalizacja sumy kwadratów reszt

$$\mathbf{e}^T \mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\mathbf{b})$$

ze względu na składowe b_1, \dots, b_k wektora \mathbf{b} prowadzi do układu k równań normalnych

$$\mathbf{X}^T \mathbf{X}\mathbf{b} = \mathbf{X}^T \mathbf{y}.$$

W przypadku, gdy istnieje macierz odwrotna do $\mathbf{X}^T \mathbf{X}$, jednoznacznym rozwiązaniem tego układu jest estymator metody najmniejszych kwadratów

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}. \quad (12)$$

Wzór ten jest odpowiednikiem wzorów (5) – (6).

Współczynnik determinacji

$$R^2 = 1 - \frac{\mathbf{e}^T \mathbf{e}}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (13)$$

ma, w przypadku występowania w (10) wyrazu wolnego, te same własności i interpretację, co w przypadku modelu regresji prostej.

2.3. Statystyczne własności estymatorów.

Podstawiając (11) w (12) otrzymujemy

$$\mathbf{b} = \boldsymbol{\beta} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\xi}, \quad (14)$$

skąd wynika, że wektor \mathbf{b} ma rozkład normalny z wartością oczekiwaną $\boldsymbol{\beta}$ i macierzą kowariancji $\sigma^2(\mathbf{X}^T \mathbf{X})^{-1}$.

Można pokazać¹, że nieobciążonym estymatorem wariancji składnika losowego σ^2 jest

$$s^2 = \frac{\mathbf{e}^T \mathbf{e}}{n - k}, \quad (15)$$

nieobciążonymi estymatorami wariancji estymatorów b_j są diagonalne elementy s_j^2 macierzy $s^2(\mathbf{X}^T \mathbf{X})^{-1}$, statystyka

$$(n - k)s^2/\sigma^2 \quad (16)$$

ma rozkład χ^2 o $n - k$ stopniach swobody, a statystyki

$$t_j = \frac{b_j - \beta_j}{s_j}, \quad j = 1, \dots, k,$$

mają rozkład t Studenta o $n - k$ stopniach swobody.

2.4. Wnioskowanie statystyczne.

Wszystkie fakty dotyczące wnioskowania statystycznego, przedstawione w punkcie 1.4, są prawdziwe w odniesieniu do modelu regresji liniowej k zmiennych objaśniających. Należy jedynie we wzorach (7), (8) i (9) zastąpić liczbę stopni swobody $n - 2$ liczbą $n - k$.

Przykład 4. Zajmijmy się modelem (10) dla danych z tabeli 1, przyjmując $n = 29$, $k = 5$, $x_1 \equiv 1$, a za x_2, x_3, x_4, x_5 biorąc odpowiednio wartość produktu narodowego brutto w bieżącym, poprzednim i dwóch kolejno wcześniejszych kwartałach. Pomijamy obserwacje na zmiennej objaśnianej z pierwszych trzech kwartałów 1995 r., gdyż nie mamy danych o produkcji krajowym brutto w trzech ostatnich kwartałach 1994 r. Wyniki oszacowania parametrów modelu przedstawia tabela 2.

W stosunku do modelu z jedną zmienną objaśniającą (1), otrzymaliśmy znacznie większą wartość współczynnika determinacji: kształtowanie się spożycia jest przez model wyjaśnione w 99,6%. Wartości teoretyczne spożycia

¹Patrz na przykład [16, s. 70].

Tabela 2. Wyniki oszacowania parametrów modelu objaśniającego zależność spożycia od produktu krajowego brutto w bieżącym i trzech bezpośrednio poprzedzających kwartałach.

Zmienna	Parametr	Błąd szacunku	Statystyka t
1	8,6684	1,7075	5,0767
x_2	0,00495	0,03094	0,1600
x_3	0,26053	0,03271	7,9651
x_4	0,32693	0,03215	10,1704
x_5	0,20837	0,03122	6,6733
Liczba obserwacji:			29
Suma kwadratów reszt:			98,1905
Współczynnik determinacji:			0,9961
Standardowy błąd regresji:			2,0227
Statystyka istotności regresji:			1527,3
Statystyka d Durbina-Watsona:			1,4294

Źródło: Obliczenia własne na podstawie danych zawartych w tabeli 1.

odchylają się od rzeczywistych przeciętnie nie o 9,9492 mld zł, lecz tylko o 2,0227 mld zł. Wartością krytyczną statystyki t Studenta przy poziomie istotności 0,05 i 24 stopniach swobody jest 2,064. Ponieważ wartością statystyki t dla sprawdzenia hipotezy, że zmienna x_2 wpływa istotnie na spożycie, jest 0,16, nie mamy podstaw do odrzucenia tej hipotezy. Natomiast odrzucamy analogiczne hipotezy w stosunku do pozostałych zmiennych. W świetle rozważanego modelu istotny wpływ na wielkość spożycia ma produkt krajowy brutto w trzech wcześniejszych kwartałach, nie ma zaś wpływu jego bieżąca wartość.

2.5. Test istotności regresji.

W następnym rozdziale wykażemy, że gdy w modelu (10) uwzględniono wyraz wolny $x_{i1} \equiv 1$, to przy prawdziwości hipotezy

$$H_0: \beta_2 = \beta_3 = \dots = \beta_k = 0$$

statystyka

$$\frac{R^2/(k-1)}{(1-R^2)/(n-k)} \quad (17)$$

ma rozkład $F(k-1, n-k)$. Statystykę tę wykorzystuje się do testowania hipotezy o łącznym wpływie zmiennych objaśniających na zmienną objaśnianą, czyli do testowania *istotności regresji*. Duże wartości R^2 towarzyszą dużym wartościom statystyki testowej i na odwrót: ze znikania współczynnika determinacji wynika zerowość statystyki (17).

Jeśli przy przyjętym poziomie istotności α wartość statystyki (17) przekroczy wartość krytyczną $F(\alpha, k-1, n-k)$ statystyki F o $k-1$ stopniach swobody licznika i $n-k$ stopniach swobody mianownika², to hipotezę H_0 należy odrzucić. W przeciwnym przypadku nie ma podstaw do odrzucenia hipotezy H_0 . Uznajemy, że zmienne objaśniające *łącznie istotnie* wpływają na zmienną objaśnianą.

Przykład 5. W przykładzie w punkcie 2.4 wartość statystyki istotności regresji (17) wynosi 1527,3 i jest znacznie większa od wartości krytycznej $F(0,05, 4, 24) = 2,776$, a więc hipotezę o braku łącznego wpływu zmiennych objaśniających na spożycie odrzucamy. Jest tak mimo że nie odrzuciliśmy wyżej hipotezy o braku wpływu bieżącej wartości produktu krajowego brutto.

3. Sprawdzanie hipotez.

3.1. Testowanie ograniczeń liniowych.

Obecnie zajmiemy się weryfikacją hipotezy

$$H_0: \mathbf{R}\boldsymbol{\beta} = \mathbf{r}, \quad (18)$$

gdzie $\mathbf{R} = [r_{ij}]_{m \times k}$ jest macierzą rzędu $1 \leq m \leq k$, a $\mathbf{r} = [r_i]_{m \times 1}$ jest wektorem. Jeśli hipoteza (18) jest prawdziwa, to z (14) wnioskujemy, że zmienna losowa $\mathbf{R}\mathbf{b} - \mathbf{r}$ ma rozkład normalny $N(\mathbf{0}, \mathbf{V})$, gdzie $\mathbf{V} = \sigma^2 \mathbf{R}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{R}^T$. Niech $\mathbf{V}^{-1} = \mathbf{P}^T \boldsymbol{\Lambda}^T \boldsymbol{\Lambda} \mathbf{P}$, gdzie $\mathbf{P}^T = \mathbf{P}^{-1}$, a $\boldsymbol{\Lambda}$ jest macierzą diagonalną. Wówczas

$$(\mathbf{R}\mathbf{b} - \mathbf{r})^T \mathbf{V}^{-1} (\mathbf{R}\mathbf{b} - \mathbf{r}) = [\boldsymbol{\Lambda} \mathbf{P} (\mathbf{R}\mathbf{b} - \mathbf{r})]^T [\boldsymbol{\Lambda} \mathbf{P} (\mathbf{R}\mathbf{b} - \mathbf{r})],$$

²Patrz tablica III.

a – jak łatwo sprawdzić – $\Lambda P(\mathbf{Rb} - \mathbf{r})$ ma rozkład $N(\mathbf{0}, \mathbf{I})$. Z definicji rozkładu χ^2 wynika więc, że zmienna

$$\frac{(\mathbf{Rb} - \mathbf{r})^T [\mathbf{R}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{R}^T]^{-1} (\mathbf{Rb} - \mathbf{r})}{\sigma^2} \quad (19)$$

ma rozkład χ^2 o m stopniach swobody. Z punktu 2.3 wiemy, że (16) ma rozkład χ^2 o $n - k$ stopniach swobody. Łatwo też wykazać, że kowariancja zmiennych $\mathbf{Rb} - \mathbf{r} = \mathbf{R}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\xi}$ i $\mathbf{e} = [\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T] \boldsymbol{\xi}$ jest równa $\mathbf{0}$, więc są one niezależne. Zatem z definicji rozkładu F wynika, że iloraz zmiennych (19) i (16), tzn.

$$\frac{(\mathbf{Rb} - \mathbf{r})^T [\mathbf{R}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{R}^T]^{-1} (\mathbf{Rb} - \mathbf{r}) / m}{\mathbf{e}^T \mathbf{e}} \quad (20)$$

ma rozkład F o m stopniach swobody licznika i $n - k$ stopniach swobody mianownika. Hipotezę (18) należy więc przy poziomie istotności α odrzucić, gdy wartość statystyki (20) przekracza wartość krytyczną $F(\alpha, m, n - k)$.

