Barbara Gładysz
Jacek Mercik

Modelowanie ekonometryczne

Studium przypadku

Wydanie II

Oficyna Wydawnicza Politechniki Wrocławskiej
Wrocław 2007

Recenzent
Paweł DITTMANN

Opracowanie redakcyjne i korekta
Alina KACZAK

Projekt okładki
Justyna GODLEWSKA-ISKIERKA

OFICYNA WYDAWNICZA POLITECHNIKI WROCŁAWSKIEJ
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław
http://www.pwr.wroc.pl/~oficwyd
e-mail: oficwyd@pwr.wroc.pl

ISBN 978-83-7493-354-4

Drukarnia Oficyny Wydawniczej Politechniki Wrocławskiej. Zam. nr 765/2007.

SPIS RZECZY

WSTĘP

Modele ekonometryczne to modele opisujące wzajemne zależności między badanymi cechami, które umożliwiają lepsze zrozumienie mechanizmów rządzących analizowanym fragmentem rzeczywistości, a także przewidywanie zachowania modelowanych procesów. Ekonometria jest stosowana dziś w wielu dziedzinach, takich jak ekonomia, medycyna, meteorologia, finanse czy technika. Rozwój informatyki umożliwia analizowanie nawet bardzo złożonych wycinków rzeczywistości. W książce zaprezentowano procesy modelowania ekonometrycznego wybranych fragmentów rzeczywistości.

Modelowanie ekonometryczne wymaga od ekonometryka uwzględnienia specyfiki analizowanego problemu. Dobór odpowiedniej postaci analitycznej modelu ekonometrycznego, właściwych testów statystycznych to klucz do sukcesu. Chcemy pokazać jak można budować modele różnych zjawisk, starając się, niejako przy okazji, pokazać cały rygoryzm statystyczny z tym związany.

W rozdziale pierwszym opisano podstawowe etapy modelowania ekonometrycznego. Przedstawiono klasyczną metodę najmniejszych kwadratów służącą do estymacji współczynników równania regresji. Podano warunki Gaussa–Markowa oraz wymieniono testy statystyczne stosowane do ich weryfikacji. Przedstawiono także metody predykcji ekonometrycznej (punktowej i przedziałowej).

W rozdziale drugim podano testy statystyczne stosowane w procesie weryfikacji modeli ekonometrycznych. Omówiono testy istotności współczynników regresji oraz testy badania własności składników losowych modeli (normalność, losowość, symetria, autokorelacja, homoskedastyczność). Zaprezentowane testy uwzględniają przypadki modeli liniowych i nieliniowych, danych chronologicznych i przekrojowych, modeli ze zmiennymi opóźnionymi, wielkość próby statystycznej.

Etapy budowania i weryfikacji modeli ekonometrycznych opisujących wybrane fragmenty rzeczywistości omówiono w rozdziale trzecim. W celu zaprezentowania czytelnikowi szerokich możliwości stosowania ekonometrii starano się dobrać modele z różnych klas i z różnych dziedzin. Przykłady modeli tak dobrano, aby zaprezentować różne warianty postępowania przy konstrukcji modeli ekonometrycznych:

- Model opisujący zależność czasu podróży samochodem od długości trasy – model liniowy z jedną zmienną objaśniającą.
- Cena mieszkań jako funkcja powierzchni – model nieliniowy (krzywa Tőrquista) z jedną zmienną objaśniającą.
- Wzrost dzieci jako funkcja wieku i płci – model liniowy z dwiema zmiennymi objaśniającymi (ilościową i jakościową).
- Podaż pieniądza w Polsce – model autoregresyjny.

• Stopa bezrobocia – model nieliniowy, autoregresyjny, okresowy ze zmienną opóźnioną w czasie i funkcją harmoniczną.

• Średnia temperatura we Wrocławiu – wielomian w okresie styczeń–sierpień i funkcja liniowa dla miesięcy wrzesień–grudzień.

Są to więc modele liniowe i nieliniowe, jedno- i wielorównaniowe, z jedną i wieloma zmiennymi, ze zmiennymi ilościowymi i jakościowymi oraz ze zmiennymi opóźnionymi w czasie. Analizowane modele różnią się ponadto strukturą danych. Zaprezentowano modele o danych przekrojowych oraz modele skonstruowane na podstawie szeregów czasowych.

Każdy model poddano weryfikacji statystycznej. Szczególny nacisk położono na zaprezentowanie, w jaki sposób w procesie modelowania wykorzystać niepomyślny dla weryfikowanego modelu ekonometrycznego wynik testu statystycznego. Występowanie autokorelacji implikuje często konieczność uwzględnienia w modelu zmiennych opóźnionych w czasie. Brak losowości lub symetrii reszt może wynikać z cykliczności badanej zmiennej lub nieliniowej zależności między zmienną objaśnianą a zmiennymi objaśniającymi. Heteroskedastyczność może być skutkiem nieliniowej zależności zmiennych lub niewłaściwie dobranej postaci analitycznej modelu. Brak istotności stałej modelu świadczy o braku liniowej zależności zmiennej objaśnianej od zmiennych objaśniających lub występowania współzależności liniowej zmiennych objaśniających. Brak koincydencji często świadczy o współliniowości zmiennych objaśniających.