Można pokazać, że gdy testowaną hipotezą jest $H_0: \boldsymbol{\beta}_1 = \mathbf{0}$, gdzie $\boldsymbol{\beta}_1$ jest podwektorem wektora $\boldsymbol{\beta}$ złożonym z jego m pierwszych składowych, to znaczy hipoteza (18) przy $\mathbf{R} = [\mathbf{I}_m \ \mathbf{0}]$ i $\mathbf{r} = \mathbf{0}$, to statystyka (20) jest równa statystyce

$$\frac{(\mathbf{e}_*^T \mathbf{e}_* - \mathbf{e}^T \mathbf{e}) / m}{\mathbf{e}^T \mathbf{e} / (n - k)}, \quad (21)$$

gdzie \mathbf{e}_* jest wektorem reszt z regresji \mathbf{y} względem ostatnich $k - m$ zmiennych objaśniających. Statystykę tę można też zapisać przy pomocy współczynników determinacji z obu regresji, przyjmując $R^2 = 1 - \mathbf{e}^T \mathbf{e} / \sum_{i=1}^n (y_i - \bar{y})^2$ i $R_*^2 = 1 - \mathbf{e}_*^T \mathbf{e}_* / \sum_{i=1}^n (y_i - \bar{y})^2$:

$$\frac{(R^2 - R_*^2) / m}{(1 - R^2) / (n - k)}, \quad (22)$$

3.2. Testowanie zmian strukturalnych.

Testowanie zmian strukturalnych zwane jest także testowaniem stabilności modelu. W dwóch następnych podpunktach opiszemy test Chowa stabilności wektora współczynników regresji i test Goldfelda-Quandt stabilności wariancji składników losowych.

3.2.1. Różne wektory parametrów.

Przykładem zastosowania testu opartego na statystyce (21) jest zagadnienie równości parametrów w dwóch podzbiorach zbioru obserwacji. Założ-

my, że pierwszy podzbiór tworzy n_1 obserwacji, drugi – n_2 obserwacji, przy czym zachowany jest warunek $n_1 + n_2 = n$. Niech \mathbf{y}_i , \mathbf{X}_i , $\boldsymbol{\beta}_i$ i $\boldsymbol{\xi}_i$, $i = 1, 2$, będą odpowiednio wektorami obserwacji na zmiennej objaśnianej, macierzami obserwacji na zmiennych objaśniających, wektorami parametrów i wektorami składników losowych. Chcemy sprawdzić hipotezę

$$H_0: \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2. \quad (23)$$

Zapisując model regresji blokowo:

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_2 \end{bmatrix} \quad (24)$$

widzimy, że sprawdzanie hipotezy (23) jest równoważne sprawdzaniu hipotezy (18) przy $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$, $\mathbf{R} = [\mathbf{I}_k \quad -\mathbf{I}_k]$ i $\mathbf{r} = \mathbf{0}$. Wobec tego statystyką testową jest statystyka (21)

$$\frac{(\mathbf{e}_*^T \mathbf{e}_* - \mathbf{e}^T \mathbf{e})/k}{\mathbf{e}^T \mathbf{e}/(n - 2k)},$$

gdzie \mathbf{e} jest wektorem reszt modelu (24), a \mathbf{e}_* jest wektorem reszt tego modelu przy założeniu, że $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$.

3.2.2. Różne wariancje składników losowych.

Załóżmy, że możemy wyróżnić dwa podzbiory zbioru obserwacji o licznosciach n_1 i n_2 , w których składniki losowe mają łączny rozkład normalny o wartości oczekiwanej $\mathbf{0}$ i macierzy kowariancji

$$\begin{bmatrix} \sigma_1^2 \mathbf{I}_{n_1} & \mathbf{0} \\ \mathbf{0} & \sigma_2^2 \mathbf{I}_{n_2} \end{bmatrix}.$$

Zgodnie z punktem 2.3, zmienne $\mathbf{e}_i^T \mathbf{e}_i / \sigma_i^2$, $i = 1, 2$, mają rozkłady χ^2 o $n_i - k$ stopniach swobody. Z ich niezależności wynika zatem, że zmienna

$$\frac{(\mathbf{e}_1^T \mathbf{e}_1 / \sigma_1^2) / (n_1 - k)}{(\mathbf{e}_2^T \mathbf{e}_2 / \sigma_2^2) / (n_2 - k)}$$

ma rozkład F o $n_1 - k$ stopniach swobody licznika i $n_2 - k$ stopniach swobody mianownika. Przy prawdziwości hipotezy $H_0: \sigma_1^2 = \sigma_2^2$, zmienna ta redukuje się do statystyki

$$s_1^2 / s_2^2, \quad (25)$$

mającej ten sam rozkład.

Przykład 6. W przykładzie w punkcie 2.4 postawiliśmy hipotezę, że na wielkość spożycia w danym kwartale ma wpływ produkt krajowy brutto z trzech kwartałów poprzedzających, a nie ma wpływu wielkość produktu krajowego brutto osiągnięta w tym kwartale. Mamy więc dwa konkurencyjne modele funkcji konsumpcji, zakładające, że w modelu

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \xi_i \quad (26)$$

jest albo

$$\beta_3 = \beta_4 = \beta_5 = 0 \quad (\text{model Keynes'a}), \quad (27)$$

albo

$$\beta_2 = 0. \quad (28)$$

Zastosujemy test oparty na statystyce (21). Suma kwadratów reszt modelu (26) wynosi 98,1905, suma kwadratów reszt modelu z ograniczeniami (27) wynosi 2916,9 a suma kwadratów reszt modelu z ograniczeniem (28) – 98,2953. Wartość statystyki (21) w przypadku modelu (27) jest równa 229,65, co przy wartości krytycznej $F(0,05, 3, 24) = 3,009$ nakazuje odrzucić hipotezę (27). W przypadku modelu (28) wartością statystyki testowej jest 0,0256 i przy wartości krytycznej $F(0,05, 1, 24) = 4,260$ nie ma podstaw do odrzucenia hipotezy (28). Przeprowadzone w tym przykładzie rozumowanie wskazuje, że lepsza jest funkcja konsumpcji spełniająca warunek (28).

4. Autokorelacja składników losowych.

4.1. Definicja i konsekwencje autokorelacji składników losowych.

W odniesieniu do modelu regresji liniowej, autokorelacja składników losowych oznacza naruszenie założenia o niezależności składników losowych bez naruszenia założenia o ich jednakowej normalności. Mówimy, że model (11)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi}$$

charakteryzuje się występowaniem autokorelacji, jeśli $\boldsymbol{\xi}$ ma n -wymiarowy rozkład normalny o wartości oczekiwanej $\mathbf{0}$ i macierzy kowariancji $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{\Omega}$, gdzie $\boldsymbol{\Omega} = [\omega_{ij}]_{n \times n}$ jest dodatnio określoną macierzą symetryczną taką, że $\omega_{ii} = 1$ dla $i = 1, \dots, n$ i $\omega_{ij} \neq 0$ dla pewnych $1 \leq i < j \leq n$.

Najczęściej rozważa się szczególny przypadek autokorelacji, w którym zakłada się, że dla $p \geq 1$ istnieją liczby ρ_p takie, że

$$\forall(i, j = 1, \dots, n) (|i - j| = p \Rightarrow E(\xi_i \xi_j) = \sigma^2 \rho_p).$$

Liczby ρ_p są współczynnikami korelacji par zmiennych losowych ξ_i, ξ_j dla $|i - j| = p$ i nazywają się współczynnikami autokorelacji rzędu p . Występowanie autokorelacji rzędu p oznacza, że $\rho_p \neq 0$. Macierz $\mathbf{\Omega}$ ma w tym przypadku postać

$$\mathbf{\Omega} = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{n-2} & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \dots & \rho_{n-3} & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \dots & \rho_{n-4} & \rho_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{n-2} & \rho_{n-3} & \rho_{n-4} & \dots & 1 & \rho_1 \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \dots & \rho_1 & 1 \end{bmatrix}.$$

Można pokazać, że gdy

$$\xi_i = \rho \xi_{i-1} + \varepsilon_i, \quad -\infty < i < +\infty, \quad (29)$$

gdzie $|\rho| < 1$, a ε_i są niezależnymi zmiennymi losowymi o wartości oczekiwanej 0 i wariancji σ_ε^2 , to $\rho_p = \rho^p$ dla $0 \leq p < n$, a

$$\sigma^2 = \frac{\sigma_\varepsilon^2}{1 - \rho^2}.$$

Występowanie autokorelacji powoduje, że estymatory metody najmniejszych kwadratów tracą niektóre spośród swoich pożądaných właściwości statystycznych. Mianowicie estymatory współczynników β , mimo że pozostają nieobciążone, nie są estymatorami o najmniejszej wariancji. Diagonalne elementy macierzy $\sigma^2(\mathbf{X}^T \mathbf{X})^{-1}$ nie doszacowują prawdziwej wariancji estymatorów metody najmniejszych kwadratów. Ponadto rozkłady t Studenta i F , używane do wnioskowania statystycznego w modelu regresji liniowej, tracą swoje uzasadnienie.

4.2. Testowanie autokorelacji składników losowych.

Często stosowanym testem hipotezy $H_0: \rho = 0$, gdzie ρ jest określone przez (29), jest test Durбина-Watsona [5, 6] oparty na statystyce

$$d = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}, \quad (30)$$

gdzie e_i są resztami z regresji (11), w której $x_{i1} \equiv 1$. Dla różnych liczb obserwacji n i zmiennych objaśniających $k' = k - 1$ oraz poziomów istotności α równych 0,05 i 0,01, tablice³ podają wartości $d_L(\alpha, n, k')$ i $d_U(\alpha, n, k')$, zwane odpowiednio kresem dolnym i górnym. Testowanie hipotezy H_0 wobec hipotezy alternatywnej $H_1: \rho > 0$ przebiega w ten sposób, że gdy $d < d_L(\alpha, n, k')$, to hipotezę H_0 należy odrzucić na rzecz hipotezy H_1 , a gdy $d > d_U(\alpha, n, k')$, to nie ma podstaw do odrzucenia hipotezy H_0 . W przypadku, gdy $d_L(\alpha, n, k') \leq d \leq d_U(\alpha, n, k')$, test nie rozstrzyga hipotezy. Testowanie hipotezy H_0 wobec hipotezy alternatywnej $H_1: \rho < 0$ przebiega podobnie, przy czym zamiast statystyki testowej d należy wykorzystać statystykę $d' = 4 - d$: jeśli $d' < d_L(\alpha, n, k')$, to hipotezę H_0 należy odrzucić na rzecz hipotezy H_1 , a gdy $d' > d_U(\alpha, n, k')$, to nie ma podstaw do odrzucenia hipotezy H_0 . Przypadek $d_L(\alpha, n, k') \leq d' \leq d_U(\alpha, n, k')$ również nie rozstrzyga hipotezy.

Przykład 7. Kontynuujemy przykłady z punktów 1.2 i 2.4. Wartość statystyki Durбина-Watsona dla modelu

$$\hat{y}_i = 3,8834 + 0,78744x_{i2}$$

wynosi $d = 2,36$, a kresy dolne i górne, odczytane z tablic, są równe

$$\begin{aligned} d_L(0,05, 32, 1) &= 1,373, \\ d_U(0,05, 32, 1) &= 1,502, \\ d_L(0,01, 32, 1) &= 1,160, \\ d_U(0,01, 32, 1) &= 1,282. \end{aligned}$$

Ponieważ, przy obu poziomach istotności, zarówno d jak i d' są większe od odpowiednich kresów górnych, nie ma podstaw do odrzucenia hipotezy o braku autokorelacji pierwszego rzędu składników losowych.

Wartość statystyki Durбина-Watsona dla składników losowych modelu

$$\hat{y}_i = 8,6684 + 0,00495x_{i2} + 0,26053x_{i3} + 0,32693x_{i4} + 0,20837x_{i5},$$

przytoczona w tabeli 2, jest równa $d = 1,4294$, a jej kresy wynoszą

$$\begin{aligned} d_L(0,05, 29, 4) &= 1,124, \\ d_U(0,05, 29, 4) &= 1,743, \\ d_L(0,01, 29, 4) &= 0,921, \\ d_U(0,01, 29, 4) &= 1,512. \end{aligned}$$

³Patrz [15, str. 195 – 200], [4, str. 492 – 493], [19, str. 279 – 284], [14, str. 1992 – 1995].

Wartość d dla obu poziomów istotności leży między kresem górnym a dolnym, test nie rozstrzyga więc hipotezy o braku autokorelacji lub występowaniu autokorelacji dodatniej. Ponieważ $d' = 2,5706$ jest większe od kresów górnych dla obu poziomów istotności, nie ma podstaw do odrzucenia hipotezy o braku autokorelacji na rzecz hipotezy o występowaniu autokorelacji ujemnej.

4.3. Estymacja.

Załóżmy, że macierz Ω jest znana, a $P_{n \times n}$ jest taką macierzą, że $P^T P = \Omega^{-1}$. Mnożąc równanie (11)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi}$$

lewostronnie przez P dostajemy

$$\mathbf{y}' = \mathbf{X}'\boldsymbol{\beta} + \boldsymbol{\xi}', \quad (31)$$

gdzie $\mathbf{y}' = P\mathbf{y}$, $\mathbf{X}' = P\mathbf{X}$, $\boldsymbol{\xi}' = P\boldsymbol{\xi}$, przy czym macierzą wariancji i kowariancji wektora $\boldsymbol{\xi}'$ jest

$$E\boldsymbol{\xi}'\boldsymbol{\xi}'^T = P \cdot E\boldsymbol{\xi}\boldsymbol{\xi}^T \cdot P^T = \sigma^2 P\Omega P^T = \sigma^2 I_n.$$

Równanie 31 spełnia zatem założenia o niezależności i jednakowej normalności składników losowych $\boldsymbol{\xi}'$, a estymatorem metody najmniejszych kwadratów wektora $\boldsymbol{\beta}$ jest

$$\begin{aligned} \mathbf{b} &= (\mathbf{X}'^T \mathbf{X}')^{-1} \mathbf{X}'^T \mathbf{y}' \\ &= (\mathbf{X}^T P^T P \mathbf{X})^{-1} \mathbf{X}^T P^T P \mathbf{y} \\ &= (\mathbf{X}^T \Omega^{-1} \mathbf{X})^{-1} \mathbf{X}^T \Omega^{-1} \mathbf{y}. \end{aligned} \quad (32)$$

Łatwo sprawdzić, że jest on estymatorem nieobciążonym o macierzy wariancji i kowariancji

$$\sigma^2 (\mathbf{X}^T \Omega^{-1} \mathbf{X})^{-1}. \quad (33)$$

Estymator ten jest zwany estymatorem *uogólnionej metody najmniejszych kwadratów*.

Na ogół jednak macierze Ω , a zatem i P , są nieznane. Naturalne jest zastąpienie obu tych macierzy ich estymatorami. W przypadku rozpatrywanego wcześniej schematu autokorelacji (29), macierz Ω zależy od jednego nieznanego parametru ρ . Zastępując go jego zgodnym estymatorem

$$\hat{\rho} = \frac{\sum_{i=2}^n (e_i e_{i-1})}{\sum_{i=1}^n e_i^2},$$

dostajemy pewien estymator $\hat{\Omega}$ macierzy Ω . Można pokazać⁴, że granicznym rozkładem estymatora

$$\mathbf{b}_n = (\mathbf{X}^T \hat{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\Omega}^{-1} \mathbf{y} \quad (34)$$

przy $n \rightarrow \infty$ jest rozkład normalny o wartości oczekiwanej β i macierzy wariancji i kowariancji (33). Estymator \mathbf{b}_n jest więc zgodny.

5. Heteroskedastyczność składników losowych.

5.1. Definicja i konsekwencje heteroskedastyczności składników losowych.

Heteroskedastyczność składników losowych modelu regresji liniowej (11)

$$\mathbf{y} = \mathbf{X}\beta + \xi$$

oznacza niejednakowość wariancji tych składników. W przeciwieństwie do założenia o stałości wariancji składników losowych (*homoskedastyczności*), przyjętego przez nas w poprzednich rozdziałach, będziemy teraz zakładali, że macierz wariancji i kowariancji składników losowych jest postaci $\Sigma = \sigma^2 \Omega$, gdzie

$$\Omega = \begin{bmatrix} \omega_1 & & \\ & \ddots & \\ & & \omega_n \end{bmatrix},$$

przy czym

$$\sum_{i=1}^n \omega_i = n, \quad \omega_i > 0,$$

a elementy poza główną przekątną macierzy Ω są równe 0, co oznacza nieskorelowanie składników losowych.

Podobnie jak występowanie autokorelacji, występowanie heteroskedastyczności powoduje, że estymator metody najmniejszych kwadratów (12)

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

wektora β , w dalszym ciągu nieobciążony, ma macierz wariancji i kowariancji

$$\sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \Omega \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}.$$

⁴Patrz na przykład [4, rozdział 3.2].