Modele, które przeszły pozytywnie przez wszystkie etapy weryfikacji statystycznej zastosowano do budowy prognoz.

W rozdziale czwartym przedstawiono próbę konstrukcji modelu frekwencji w wyborach prezydenta RP. Jest to zarazem przykład modelowania w dziedzinie nauk społecznych, które się nie powiodło. Wynika z tego, że nie zawsze proces konstrukcji modelu ekonometrycznego kończy się sukcesem. Przyczyną klęski może być np: losowość badanej cechy i brak jej zależności od innych czynników, nieumiejętność dobrania postaci modelu ekonometrycznego lub zmiennych objaśniających. Co więcej, ekonometryk w swojej pracy spotyka się z przypadkami modeli pozytywnie zweryfikowanych statystycznie, które okazują się nieefektywne w praktyce.

Zaprezentowano możliwości zastosowania w modelowaniu ekonometrycznym arkusza kalkulacyjnego *Excel* (rozdz. 4). Chcieliśmy pokazać Czytelnikowi, że z wieloma problemami w modelowaniu ekonometrycznym można się zmagać, będąc wspomaganym przez tak popularny arkusz kalkulacyjny jakim jest *Excel*.

Książka jest przeznaczona dla studentów różnych kierunków studiów ekonomicznych, ale także może służyć pomocą osobom zajmującym się modelowaniem ekonometrycznym w praktyce zawodowej. Stanowi uzupełnienie bogatej literatury z zakresu teorii ekonometrii oraz zbiorów zadań ekonometrycznych. Do pełnego zrozumienia prezentowanych w książce zagadnień konieczna jest wiedza statystyczna. Założyliśmy, że odpowiada ona standardowemu kursowi statystyki i ekonometrii, który koń-

czą studenci Wydziału Informatyki i Zarządzania Politechniki Wrocławskiej. Studentom, z którymi wspólnie zmagaliśmy się przy konstrukcji różnorakich modeli ekonometrycznych tą drogą składamy podziękowanie, wierząc, że i oni w swojej pracy zawodowej sięgną w przyszłości po tę książkę.

Autorzy

ROZDZIAŁ 1

OGÓLNY SCHEMAT MODELOWANIA I PROGNOZOWANIA EKONOMETRYCZNEGO

W pewnym uproszczeniu modelowanie ekonometryczne może być rozumiane jako ciąg kolejno następujących po sobie procedur, których wykonanie prowadzi do wyniku, jakim jest model ekonometryczny. W praktyce modelowania zdarza się często, że wiele z tych procedur trzeba powtórzyć wielokrotnie. Jeżeli bowiem skonstruowany model nie przejdzie pomyślnie weryfikacji statystycznej, to może się okazać, że badane zjawisko lepiej opisuje inna funkcja lub inny układ zmiennych objaśniających. Wymusza, to ponowną konstrukcję modelu i jego weryfikację. W dalszej części przedstawiono podstawową sekwencję procedur modelowania ekonometrycznego. Podano też metody konstrukcji prognoz ekonometrycznych.

1.1. Krok I. Określenie celu badań modelowych

Określenie celu badań modelowych wymaga sprecyzowania dziedziny i rodzaju badań, a więc np.: zdefiniowania czy naszym celem jest poznanie kształtowania się badanego zjawiska w czasie, czy też określenie charakteru i rodzaju zależności przyczynowo-skutkowych. W początkowym etapie modelowania ekonometrycznego musimy starać się odpowiedzieć na pytania, jakie są nasze rzeczywiste potrzeby, czego oczekujemy po modelowaniu i do czego będziemy używać skonstruowane modele? Od tego zależy, czy zbudowany model uznamy za istotnie poprawny i czy wnioski, jakie na jego podstawie będziemy wyciągać będą mogły być zaakceptowane. Zdarza się często, że modelujący, zadowolony z poprawności formalnej modelu ekonometrycznego, zapomina o celu jego budowy i formułuje wnioski, które w żadnym razie nie powinny być z niego wyprowadzone.

Chcemy zaznaczyć, że jest to jeden z ważniejszych etapów modelowania, który wymaga od modelującego znacznej wiedzy o badanym zjawisku. Nie można się tutaj ograniczyć wyłącznie do podejścia czysto formalnego, które często sprowadza się do

analizy zbioru danych bez jego zrozumienia. Takie formalne podejście nie pozwala zrozumieć istoty badanych zależności, a więc w konsekwencji może prowadzić do budowy fałszywych modeli lub wyciągania fałszywych wniosków. Z naszej praktyki związanej z modelowaniem ekonometrycznym wynika, że pierwsze trzy kroki (w tym określenie celu badań modelowych) zajmują ok. 80–90% czasu poświęconego na zbudowanie poprawnego modelu ekonometrycznego.