Można pokazać, że estymator uogólnionej metody najmniejszych kwadratów (32)

$$\mathbf{b} = (\mathbf{X}^T \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Omega}^{-1} \mathbf{y}.$$

ma macierz wariancji i kowariancji mniejszą o pewną macierz określoną nieujemnie, jest więc efektywniejszy.

5.2. Testowanie heteroskedastyczności składników losowych.

Do weryfikacji hipotezy o homoskedastyczności składników losowych stosuje się między innymi test Goldfelda-Quandta [8]. Wyróżniwszy dwa podzbiory obserwacji o liczebnościach n_1 , n_2 i wariancjach składników losowych σ_1^2 , σ_2^2 , szacujemy metodą najmniejszych kwadratów parametry równania (11) dla każdego podzbioru oddzielnie, otrzymując estymatory s_1^2 , s_2^2 odpowiednich wariancji składników losowych. Jak wiadomo, statystyki

$$(n_i - k)s_i^2/\sigma_i^2, \quad i = 1, 2,$$

mają rozkłady χ^2 o $n_i - k$ stopniach swobody (por. (16)) i są niezależne, zatem iloraz

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

ma rozkład F o $n_1 - k$ stopniach swobody licznika i $n_2 - k$ stopniach swobody mianownika. Przy założeniu prawdziwości hipotezy $H_0: \sigma_1^2 = \sigma_2^2$, zbyt duże wartości statystyki testowej

$$\frac{s_1^2}{s_2^2} \tag{35}$$

powinny powodować poddanie w wątpliwość hipotezy H_0 . Dokładniej, jeśli $s_1^2/s_2^2 > F(\alpha, n_1 - k, n_2 - k)$, gdzie α jest przyjętym poziomem istotności, to hipotezę o jednakowości wariancji w obu podpróbach należy odrzucić.

Zauważmy, że ponieważ odrzucenie hipotezy H_0 powodują wartości statystyki (35) większe od 1, to w razie potrzeby podpróby należy przenumerać.

Przykład 8. W dalszym ciągu kontynuujemy przykład z punktu 1.2. Patrząc na wykres 1 można zauważyć, że im większe są wartości produktu krajowego brutto, tym większy wydaje się rozrzut spożycia po obu stronach prostej regresji

$$\hat{y}_i = 3,8834 + 0,78744x_{i2}.$$

Być może jest więc tak, że dużym wartościom zmiennej objaśniającej towarzyszy duża wariancja składnika losowego σ_1^2 , a małym – mniejsza wariancja σ_2^2 . Aby to zbadać, posłużymy się testem Goldfelda-Quandta.

Podzielimy zbiór obserwacji na dwa podzbiory: pierwszy, liczący $n_1 = 14$ obserwacji, do którego włączamy te kwartały, w których produkt krajowy brutto był większy niż 160 mld zł i drugi, zawierający $n_2 = 10$ obserwacji, w których produkt krajowy był mniejszy od 120 mld zł. Szacujemy parametry modelu (1) dla każdego podzbioru obserwacji oddzielnie i uzyskujemy standardowe błędy regresji odpowiednio równe $s_1 = 12,6986$ i $s_2 = 4,8492$. Wartością statystyki (35), mającej rozkład F o 12 stopniach swobody licznika i 8 stopniach swobody mianownika jest 6,8576 i wartość ta jest większa od wartości krytycznej $F(0,01, 12, 8) = 5,667$, odczytanej z tablic. Przy poziomie istotności 0,01 hipotezę o jednakowości wariancji składników losowych modelu (1) w obu podpróbach należy odrzucić. Heteroskedastyczność występująca w modelu (1) podważa słuszność wnioskowania statystycznego przeprowadzonego dotychczas na podstawie tego modelu.

5.3. Estymacja.

W przypadku, gdy elementy $\omega_1, \omega_2, \dots, \omega_n$ macierzy $\mathbf{\Omega}$ są znane, nieobciążonym estymatorem wektora β o najmniejszej wariancji jest estymator uogólnionej metody najmniejszych kwadratów (32). W tym przypadku możliwe jest także inne podejście. Załóżmy na przykład [7, str. 304], że wiemy, iż w modelu

$$y_i = \alpha + \beta x_i + \xi_i \quad (36)$$

wariancje składników losowych są proporcjonalne do kwadratów wartości zmiennej objaśniającej x_i , tj. $\sigma_i^2 = \sigma^2 x_i^2$. Dzieląc obie strony (36) przez x_i , otrzymujemy model

$$y'_i = \beta + \alpha x'_i + \xi'_i,$$

gdzie $y'_i = y_i/x_i$, $x'_i = 1/x_i$, $\xi'_i = \xi_i/x_i$. Wariancje składników ξ'_i są stałe, a więc do oszacowania parametrów modelu można użyć metody najmniejszych kwadratów. Przypadek modelu (36) ma częste zastosowanie. Jego uzasadnienie jest zależne od kontekstu, na przykład [16, str. 108]:

wariancja konsumpcji wśród rodzin o wysokich dochodach jest wyższa niż wśród gospodarstw domowych o niskich dochodach; wariancja zysków rośnie wraz ze wzrostem rozmiarów firmy; wariancja płac jest wyższa w regionach bogatszych niż biednych etc.

Na ogół elementy macierzy Ω są nieznane. Macierz ta zależy od zbyt wielu parametrów, aby można było sobie pozwolić na ich szacowanie. Dlatego niezbędne są jakieś założenia o ich strukturze. Zajmiemy się przypadkiem, gdy wariancje składników losowych są stałe w m podpróbach. Szacujemy wówczas m regresji, otrzymując m zgodnych estymatorów s_1^2, \dots, s_m^2 nieznanymi wariancjami. Estymatory te wykorzystujemy do otrzymania estymatora $\hat{\Omega}$ macierzy Ω , który wykorzystujemy do obliczenia zgodnego estymatora (34)

$$\mathbf{b}_n = (\mathbf{X}^T \hat{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\Omega}^{-1} \mathbf{y}.$$

6. Inne zagadnienia dotyczące regresji liniowej.

6.1. Postać analityczna.

Załóżmy, że dysponujemy obserwacjami na zmiennej objaśnianej y i k zmiennych objaśniających x_1, \dots, x_k . Utwórzmy przy pomocy pewnych funkcji $\psi, \phi_1, \dots, \phi_k$, nową zmienną objaśnianą

$$y^* = \psi(x_1, \dots, x_k, y) \quad (37)$$

i nowe zmienne objaśniające

$$x_i^* = \phi(x_1, \dots, x_k), \quad i = 1, \dots, k. \quad (38)$$

Zakładamy, że funkcje $\psi, \phi_1, \dots, \phi_k$ mają wszystkie potrzebne w dalszym ciągu własności, m. in. własność liniowej niezależności. Możemy wówczas rozpatrywać model regresji liniowej

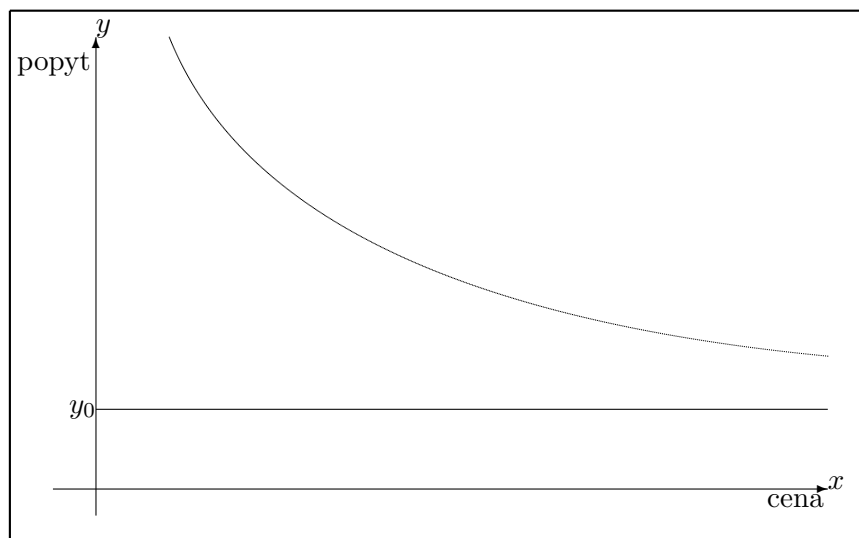
$$\mathbf{y}^* = \mathbf{X}^* \boldsymbol{\beta}^* + \boldsymbol{\xi}^*,$$

gdzie \mathbf{y}^* , \mathbf{X}^* , $\boldsymbol{\beta}^*$, $\boldsymbol{\xi}^*$ mają znaczenie analogiczne do znaczenia \mathbf{y} , \mathbf{X} , $\boldsymbol{\beta}$, $\boldsymbol{\xi}$ modelu (11).