1.2. Krok II. Specyfikacja zmiennych wraz z gromadzeniem danych

Specyfikacja zmiennych wraz zgromadzeniem danych obejmuje:

- Zebranie informacji o wartościach zmiennych objaśnianych i objaśniających.
- Graficzną analizę kształtowania się poszczególnych zmiennych oraz zależności zmiennych objaśnianych od zmiennych objaśniających.
- Eliminację zmiennych objaśniających o małym współczynniku zmienności.
- Eliminację liniowo zależnych zmiennych objaśniających.
- Dobór zmiennych objaśniających do modelu ekonometrycznego (techniki doboru zmiennych – metoda pojemności informacji, metoda grafowa, procedura eliminacji *a posteriori*, procedura selekcji *a priori*, procedury regresji krokowej).

1.3. Krok III. Wybór klasy modelu

Wybór klasy modelu ekonometrycznego wymaga:

- Zdefiniowania postaci analitycznej modelu (liniowa, nieliniowa),
- Określenia liczby funkcji w modelu (modele jedno lub wielorównaniowe),
- Ustalenia liczby i rodzaju zmiennych objaśniających (modele z jedną lub wieloma zmiennymi objaśniającymi; zmienne ilościowe i jakościowe),
- Wyznaczenia roli czynnika czasu w modelowaniu (modele statyczne, dynamiczne).

1.4. Krok IV. Estymacja parametrów strukturalnych

Parametry modelu liniowego[1]

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_k x_k + \varepsilon_t$$

[1] Jeżeli przyjęta funkcja jest nieliniowa, należy transformować ją do postaci liniowej.

szacujemy klasyczną metodą najmniejszych kwadratów (KMNK), otrzymując równanie liniowe

$$\hat{y} = a_0 + a_1 x_1 + a_2 x_2 + ... + a_k x_k ,$$

w którym współczynniki a_j są estymatorami nieznanych parametrów α_j (j = 0, 1, 2, ..., k) podanej funkcji.

W metodzie najmniejszych kwadratów współczynniki a_j dobiera się tak, aby suma kwadratów odchyleń estymowanych wartości zmiennej objaśnianej $\hat{y}$ od jej rzeczywistych wartości y była minimalna

$$\sum_{i=1}^{n} e_t^2 = \sum_{i=1}^{n} \left(y_t - \hat{y}_t\right)^2 \to \min$$

Funkcja przyjmuje minimum w punkcie

$$\mathbf{a} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y} ,$$

gdzie

$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & ... & x_{1k} \\ 1 & x_{21} & x_{22} & ... & x_{2k} \\ ... & ... & ... & ... & ... \\ 1 & x_{n1} & x_{n2} & ... & x_{nk} \end{bmatrix}$ – macierz obserwacji zmiennych objaśniających,

$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_k \end{bmatrix}$ – wektor obserwacji zmiennej objaśnianej,

$\mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \\ \dots \\ a_k \end{bmatrix}$ – wektor estymatorów współczynników równania regresji.

Za estymator wariancji składnika losowego ε równania regresji przyjmujemy

$$S_\varepsilon^2 = \frac{\sum_{t=1}^{n} \left(y_t - \hat{y}_t\right)^2}{n - k - 1} ,$$

a za estymatory wariancji i kowariancji współczynników regresji elementy leżące odpowiednio na i poza główną przekątną macierzy

$$S^2(a) = \begin{bmatrix} d_{00} & d_{01} & \dots & d_{0k} \\ d_{10} & d_{11} & \dots & d_{1k} \\ \dots & \dots & \dots & \dots \\ d_{k0} & d_{k1} & \dots & d_{kk} \end{bmatrix} = S_\varepsilon^2 \left(\mathbf{X}^T \mathbf{X}\right)^{-1}.$$

1.5. Krok V. Weryfikacja modelu

Aby otrzymane metodą najmniejszych kwadratów estymatory a_j współczynników α_j (j = 0, 1, 2,..., k) były efektywne, muszą być spełnione założenia Gaussa–Markowa, a mianowicie:

- Związek między zmienną objaśnianą y a zmiennymi objaśniającymi x_1, x_2, ..., x_k ma charakter liniowy.
- Wartości zmiennych objaśniających są ustalone (nie są losowe) – losowość wartości zmiennej objaśnianej y wynika z losowości składnika ε.
- Składniki losowe ε dla poszczególnych wartości zmiennych objaśniających mają rozkład normalny (lub bardzo silnie zbliżony do normalnego) o wartości oczekiwanej zero i stałej wariancji: $N(0, \delta_\varepsilon)$.
- Składniki losowe nie są ze sobą skorelowane.

Spełnienie założeń Gaussa–Markowa weryfikuje się za pomocą odpowiednich testów statystycznych.

Liniowy charakter zależności między zmienną objaśnianą y a zmiennymi objaśniającymi x_1, x_2, ..., x_k weryfikujemy na podstawie wartości takich statystyk, jak współczynnik determinacji lub współczynnik zbieżności modelu.

Do weryfikacji losowości rozkładu reszt modelu względem równania regresji $\hat{y}$ można zastosować między innymi testy serii (test liczby serii, test maksymalnej długości serii).