Rozpatrywanie zamiany zmiennych (37 – 38) wynika z potrzeby zastosowania modelu regresji liniowej do zależności nieliniowej, wynikającej z ekonomii. Przykładowo, nieliniowa funkcja popytu [2, rys. 4.2] mogłaby być przybliżana funkcją $y = y_0 + \alpha/x$, gdzie y i x są odpowiednio popytem i ceną, a y_0 i α – dodatnimi parametrami (patrz rysunek 2). Przyjmując w (37 – 38) $\psi(x, y) = y$, $\phi(x) = 1/x$, otrzymujemy model regresji liniowej

$$y_i^* = \beta_0^* + \beta_1^* x_i^* + \xi_i^*,$$

gdzie $y_i^* = y_i$, $x_i^* = 1/x_i$. Zauważmy, że parametry β_0^* i β_1^* odpowiadają wzajemnie jednoznacznie parametrom y_0 i α .

Rysunek 2. Nieliniowa funkcja popytu $y = y_0 + \alpha/x$.

Do najczęściej stosowanych postaci analitycznych modeli należy postać liniowa względem logarytmów zmiennych:

$$\ln y = \beta_0 + \beta_1 \ln x_1 + \dots + \beta_k \ln x_k, \quad (39)$$

odpowiadająca funkcji potęgowej

$$y = e^{\beta_0} x_1^{\beta_1} \dots x_k^{\beta_k}. \quad (40)$$

Funkcja ta charakteryzuje się stałą elastycznością względem zmiennych objaśniających. Przypomnijmy, że elastycznością w punkcie x różniczkowalnej i niezerowej w tym punkcie funkcji jednej zmiennej f nazywamy granicę ilorazu względnego przyrostu funkcji i względnego przyrostu argumentu przy przyroście argumentu dążącym do 0:

$$\lim_{h \rightarrow 0} \frac{\frac{f(x+h) - f(x)}{f(x)}}{\frac{x+h-x}{x}} = f'(x) \cdot \frac{x}{f(x)}.$$

Uogólnienie tej definicji na funkcje wielu zmiennych jest natychmiastowe. W przypadku funkcji potęgowej (40) elastyczność względem i -tej zmiennej jest równa β_i . Interpretacja parametru β_i modelu (39) i (40) może więc być taka, że przyrost wartości x_i o 1 procent powoduje przyrost wartości y o β_i procent.

6.2. Zmienne zero-jedynkowe.

Zmienne zero-jedynkowe są zmiennymi przyjmującymi dokładnie dwie wartości: 0 i 1. Ograniczymy się do zmiennych zero-jedynkowych występujących w roli zmiennych objaśniających. Każda zmienna zero-jedynkowa pozwala wyróżnić dwa podzbiory obserwacji: podzbiór tych obserwacji, w których przyjmuje ona wartość 0 oraz tych, w których przyjmuje wartość 1. Odnotujemy trzy przykłady zastosowania zmiennych zero-jedynkowych. Wyjściowym modelem w trzech następujących podpunktach jest model

$$y_i = \alpha_0 + \alpha_1 x_i + \eta_i, \quad (41)$$

a dodatkowo zakładamy, że liczba obserwacji jest wystarczająco duża.

6.2.1. Dwie grupy obserwacji.

Załóżmy, że część obserwacji zmiennych modelu (41) dotyczy osób płci męskiej, a część – żeńskiej. Przypuszczając, że płeć – obok x – może mieć wpływ na wartości zmiennej objaśnianej, możemy uwzględnić to przypuszczenie wprowadzając do modelu (41) zmienną z określoną następująco:

$$z_i = \begin{cases} 1 & \text{gdy } i\text{-ta obserwacja dotyczy mężczyzny,} \\ 0 & \text{w przeciwnym przypadku.} \end{cases}$$

Rozszerzony model ma postać

$$y_i = \beta_0 + \beta_1 x_i + \gamma z_i + \xi_i,$$

a parametr γ informuje, o ile przeciętnie większa jest oczekiwana wartość zmiennej y w przypadku mężczyzn.

6.2.2. Kilka grup obserwacji.

Do wyróżnienia m rozłącznych podzbiorów zbioru obserwacji wystarcza $m - 1$ zmiennych zero-jedynkowych

$$z_{ij} = \begin{cases} 1 & \text{gdy } i\text{-ta obserwacja należy do } j\text{-tej grupy,} \\ 0 & \text{w przeciwnym przypadku,} \end{cases} \quad j = 1, \dots, m - 1.$$

Nienależenie i -tej obserwacji do żadnej z $m - 1$ grup, tzn. warunek $z_{i1} = \dots = z_{i,m-1} = 0$, oznacza należenie tej obserwacji do m -tej grupy. Po włączeniu do modelu (41) zmiennych z_{ij} dostajemy model

$$y_i = \beta_0 + \beta_1 x_i + \gamma_1 z_{i1} + \dots + \gamma_{m-1} z_{i,m-1} + \xi_i, \quad (42)$$

7. Prognozowanie na podstawie modelu regresji liniowej. 25

Parametr γ_j określa, o ile przeciętnie większa jest wartość oczekiwana zmiennej objaśnianej w j -tej grupie od jej wartości oczekiwanej w m -tej grupie.

6.2.3. Sezonowość.

Szczególnym przypadkiem modelu (42) jest model tendencji rozwojowej z uwzględnieniem wahań sezonowych. Załóżmy, że trend rozwoju wielkości y_i jest liniowy, tzn. w (42) jest $x_i = i$, a w tych samych kwartałach występują podobne odchylenia od trendu. Wybierając za punkt odniesienia pierwszy kwartał, możemy napisać następujący model:

$$y_i = \beta_0 + \beta_1 i + \gamma_2 z_{i2} + \gamma_3 z_{i3} + \gamma_4 z_{i4} + \xi_i, \quad (43)$$

gdzie

$$z_{ij} = \begin{cases} 1 & \text{gdy } i\text{-ta obserwacja dotyczy } j\text{-tego kwartału,} \\ 0 & \text{w przeciwnym przypadku,} \end{cases} \quad j = 2, 3, 4. \quad (44)$$

6.3. Przykład.

Kontynuujemy przykład z punktu 1.2. Rozpatrujemy następujący model zależności spożycia y_t od produktu krajowego brutto x_t z uwzględnieniem odchyłeń kwartalnych z_{tj} , zdefiniowanych przez (44):

$$y_t = \beta_0 + \beta_1 x_t + \gamma_2 z_{t2} + \gamma_3 z_{t3} + \gamma_4 z_{t4} + \xi_t. \quad (45)$$

Wyniki oszacowania parametrów tego modelu przedstawia tabela 3.

Wnioskujemy, że parametry γ_2 , γ_3 i γ_4 są istotnie różne od 0. Biorąc pod uwagę znak tych parametrów możemy stwierdzić, że w drugim, trzecim i czwartym kwartale wydatki na spożycie są istotnie niższe od wynikających z wielkości produktu krajowego brutto.

7. Prognozowanie na podstawie modelu regresji liniowej.

7.1. Definicja i własności prognoz.

Założmy, że w warunkach modelu regresji liniowej (11)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi}$$

Tabela 3. Wyniki oszacowania parametrów modelu (45).

Zmienna	Parametr	Błąd szacunku	Statystyka t
1	4,2104	3,1292	1,3455
x_t	0,8542	0,0206	41,5668
z_{t2}	-5,7174	2,2681	-2,5208
z_{t3}	-9,9035	2,2808	-4,3420
z_{t4}	-24,6150	2,3619	-10,4216
Liczba obserwacji:			32
Suma kwadratów reszt:			551,0050
Współczynnik determinacji:			0,9847
Standardowy błąd regresji:			4,5175
Statystyka istotności regresji:			432,8761
Statystyka d Durbina-Watsona:			2,0725

Źródło: Obliczenia własne na podstawie danych zawartych w tabeli 1.

prognozujemy wartość

$$y_* = \mathbf{X}_*^T \boldsymbol{\beta} + \xi_*, \quad (46)$$

gdzie $\mathbf{X}_* = [x_{i*}]_{k \times 1}$ jest wektorem wartości przyjmowanych przez zmienne objaśniające, a ξ_* jest składnikiem losowym o takim samym rozkładzie jak składniki losowe z próby (tzn. o rozkładzie normalnym o wartości oczekiwanej 0 i wariancji σ^2) i od nich niezależnym (tzn. spełniającym warunek $E\xi\xi_* = \mathbf{0}$). Załóżmy, że prognozy dokonujemy przy pomocy wartości

$$\hat{y}_* = \mathbf{X}_*^T \mathbf{b}. \quad (47)$$

Wówczas błąd prognozy jest równy

$$f = y_* - \hat{y}_* = \mathbf{X}_*^T (\boldsymbol{\beta} - \mathbf{b}) + \xi_*,$$

jego wartość oczekiwana wynosi

$$Ef = \mathbf{X}_*^T E(\boldsymbol{\beta} - \mathbf{b}) + E\xi_* = 0,$$

a wariancja jest równa

$$\sigma_f^2 = Ef^2 = \underbrace{\mathbf{X}_*^T E(\boldsymbol{\beta} - \mathbf{b})(\boldsymbol{\beta} - \mathbf{b})^T \mathbf{X}_*}_{\sigma_{\hat{y}_*}^2} + \underbrace{E\xi_*^2}_{\sigma^2} = \sigma_{\hat{y}_*}^2 + \sigma^2.$$

7. Prognozowanie na podstawie modelu regresji liniowej. 27

Wariancja błędu prognozy jest więc sumą dwóch wariancji: wariancji prognozy $\sigma_{\hat{y}_*}^2$ i wariancji składnika losowego σ^2 .

Łatwo pokazać, że nieobciążonym estymatorem wariancji błędu prognozy jest

$$s_f^2 = s^2 + s^2 \mathbf{X}_*^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_*. \quad (48)$$

Analogicznie jak w punkcie (1.4), zmienna losowa f/s_f ma rozkład t Studenta o $n - k$ stopniach swobody. Pozwala to na dokonywanie prognoz przedziałowych:

$$P \left(\left| \frac{y_* - \hat{y}_*}{s_f} \right| \leq t(\alpha, n - k) \right) = 1 - \alpha. \quad (49)$$

7.2. Miary dokładności prognoz.

Oznaczmy przez y_1, \dots, y_{n_*} i $\hat{y}_1, \dots, \hat{y}_{n_*}$ odpowiednio dokładne i prognozowane wartości zmiennej objaśnianej. Do głównych miar dokładności prognoz należą *średniokwadratowy błąd prognozy*

$$\frac{1}{n_*} \sum_{i=1}^{n_*} (y_i - \hat{y}_i)^2 \quad (50)$$

i *średni bezwzględny błąd prognozy*

$$\frac{1}{n_*} \sum_{i=1}^{n_*} |y_i - \hat{y}_i|. \quad (51)$$

Pierwiastek kwadratowy błędu średniokwadratowego i średni błąd bezwzględny są używane najczęściej jako miary opisowe ex post.

7.3. Przykład.

Poniżej przytaczamy dane GUS [13] o wielkości spożycia (y) i produktu krajowego brutto (x) w Polsce w pierwszych trzech kwartałach 2003 r. w cenach bieżących w mld zł, stanowiące kontynuację danych zawartych w tabeli 1.

Kwartał	y	x
I	166,8	184,5
II	168,4	197,6
III	166,9	198,9

7. Prognozowanie na podstawie modelu regresji liniowej. 28

Wykorzystamy dane o produkcie krajowym brutto do sporządzenia prognozy spożycia. Do prognozowania wykorzystamy model (45). Z tabeli 3 wynika, że

$$\mathbf{b}^T = \begin{bmatrix} 4,2104 & 0,8542 & -5,7174 & -9,9035 & -24,6150 \end{bmatrix},$$

a ponadto wiadomo, że oszacowaniem macierzy wariancji i kowariancji wektora \mathbf{b} jest

$$s^2(\mathbf{X}^T \mathbf{X})^{-1} = \begin{bmatrix} 9,7919 & -0,0553 & -1,9979 & -1,6986 & -0,6928 \\ -0,0553 & 0,0004 & -0,0042 & -0,0065 & -0,0142 \\ -1,9979 & -0,0042 & 5,1441 & 2,6160 & 2,6929 \\ -1,6986 & -0,0065 & 2,6160 & 5,2022 & 2,7697 \\ -0,6928 & -0,0142 & 2,6929 & 2,7697 & 5,5787 \end{bmatrix}.$$

Wektor wartości zmiennych objaśniających w pierwszym kwartale 2003 r. jest równy

$$\mathbf{X}_*^T = \begin{bmatrix} 1 & 184,5 & 0 & 0 & 0 \end{bmatrix}.$$

Prognoza obliczona według wzoru (47) jest równa 161,8178, oszacowaniem odchylenia standardowego błędu prognozy według wzoru (48) jest 4,9163, a 95-procentowym przedziałem ufności dla spożycia, obliczonym według wzoru (49), jest $151,73 \leq \hat{y}_* \leq 171,91$. Prognoza mówi więc, że przy produkcie krajowym brutto w pierwszym kwartale 2003 r. równym 184,5 mld zł wartość spożycia będzie wynosiła 161,8178 mld zł z błędem 4,9163 mld zł, a dokładniej, z prawdopodobieństwem 95% będzie znajdowała się w przedziale od 151,73 do 171,91 mld zł.

Prognozy dla II i III kwartału 2003 r. są odpowiednio równe 167,2909 i 164,2153 mld zł. Dla trzech prognoz kwartalnych na rok 2003 pierwiastek średniokwadratowego błędu prognozy jest równy 3,3297, a średni bezwzględny błąd prognozy – 2,9253. Oznacza to, że – przeciętnie rzecz biorąc – prognozy sporządzane według modelu (45) odchylają się od wartości rzeczywistych o około 3 mld zł.

8. Zadanie przygotowawcze do sprawdzianu.

Na podstawie danych o produkcji sprzedanej w mln zł (y_i), wartości środków trwałych w mln zł (x_{i1}) i liczbie pracujących w tys. osób (x_{i2}) w przemyśle w 16 województwach Polski w 2001 r. oszacowano metodą najmniejszych kwadratów parametry modelu

$$\ln y_i = \beta_0 + \beta_1 \ln x_{i1} + \beta_2 \ln x_{i2} + \xi_i, \quad i = 1, \dots, 16.$$

Wyniki oszacowań są następujące.

Parametr	Oszacowanie	Błąd szacunku	Statystyka t
β_0	2,7074	0,81575	3,3189
β_1	0,33733	0,15756	2,1409
β_2	0,79310	0,17881	4,4354
Liczba obserwacji:			16
Suma kwadratów reszt:			0,32163
Współczynnik determinacji:			0,96089
Standardowy błąd regresji:			0,15729
Statystyka istotności regresji:			159,7100
Statystyka d Durбина-Watsona:			2,2231

Jest podany fragment tablicy wartości krytycznych $t(\alpha, n)$ rozkładu t Studenta.

n	α				
	0,20	0,10	0,05	0,02	0,01
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921

W każdym z poniższych punktów zaznaczyć dokładnie jedną odpowiedź najbliższą rzeczywistości (w zadaniach przygotowawczych prawidłowe są wszystkie pierwsze odpowiedzi).

8. Zadanie przygotowawcze do sprawdzianu.

30

1. Parametr β_1 oszacowano na poziomie 0,34 z błędem
 - (a) 0,16,
 - (b) 0,16%,
 - (c) 2,14,
 - (d) 0,32.
2. Przy ustalonej wartości środków trwałych, wzrost liczby pracujących o 1% powoduje wzrost produkcji w przybliżeniu
 - (a) o 0,79%,
 - (b) o 0,46%,
 - (c) o 1,91%,
 - (d) o 1,13%.
3. Wartości teoretyczne logarytmu produkcji sprzedanej różnią się od wartości rzeczywistych przeciętnie
 - (a) o 0,16,
 - (b) o 0,16%,
 - (c) o 0,96%,
 - (d) o 0,32%.
4. Przy poziomie istotności 0,05, parametr
 - (a) β_1 jest nieistotnie różny od 0,
 - (b) β_1 jest istotnie różny od 0,
 - (c) β_2 jest nieistotnie różny od 0,
 - (d) β_0 jest nieistotnie różny od 0.
5. Model objaśnia kształtowanie się produkcji sprzedanej
 - (a) w 96%,
 - (b) w 4%,
 - (c) w 32%,
 - (d) w 68%.
6. Z prawdopodobieństwem 0,95 elastyczność produkcji względem zatrudnienia zawiera się w przedziale

8. Zadanie przygotowawcze do sprawdzianu.

31

- (a) od 0,41 do 1,18,
 - (b) od $-0,41$ do 1,18,
 - (c) od 0,48 do 1,11,
 - (d) od 0,27 do 1,32.
7. Dla logarytmu wartości środków trwałych równego 11,34 i logarytmu zatrudnienia równego 5,91, punktowa prognoza logarytmu produkcji wynosi
- (a) 11,22,
 - (b) 8,52,
 - (c) 17,25,
 - (d) 5,43.

A. Tablice statystyczne.

Układ tablic jest wzorowany na tablicach Zielińskiego [18]. Tablice ułożono wykorzystując m. in. algorytmy opublikowane w [9], [1] i [17].