Zaprezentowane w pracy testy weryfikacji normalności rozkładu składnika losowego to: testy zgodności χ^2, λ Kołmogorowa, Shapiro–Wilka, Dawida–Hellwiga.

Równość wariancji składnika losowego można weryfikować między innymi za pomocą testów: Goldfelda–Quandta, korelacji rangowej Spearmana oraz korelacji modułów składników losowych i czasu.

Zjawisko autokorelacji pierwszego rzędu składników losowych można weryfikować między innymi za pomocą testów Durbina–Watsona, von Neumanna, Durbina, a występowanie autokorelacji dowolnego rzędu testem istotności współczynników autokorelacji.

1.6. Krok VI. Wnioskowanie na podstawie modelu

Skonstruowany model może być stosowany między innymi do budowy prognoz. Wyróżnia się trzy rodzaje prognoz (predykcji ekonometrycznych).

Prognoza punktowa. Jest to prognoza warunkowej wartości oczekiwanej zmiennej objaśnianej y dla ustalonych wartości zmiennych objaśniających $x_0 = (x_{01}, x_{02}, ..., x_{0k})$ na podstawie zbudowanego równania regresji

$$\hat{y}_0 = a_0 + a_1 x_{01} + a_2 x_{02} + ... + a_k x_{0k}.$$

Prognoza przedziałowa wartości zmiennej objaśnianej *y*. Jest to przedział losowy postaci:

$$\left(\hat{y}_0 - t_\alpha S_\varepsilon \sqrt{1 + \mathbf{x}_0^T \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{x}_0}\,,\ \hat{y}_0 + t_\alpha S_\varepsilon \sqrt{1 + \mathbf{x}_0^T \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{x}_0} \right),$$

gdzie: t_α – wartość krytyczna rozkładu t Studenta o $n - k - 1$ stopniach swobody odpowiadająca przyjętemu poziomowi ufności $1 - \alpha$ taka, że

$$\left\{ P\left(|t| \geq t_\alpha \right) = \alpha \right\},$$

S_ε – estymator odchylenia standardowego składnika losowego modelu ekonometrycznego.

Prognoza przedziałowa wartości oczekiwanej zmiennej objaśnianej *y*. Dla ustalonego poziomu ufności $1 - \alpha$ jest to przedział losowy postaci:

$$\left(\hat{y}_0 - t_\alpha S_\varepsilon \sqrt{\mathbf{x}_0^T \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{x}_0}\,,\ \hat{y}_0 + t_\alpha S_\varepsilon \sqrt{\mathbf{x}_0^T \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{x}_0} \right),$$

gdzie: t_α – wartość krytyczna rozkładu t Studenta o $n - k -$ 1stopniach swobody odpowiadająca przyjętemu poziomowi ufności $1 - \alpha$ taka, że

$$\left\{ P\left(|t| \geq t_\alpha \right) = \alpha \right\},$$

S_ε – estymator odchylenia standardowego składnika losowego modelu ekonometrycznego.

ROZDZIAŁ 2

SCHEMAT WERYFIKACJI STATYSTYCZNEJ MODELU EKONOMETRYCZNEGO

Wyznaczony metodą najmniejszych kwadratów model ekonometryczny

$$\hat{y} = a_0 + a_1 x_1 + a_2 x_2 + ... + a_k x_k$$

musi być poddany weryfikacji statystycznej. W rozdziale tym omówiono podstawowe statystyki wykorzystywane do określenia stopnia dopasowania modelu do danych rzeczywistych, testy statystyczne weryfikujące istotność współczynników modelu ekonometrycznego oraz testy weryfikujące spełnienie założeń Gaussa–Markowa.

2.1. Dopasowanie modelu do danych empirycznych

Podstawowe miary dopasowania modelu do danych rzeczywistych to:

- błąd standardowy składnika losowego równania regresji S_ε

$$S_\varepsilon = \sqrt{\frac{\sum_{t=1}^{n} e_t^2}{n-k-1}} = \sqrt{\frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{n-k-1}},$$

przy czym: y_t – rzeczywista wartość zmiennej objaśnianej,

$\hat{y}_t$ – wartość zmiennej objaśnianej wyznaczona na podstawie modelu,

$e_t = y_t - \hat{y}_t$ – reszty modelu.

Im mniejsza wartość s_ε, tym model lepiej opisuje rzeczywistość

- współczynnik zbieżności φ^2:

$$\varphi^2 = \frac{\sum_{t=1}^{n} e_t^2}{\sum_{t=1}^{n} (y_t - \bar{y})^2},$$

gdzie $\bar{y}$ – wartość średnia zmiennej objaśnianej y.

- współczynnik determinacji:

$$R^2 = 1 - \varphi^2.$$

Arbitralnie ustala się dopuszczalną wartość graniczną R^2 (jest to zazwyczaj wielkość około 0,6)[2].

Miarą dopasowania modeli nieliniowych jest ponadto

- wskaźnik średniego względnego dopasowania modelu Ψ:

$$\Psi = \frac{1}{n} \sum_{t=1}^{n} \frac{|E_t|}{|\hat{y}_t|},$$

gdzie E_t – reszty modelu nieliniowego.
W sposób arbitralny ustala się dopuszczalną wartość graniczną Ψ (jest to zazwyczaj wielkość około 0,1).

W przypadku modeli ekonometrycznych z wieloma zmiennymi objaśniającymi należy ponadto sprawdzić, czy spełnione są warunki:

- koincydencji:

$$\operatorname{sign}\left(r\left(x_j, y\right)\right) = \operatorname{sign}\left(a_j\right),$$

gdzie: sign($r(x_j, y)$) – znak współczynnika korelacji pomiędzy zmienną objaśniającą x_j a zmienną objaśnianą y,
sign(a_j) – znak współczynnika a_j w modelu ekonometrycznym przy zmiennej x_j.

---

[2] Stosuje się także skorygowany współczynnik determinacji $\tilde{R}^2 = 1 - \left(1 - R^2\right)\frac{n-1}{n-k}$. Współczynnik ten może przyjmować wartości z przedziału (–∞, 1). Stosowany jest do porównania dopasowania modeli ekonometrycznych z różną liczbą zmiennych objaśniających.

W przypadku modeli nieliniowych, w których zmienna objaśniana y jest transformowana stosuje się także współczynnik „quasi R^2" $= 1 - \frac{\sum_{t=1}^{n} E_t^2}{\sum_{t=1}^{n} (y_t - \bar{y})^2}$. Współczynnik ten ma zastosowanie do porównania dopasowania modeli ekonometrycznych z różnymi kształtami funkcji.

Zgodność znaków współczynnika korelacji i współczynnika modelu ekonometrycznego musi zachodzić dla wszystkich zmiennych objaśniających. Jeżeli zmienne objaśniające są liniowo niezależne, to warunek ten jest spełniony.

2.2. Istotność układu współczynników regresji

W procesie weryfikacji modelu ekonometrycznego w pierwszej kolejności należy sprawdzić, czy zachodzi zależność liniowa między zmienną objaśnianą y a którąkolwiek ze zmiennych objaśniających x_j modelu.

Test 1 – istotności układu współczynników regresji. Stawiamy hipotezy:

$$H_0 : \sum_{j=1}^{n} \alpha_j^2 = 0,$$

$$H_1 : \sum_{j=1}^{n} \alpha_j^2 \neq 0.$$

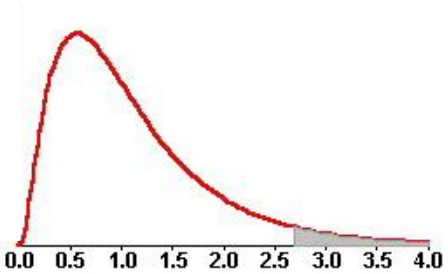
Sprawdzianem zespołu hipotez jest statystyka

$$F = \frac{R^2}{1-R^2} \frac{n-k-1}{k}.$$

Statystyka ta, przy założeniu prawdziwości hipotezy zerowej, ma rozkład F Snedecora o k stopniach swobody licznika oraz o $(n - k - 1)$ stopniach swobody mianownika.

Obszar krytyczny testu jest prawostronny

$$\Theta = \{F : P(F \geq F_\alpha) = \alpha\}.$$



Rys. 2.1. Obszar krytyczny testu

Jeżeli zatem wyznaczona wartość empiryczna statystyki F jest mniejsza od wartości krytycznej F_α ($F < F_\alpha$), to nie ma podstaw do odrzucenia hipotezy H_0 na korzyść hipotezy alternatywnej H_1. Nie zachodzi związek liniowy między zmienną objaśnianą y a żadną ze zmiennych objaśniających x_j. Oznacza to, iż badany model ekonometryczny jest niepoprawny.

W przeciwnym razie, gdy $F \geq F_\alpha$, przyjmujemy hipotezę H_1, a więc uznajemy, że między zmienną y a przynajmniej jedną ze zmiennych uwzględnionych w modelu zachodzi zależność liniowa.

2.3. Istotność poszczególnych współczynników regresji

W poprawnym modelu ekonometrycznym zmienna objaśniana y musi istotnie zależeć od każdej ze zmiennych objaśniających x_j modelu. Test weryfikujący ten fakt jest następujący.

Test 2 – istotności poszczególnych współczynników regresji. Dla każdego współczynnika równania regresji ($j = 0, 1, ..., k$) stawiamy hipotezy:

$H_0 : \alpha_j = 0$,

$H_1 : \alpha_j \neq 0$.

Sprawdzianem zespołu hipotez jest statystyka

$$t = \frac{a_j}{S(\alpha_j)},$$

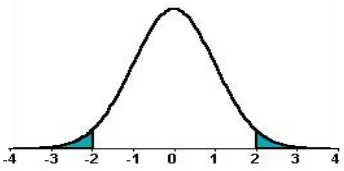
gdzie: a_j – estymator współczynnika α_j,

$S(\alpha_j) = \sqrt{d_{jj}}$ – estymator dyspersji współczynnika α_j.