Zmienna losowa ma rozkład χ^2 o $n \geq 1$ stopniach swobody, jeśli jest sumą kwadratów n niezależnych zmiennych losowych o tym samym rozkładzie normalnym o wartości oczekiwanej 0 i wariancji 1. Gęstością rozkładu χ^2 o n stopniach swobody jest

$$f(x; n) = \frac{x^{\frac{1}{2}n-1} e^{-\frac{1}{2}x}}{2^{\frac{1}{2}n} \Gamma(\frac{1}{2}n)}.$$

Zmienna losowa ma rozkład t Studenta o $n \geq 1$ stopniach swobody, jeśli jest ilorazem

$$\frac{X}{\sqrt{Y/n}},$$

gdzie zmienna X ma rozkład normalny o wartości oczekiwanej 0 i wariancji 1, Y ma rozkład χ^2 o n stopniach swobody i zmienne te są niezależne. Gęstością rozkładu t Studenta o n stopniach swobody jest

$$f(x; n) = \frac{1}{\sqrt{\pi n}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}.$$

Zmienna losowa ma rozkład F o $m \geq 1$ stopniach swobody licznika i $n \geq 1$ stopniach swobody mianownika jeśli jest ilorazem

$$\frac{X/m}{Y/n},$$

gdzie X i Y są niezależnymi zmiennymi losowymi o rozkładach χ^2 o odpowiednio m i n stopniach swobody. Gęstością rozkładu tej zmiennej jest

$$f(x; m, n) = \frac{m^{m/2} n^{n/2} \Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2}) \Gamma(\frac{n}{2})} x^{\frac{m}{2}-1} (mx+n)^{-\frac{m+n}{2}}, \quad x > 0, m \geq 1, n \geq 1.$$

Tablica I. Wartości krytyczne $\chi^2(\alpha, n)$ rozkładu χ^2 .

n	α				
	0,995	0,990	0,975	0,950	0,900
1	0,000	0,000	0,001	0,004	0,016
2	0,010	0,020	0,051	0,103	0,211
3	0,072	0,115	0,216	0,352	0,584
4	0,207	0,297	0,484	0,711	1,064
5	0,412	0,554	0,831	1,145	1,610
6	0,676	0,872	1,237	1,635	2,204
7	0,989	1,239	1,690	2,167	2,833
8	1,344	1,646	2,180	2,733	3,490
9	1,735	2,088	2,700	3,325	4,168
10	2,156	2,558	3,247	3,940	4,865
11	2,603	3,053	3,816	4,575	5,578
12	3,074	3,571	4,404	5,226	6,304
13	3,565	4,107	5,009	5,892	7,042
14	4,075	4,660	5,629	6,571	7,790
15	4,601	5,229	6,262	7,261	8,547
16	5,142	5,812	6,908	7,962	9,312
17	5,697	6,408	7,564	8,672	10,085
18	6,265	7,015	8,231	9,390	10,865
19	6,844	7,633	8,907	10,117	11,651
20	7,434	8,260	9,591	10,851	12,443
21	8,034	8,897	10,283	11,591	13,240
22	8,643	9,542	10,982	12,338	14,041
23	9,260	10,196	11,689	13,091	14,848
24	9,886	10,856	12,401	13,848	15,659
25	10,520	11,524	13,120	14,611	16,473
26	11,160	12,198	13,844	15,379	17,292
27	11,808	12,879	14,573	16,151	18,114
28	12,461	13,565	15,308	16,928	18,939
29	13,121	14,256	16,047	17,708	19,768
30	13,787	14,953	16,791	18,493	20,599
35	17,192	18,509	20,569	22,465	24,797
40	20,707	22,164	24,433	26,509	29,051
50	27,991	29,707	32,357	34,764	37,689
60	35,534	37,485	40,482	43,188	46,459
80	51,172	53,540	57,153	60,391	64,278
100	67,328	70,065	74,222	77,929	82,358

Tablica podaje liczby $\chi^2(\alpha, n)$ takie, że $P(\chi_n^2 > \chi^2(\alpha, n)) = \alpha$, gdzie χ_n^2 jest zmienną losową o rozkładzie χ^2 o n stopniach swobody.

Tablica I (cd.). Wartości krytyczne $\chi^2(\alpha, n)$ rozkładu χ^2 .

n	α				
	0,100	0,050	0,025	0,010	0,005
1	2,706	3,841	5,024	6,635	7,879
2	4,605	5,991	7,378	9,210	10,597
3	6,251	7,815	9,348	11,345	12,838
4	7,779	9,488	11,143	13,277	14,860
5	9,236	11,071	12,833	15,086	16,750
6	10,645	12,592	14,449	16,812	18,548
7	12,017	14,067	16,013	18,475	20,278
8	13,362	15,507	17,535	20,090	21,955
9	14,684	16,919	19,023	21,666	23,590
10	15,987	18,307	20,483	23,209	25,188
11	17,275	19,675	21,920	24,725	26,757
12	18,549	21,026	23,337	26,217	28,300
13	19,812	22,362	24,736	27,688	29,820
14	21,064	23,685	26,119	29,141	31,320
15	22,307	24,996	27,488	30,578	32,802
16	23,542	26,296	28,845	32,000	34,267
17	24,769	27,587	30,191	33,409	35,719
18	25,989	28,869	31,526	34,805	37,157
19	27,204	30,144	32,852	36,191	38,583
20	28,412	31,410	34,170	37,566	39,997
21	29,615	32,671	35,479	38,932	41,402
22	30,813	33,924	36,781	40,290	42,796
23	32,007	35,172	38,076	41,639	44,182
24	33,196	36,415	39,364	42,980	45,559
25	34,382	37,653	40,647	44,314	46,928
26	35,563	38,885	41,923	45,642	48,290
27	36,741	40,113	43,195	46,963	49,645
28	37,916	41,337	44,461	48,278	50,994
29	39,087	42,557	45,722	49,588	52,336
30	40,256	43,773	46,979	50,892	53,672
35	46,059	49,802	53,203	57,342	60,275
40	51,805	55,759	59,342	63,691	66,767
50	63,167	67,505	71,420	76,154	79,491
60	74,397	79,082	83,298	88,380	91,953
80	96,578	101,880	106,629	112,329	116,322
100	118,498	124,342	129,561	135,807	140,171

Tablica podaje liczby $\chi^2(\alpha, n)$ takie, że $P(\chi_n^2 > \chi^2(\alpha, n)) = \alpha$, gdzie χ_n^2 jest zmienną losową o rozkładzie χ^2 o n stopniach swobody.

Tablica II. Wartości krytyczne $t(\alpha, n)$ rozkładu t Studenta.

n	α				
	0,20	0,10	0,05	0,02	0,01
1	3,078	6,314	12,706	31,821	63,657
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750
40	1,303	1,684	2,021	2,423	2,704
50	1,299	1,676	2,009	2,403	2,678
60	1,296	1,671	2,000	2,390	2,660
70	1,294	1,667	1,994	2,381	2,648
80	1,292	1,664	1,990	2,374	2,639
90	1,291	1,662	1,987	2,368	2,632
100	1,290	1,660	1,984	2,364	2,626
∞	1,282	1,645	1,960	2,326	2,576

Tablica podaje liczby $t(\alpha, n)$ takie, że $P(|t_n| > t(\alpha, n)) = \alpha$, gdzie t_n jest zmienną losową o rozkładzie t Studenta o n stopniach swobody.

Tablica III. Wartości krytyczne $F(\alpha, m, n)$ rozkładu F .

n	m							
	1	2	3	4	5	6	7	8
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9
	4052	5000	5403	5625	5764	5859	5928	5981
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37
	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37
3	10,13	9,552	9,277	9,117	9,013	8,941	8,887	8,845
	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041
	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80
5	6,608	5,786	5,409	5,192	5,050	4,950	4,876	4,818
	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147
	13,75	10,92	9,780	9,148	8,746	8,466	8,260	8,102
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726
	12,25	9,547	8,451	7,847	7,460	7,191	6,993	6,840
8	5,318	4,459	4,066	3,838	3,687	3,581	3,500	3,438
	11,26	8,649	7,591	7,006	6,632	6,371	6,178	6,029
9	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,230
	10,56	8,022	6,992	6,422	6,057	5,802	5,613	5,467
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072
	10,04	7,559	6,552	5,994	5,636	5,386	5,200	5,057
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913	2,849
	9,330	6,927	5,953	5,412	5,064	4,821	4,640	4,499
15	4,543	3,682	3,287	3,056	2,901	2,790	2,707	2,641
	8,683	6,359	5,417	4,893	4,556	4,318	4,142	4,004
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514	2,447
	8,096	5,849	4,938	4,431	4,103	3,871	3,699	3,564
24	4,260	3,403	3,009	2,776	2,621	2,508	2,423	2,355
	7,823	5,614	4,718	4,218	3,895	3,667	3,496	3,363
30	4,171	3,316	2,922	2,690	2,534	2,421	2,334	2,266
	7,562	5,390	4,510	4,018	3,699	3,473	3,304	3,173
40	4,085	3,232	2,839	2,606	2,449	2,336	2,249	2,180
	7,314	5,179	4,313	3,828	3,514	3,291	3,124	2,993
60	4,001	3,150	2,758	2,525	2,368	2,254	2,167	2,097
	7,077	4,977	4,126	3,649	3,339	3,119	2,953	2,823
120	3,920	3,072	2,680	2,447	2,290	2,175	2,087	2,016
	6,851	4,787	3,949	3,480	3,174	2,956	2,792	2,663
∞	3,863	2,996	2,606	2,372	2,214	2,099	2,010	1,938
	6,622	4,605	3,782	3,319	3,017	2,802	2,639	2,511

Tablica podaje liczby $F(\alpha, m, n)$ takie, że $P(F > F(\alpha, m, n)) = \alpha$, gdzie F jest zmienną losową o rozkładzie F o m stopniach swobody licznika i n stopniach swobody mianownika. Wartości krytyczne dla $\alpha = 0,05$ podano u góry, a dla $\alpha = 0,01$ – u dołu wiersza.