Statystyka ta, przy prawdziwości hipotezy zerowej, ma rozkład t Studenta o $(n - k - 1)$ stopniach swobody.

Obszar krytyczny testu jest dwustronny

$$\Theta = \{t : P(|t| \geq t_\alpha) = \alpha\}.$$



Rys. 2.2. Obszar krytyczny testu

Jeżeli zatem dla którejkolwiek zmiennej objaśniającej wyznaczona wartość empiryczna statystyki t jest mniejsza w module od wartości krytycznej t_α($|t| < t_\alpha$), to nie ma podstaw do odrzucenia hipotezy H_0 na korzyść hipotezy H_1. Oznacza to, że zmienna ta jest nieistotna (nie ma zależności liniowej między tą zmienną a zmienną objaśnianą). Nieistotność jakiejkolwiek zmiennej objaśniającej wymaga powtórnego sformułowania modelu.

Jeżeli dla wszystkich zmiennych objaśniających x_1, x_2, ..., x_k zachodzi $|t| \geq t_\alpha$, to przyjmujemy hipotezę H_1, a więc mamy podstawę do przyjęcia, że między zmienną objaśnianą y a wszystkimi zmiennymi objaśniającymi uwzględnionymi w modelu zachodzi zależność liniowa.

2.4. Własności składników losowych

Trzeci i czwarty warunek Gaussa–Markowa formułują własności składnika losowego modelu ekonometrycznego, których spełnienie jest wymagane dla zapewnienia efektywności estymatorów współczynników modelu, tj.:

- Składniki losowe dla poszczególnych wartości zmiennych objaśniających mają rozkłady normalne o wartości oczekiwanej zero i stałej wariancji: $N(0, \delta_\varepsilon)$.
- Składniki losowe nie są ze sobą skorelowane.

Przedstawimy niektóre z testów statystycznych stosowanych do weryfikacji spełnienia warunków Gaussa–Markowa.

2.4.1. Normalność

Wybór testu zależy od wielkości próby (liczba obserwacji). W przypadku dużej próby hipotezę o normalności składników losowych weryfikujemy testem zgodności χ^2 lub testem λ Kołmogorowa[3]. Dla małych prób możemy stosować test Shapiro–Wilka lub test Dawida–Hellwiga.

TESTY DLA DUŻEJ LICZBY OBSERWACJI

Test 3 χ^2. Stawiamy hipotezę

H_0: składniki losowe mają rozkład $N(0, S_\varepsilon)$.

Sprawdzianem hipotezy jest statystyka

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i},$$

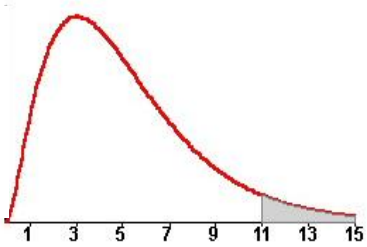
[3] W modelowaniu ekonometrycznym testy te rzadko mają zastosowanie, gdyż najczęściej równania regresji budujemy na podstawie małej próby.

gdzie: r – liczba klas szeregu rozdzielczego,
n_i – liczba obserwacji w i-tej klasie $n_i \geq 5$,
p_i – prawdopodobieństwo hipotetyczne zaobserwowania wartości składnika losowego w i-tej klasie.

Statystyka ta, przy prawdziwości hipotezy H_0, ma rozkład χ^2 o $(r - 2)$ stopniach swobody.

Obszar krytyczny testu jest prawostronny

$$\Theta = \left\{\chi^2 : P\left(\chi^2 \geq \chi^2_\alpha\right) = \alpha\right\}.$$



Rys. 2.3. Obszar krytyczny testu

Jeżeli zatem wyznaczona wartość empiryczna statystyki χ^2 jest mniejsza od wartości krytycznej $\chi^2_\alpha\left(\chi^2 < \chi^2_\alpha\right)$, to nie ma podstaw do odrzucenia hipotezy H_0 o normalności rozkładu składników losowych.

Test 4 – λ Kołmogorowa. Stawiamy hipotezę:

H_0: składniki losowe mają rozkład $N(0, S_\varepsilon)$.

Sprawdzianem tej hipotezy jest statystyka λ Kołmogorowa

$$\lambda = \sqrt{n} \cdot \sup_x \left|F^*(x) - F(x)\right|,$$

gdzie: $F^*(x)$ – dystrybuanta empiryczna składnika losowego modelu,
$F(x)$ – dystrybuanta hipotetyczna składnika losowego modelu.

Obszar krytyczny testu jest prawostronny:

$$\Theta = \left\{\lambda : P\left(\lambda \geq \lambda_\alpha\right) = \alpha\right\}.$$

Jeżeli zatem wyznaczona wartość empiryczna statystyki λ jest mniejsza od wartości krytycznej $\lambda_\alpha (\lambda < \lambda_\alpha)$, to nie ma podstaw do odrzucenia hipotezy H_0 o normalności rozkładu składników losowych.

TESTY DLA MAŁEJ PRÓBY

Test 5. Shapiro–Wilka. Stawiamy hipotezę:
H_0: składniki losowe mają rozkład $N(0, S_\varepsilon)$.

Sprawdzianem hipotezy jest statystyka

$$W = \frac{\left[\sum_{i=1}^{\left[\frac{n}{2}\right]} a_{n,i}\left(e_{(n-i+1)} - e_{(i)}\right)\right]^2}{\sum_{i=1}^{n}\left(e_i - \bar{e}\right)^2},$$

przy czym:

$$\sum_{i=1}^{n} a_{n,i} = 0 \text{ oraz } \sum_{i=1}^{n} a_{n,i}^2 = 1,$$

$$\bar{e} = 0,$$

gdzie: $a_{n,i}$ – współczynniki (stablicowane przez Shapiro–Wilka),
$e_{(1)}, e_{(2)}, ..., e_{(n)}$ – wartości reszt uporządkowane niemalejąco.

Obszar krytyczny testu jest następujący:

$$\Theta = \left\{W : P\left(W \le W_\alpha\right) = \alpha\right\}.$$

Statystyka W jest statystyką pozycyjną. Jeżeli zatem wyznaczona wartość empiryczna statystyki W jest nie mniejsza od wartości krytycznej $W_\alpha (W \ge W_\alpha)$, to nie ma podstaw do odrzucenia hipotezy H_0 o normalności rozkładu składników losowych.

Test 6 – Davida–Hellwiga. Stawiamy hipotezę:

H_0: składniki losowe mają rozkład $N(0, S_\varepsilon)$.

Test ten wykorzystuje to, że każda dystrybuanta rozkładu ciągłego ma rozkład jednostajny na odcinku [0, 1]. Procedura testowania jest następująca:

- Konstruujemy cele, dzieląc odcinek [0, 1] na n rozłącznych odcinków o długości $1/n$

$$\left[0, \frac{1}{n}\right), \left[\frac{1}{n}, \frac{2}{n}\right)\left[\frac{2}{n}, \frac{3}{n}\right), ..., \left[\frac{n-1}{n}, 1\right).$$

- Następnie wyznaczamy wartości dystrybuanty hipotetycznej dla wszystkich wartości reszt modelu $F(e_i)$ (dla $i = 1, 2, ..., n$).
- Sprawdzamy, do których cel należą wyznaczone wartości dystrybuanty. Wyznaczamy liczbę k pustych celi, do których nie wpadła żadna wartość $F(e_i)$.

Obszar krytyczny testu jest dwustronny:

$$\Theta = \left\{ k : P(k \le k_1) = \frac{\alpha}{2} \right\} \cup \left\{ k : P(k \ge k_2) = \frac{\alpha}{2} \right\}.$$

Jeżeli zatem wyznaczona wartość empiryczna statystyki k nie wpada do obszaru krytycznego ($k \in (k_1, k_2)$), to nie ma podstaw do odrzucenia hipotezy H_0 o normalności rozkładu składników losowych.

2.4.2. Autokorelacja

Autokorelacja to współzależność składników losowych i w sposób oczywisty nie jest pożądana. Podstawowe przyczyny występowania autokorelacji to:

- niewłaściwie dobrana postać modelu ekonometrycznego,
- nieuwzględnienie w modelu istotnej zmiennej (objaśnianej, objaśniającej), w szczególności opóźnionej w czasie,
- cykliczność analizowanego zjawiska.

Stopień autokorelacji τ można ustalić na podstawie analizy właściwości badanego zjawiska lub można przyjąć τ odpowiadające największej wartości współczynnika korelacji $\rho(\varepsilon_t, \varepsilon_{t-\tau})$:

$$\rho_\tau = \rho(\varepsilon_t, \varepsilon_{t-\tau}) = \frac{\operatorname{cov}(\varepsilon_t, \varepsilon_{t-\tau})}{\sqrt{D^2(\varepsilon_t) D^2(\varepsilon_{t-\tau})}}.$$

Współczynnik autokorelacji $\rho(\varepsilon_t, \varepsilon_{t-\tau})$ nosi nazwę współczynnika autokorelacji rzędu τ.

Opracowano wiele testów, które umożliwiają wykrycie autokorelacji składników losowych. Każdy z tych testów wymaga odpowiedniego uszeregowania obserwacji błędu losowego zgodnego ze zjawiskiem autokorelacji.

AUTOKORELACJA RZĘDU PIERWSZEGO

W przypadku $\tau = 1$ (proces autokorelacyjny AR(1)) hipotezę o braku autokorelacji składników losowych weryfikujemy testem Durbina–Watsona:

Test 7 – Durbina–Watsona. Stawiamy hipotezę:

$H_0 : \rho(\varepsilon_t, \varepsilon_{t-1}) = 0$,

$H_1 : \rho(\varepsilon_t, \varepsilon_{t-1}) > 0$ lub $H_1 : \rho(\varepsilon_t, \varepsilon_{t-1}) < 0$, lub $H_1 : \rho(\varepsilon_t, \varepsilon_{t-1}) \neq 0$.