Tablica III (cd.). Wartości krytyczne $F(\alpha, m, n)$ rozkładu F .

n	m							
	9	10	12	15	20	24	30	40
1	240,5	241,9	243,9	245,9	248,0	249,1	250,1	251,1
	6022	6056	6106	6157	6209	6235	6261	6287
2	19,38	19,40	19,41	19,43	19,45	19,45	19,46	19,47
	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,47
3	8,812	8,786	8,745	8,703	8,660	8,639	8,617	8,594
	27,35	27,23	27,05	26,87	26,69	26,60	26,50	26,41
4	5,999	5,964	5,912	5,858	5,803	5,774	5,746	5,717
	14,66	14,55	14,37	14,20	14,02	13,93	13,84	13,75
5	4,772	4,735	4,678	4,619	4,558	4,527	4,496	4,464
	10,16	10,05	9,888	9,722	9,553	9,466	9,379	9,291
6	4,099	4,060	4,000	3,938	3,874	3,841	3,808	3,774
	7,976	7,874	7,718	7,559	7,396	7,313	7,229	7,143
7	3,677	3,637	3,575	3,511	3,445	3,410	3,376	3,340
	6,719	6,620	6,469	6,314	6,155	6,074	5,992	5,908
8	3,388	3,347	3,284	3,218	3,150	3,115	3,079	3,043
	5,911	5,814	5,667	5,515	5,359	5,279	5,198	5,116
9	3,179	3,137	3,073	3,006	2,936	2,900	2,864	2,826
	5,351	5,257	5,111	4,962	4,808	4,729	4,649	4,567
10	3,020	2,978	2,913	2,845	2,774	2,737	2,700	2,661
	4,942	4,849	4,706	4,558	4,405	4,327	4,247	4,165
12	2,796	2,753	2,687	2,617	2,544	2,505	2,466	2,426
	4,388	4,296	4,155	4,010	3,858	3,780	3,701	3,619
15	2,588	2,544	2,475	2,403	2,328	2,288	2,247	2,204
	3,895	3,805	3,666	3,522	3,372	3,294	3,214	3,132
20	2,393	2,348	2,278	2,203	2,124	2,082	2,039	1,994
	3,457	3,368	3,231	3,088	2,938	2,859	2,778	2,695
24	2,300	2,255	2,183	2,108	2,027	1,984	1,939	1,892
	3,256	3,168	3,032	2,889	2,738	2,659	2,577	2,492
30	2,211	2,165	2,092	2,015	1,932	1,887	1,841	1,792
	3,067	2,979	2,843	2,700	2,549	2,469	2,386	2,299
40	2,124	2,077	2,003	1,924	1,839	1,793	1,744	1,693
	2,888	2,801	2,665	2,522	2,369	2,288	2,203	2,114
60	2,040	1,993	1,917	1,836	1,748	1,700	1,649	1,594
	2,718	2,632	2,496	2,352	2,198	2,115	2,028	1,936
120	1,959	1,910	1,834	1,750	1,659	1,608	1,554	1,495
	2,559	2,472	2,336	2,192	2,035	1,950	1,860	1,763
∞	1,880	1,831	1,752	1,666	1,571	1,517	1,459	1,394
	2,407	2,321	2,185	2,039	1,878	1,791	1,696	1,592

Tablica podaje liczby $F(\alpha, m, n)$ takie, że $P(F > F(\alpha, m, n)) = \alpha$, gdzie F jest zmienną losową o rozkładzie F o m stopniach swobody licznika i n stopniach swobody mianownika. Wartości krytyczne dla $\alpha = 0,05$ podano u góry, a dla $\alpha = 0,01$ – u dołu wiersza.

Tablica III (cd.). Wartości krytyczne $F(\alpha, m, n)$ rozkładu F .

n	m		
	60	120	∞
1	252,2	253,3	254,2
	6313	6339	6283
2	19,48	19,49	19,50
	99,48	99,49	99,48
3	8,572	8,549	8,527
	26,32	26,22	26,13
4	5,688	5,658	5,628
	13,65	13,56	13,46
5	4,431	4,398	4,365
	9,202	9,112	9,020
6	3,740	3,705	3,669
	7,057	6,969	6,880
7	3,304	3,267	3,230
	5,824	5,737	5,650
8	3,005	2,967	2,928
	5,032	4,946	4,859
9	2,787	2,748	2,707
	4,483	4,398	4,311
10	2,621	2,580	2,538
	4,082	3,996	3,909
12	2,384	2,341	2,296
	3,535	3,449	3,361
15	2,160	2,114	2,066
	3,047	2,959	2,868
20	1,946	1,896	1,843
	2,608	2,517	2,421
24	1,842	1,790	1,733
	2,403	2,310	2,211
30	1,740	1,683	1,622
	2,208	2,111	2,006
40	1,637	1,577	1,509
	2,019	1,917	1,805
60	1,534	1,467	1,389
	1,836	1,726	1,601
120	1,429	1,352	1,254
	1,656	1,533	1,381
∞	1,318	1,221	1,001
	1,473	1,325	1,001

Tablica podaje liczby $F(\alpha, m, n)$ takie, że $P(F > F(\alpha, m, n)) = \alpha$, gdzie F jest zmienną losową o rozkładzie F o m stopniach swobody licznika i n stopniach swobody mianownika. Wartości krytyczne dla $\alpha = 0,05$ podano u góry, a dla $\alpha = 0,01$ – u dołu wiersza.

Literatura

- [1] A. Bartkowiak, *Podstawowe algorytmy statystyki matematycznej*, PWN, Warszawa 1979.
- [2] D. Begg, S. Fischer, R. Dornbusch, *Ekonomia*, t. I, PWE, Warszawa 1993.
- [3] *Bilans produktu krajowego brutto według kwartałów w latach 1995-2000*, GUS, http://www.stat.gov.pl/serwis/bilans_pkb.htm, 23.IX.2003.
- [4] G. C. Chow, *Ekonometria*, Wydawnictwo Naukowe PWN, Warszawa 1995.
- [5] J. Durbin, G. S. Watson, *Testing for serial correlation in least squares regression. I*, *Biometrika* 37/1950, str. 409 – 428.
- [6] J. Durbin, G. S. Watson, *Testing for serial correlation in least squares regression. II*, *Biometrika* 38/1951, str. 159 – 178.
- [7] A. S. Goldberger, *Teoria ekonometrii*, PWE, Warszawa 1972.
- [8] S. M. Goldfeld, R. E. Quandt, *Some tests for homoscedasticity*, *American Statistical Association Journal* 60/1965, str. 539 – 547.
- [9] P. Griffiths, I. D. Hill (red.), *Applied Statistics Algorithms*, Royal Statistical Society, Chichester 1985.
- [10] J. M. Keynes, *Ogólna teoria zatrudnienia, procentu i pieniądza*, Wydawnictwo Naukowe PWN, Warszawa 2003.
- [11] J. M. Keynes, *The General Theory of Employment, Interest and Money*, MacMillan and Co., Limited, Londyn 1946.
- [12] *Produkt krajowy brutto w IV kwartale 2002 r.*, GUS, http://www.stat.gov.pl/serwis/monit_kwart/pkb/IV2002.htm, 24.IX.2003.
- [13] *Produkt krajowy brutto w III kwartale 2003 roku, skorygowane szacunki PKB za 2002 rok według kwartałów oraz za I i II kwartał 2003 r. (tablice)*, GUS, http://www.stat.gov.pl/serwis/monit_kwart/pkb/III2003.htm, 8.I.2004.
- [14] N. E. Savin, K. J. White, *The Durbin-Watson test for serial correlation with extreme sample sizes or many regressors*, *Econometrica* 45/1977, str. 1989 – 1996.

- [15] K. Strzała, T. Przechlewski, *Ekonometria inaczej*, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk 1994.
- [16] A. Welfe, *Ekonometria*, PWE, Warszawa 1995.
- [17] M. J. Wichura, *Algorithm AS 241. The percentage points of the normal distribution*, Applied Statistics 37/1988, str. 477 – 484.
- [18] R. Zieliński, W. Zieliński, *Podręczne tablice statystyczne*, WNT, Warszawa 1987.
- [19] R. Zieliński, W. Zieliński, *Tablice statystyczne*, wydanie II zmienione, PWN, Warszawa 1990.