Sprawdzianem zespołu hipotez jest statystyka

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}.$$

Tablice statystyczne[4] podają wartości krytyczne d_L oraz d_U dla określonej liczby obserwacji n oraz liczby zmiennych w modelu k.

- Jeżeli hipoteza alternatywna jest postaci: H_1: $\rho(\varepsilon_t, \varepsilon_{t-1}) > 0$.

Hipotezę H_0 odrzucamy, jeżeli zachodzi nierówność $d < d_L$, a zatem przyjmujemy istnienie dodatniej autokorelacji. Nie mamy podstaw do odrzucenia hipotezy H_0, gdy $d > d_U$. Nierówność $d_L \le d \le d_U$ natomiast nie umożliwia rozstrzygnięcia.

- Jeżeli hipoteza alternatywna jest postaci: H_1: $\rho(\varepsilon_t, \varepsilon_{t-1}) < 0$.

Hipotezę H_0 odrzucamy, jeżeli zachodzi nierówność $d' = (4 - d) < d_L$, a zatem przyjmujemy istnienie ujemnej autokorelacji. Nie mamy podstaw do odrzucenia hipotezy H_0, gdy $d' = (4 - d) > d_U$. Nierówność $d_L < (4 - d) \le d_U$ natomiast nie umożliwia rozstrzygnięcia.

- Jeżeli hipoteza alternatywna jest postaci H_1: $\rho(\varepsilon_t, \varepsilon_{t-1}) \neq 0$.

Gdy zachodzi nierówność $d < d_L$ lub $d' = 4 - d < d_L$ odrzucamy hipotezę zerową i przyjmujemy istnienie autokorelacji. Nie mamy podstaw do odrzucenia hipotezy H_0 o braku autokorelacji, gdy zachodzi nierówność $d > d_U$ lub $4 - d > d_L$. Nierówność $d_L \le d \le d_U$ lub $(4 - d_U) \le d \le (4 - d_L)$ nie umożliwia rozstrzygnięcia.

Jeżeli stwierdzono autokorelację składników losowych, to można próbować ją wyeliminować, stosując przekształcenie Cochrana–Orcutta polegające na przejściu od modelu

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_k x_k + \varepsilon_t$$

do modelu:

$$y' = \alpha'_0 + \alpha'_1 x'_1 + \alpha'_2 x'_2 + ... + \alpha'_k x'_k + \varepsilon'_t ,$$

przy czym dla $i = 2, 3, ..., n$; $j = 2, 3, ..., k$,

$$y'_i = y_i - r_1 y_{i-1}$$

$$x'_{ij} = x_{ij} - r_1 x_{i-1,j}$$

---

[4] Wartości krytyczne podane w tych tablicach można również wykorzystać przy testowaniu statystyką $d_4 = \dfrac{\sum_{t=5}^{n} (e_t - e_{t-4})^2}{\sum_{t=1}^{n} e_t^2}$ zjawiska autokorelacji dla modeli autoregresyjnych AR(4), np. gdy dane analizowane są w układzie kwartalnym.

gdzie r_1 jest estymatorem współczynnika autokorelacji [5] między składnikami losowymi modelu dla $\tau = 1$. Współczynnik ten nazywany jest współczynnikiem autokorelacji.

Procedurę stosujemy iteracyjnie aż do usunięcia autokorelacji z modelu.

Analogicznym do testu Durbina–Watsona jest test von Neumanna.

Test 8 – von Neumanna. Stawiamy hipotezy:
H_1: $\rho(\varepsilon_t, \varepsilon_{t-1}) = 0$,
H_1: $\rho(\varepsilon_t, \varepsilon_{t-1}) > 0$ (H_1: $\rho(\varepsilon_t, \varepsilon_{t-1}) < 0$; H_1: $\rho(\varepsilon_t, \varepsilon_{t-1}) \neq 0$).
Sprawdzianem zespołu hipotez jest statystyka

$$Q = \frac{n\sum_{t=2}^{n}(e_t - e_{t-1})^2}{(n-1)\sum_{t=1}^{n} e_t^2}.$$

Obszar krytyczny testu jest lewostronny (prawostronny, dwustronny)

$$\Theta = \{Q : P(Q \leq Q_\alpha) = \alpha\}.$$

Jeżeli zatem wyznaczona wartość empiryczna statystyki jest mniejsza od wartości krytycznej $Q > Q_a$, to nie ma podstaw do odrzucenia hipotezy H_0 o braku autokorelacji składników losowych rzędu τ na korzyść hipotezy H_1.

Dla dużej liczby obserwacji ($n > 60$) statystyka Q ma asymptotyczny rozkład normalny $N\left(\frac{2n}{n-1}, \sqrt{\frac{4}{n}}\right)$.

Test 9 – Durbina. Dla modeli autoregresyjnych AR(1), w których opóźniona o okres zmienna objaśniana jest jedną ze zmiennych objaśniających statystyka Durbina–Watsona jest statystyką obciążoną. W tym przypadku do zbadania zjawiska autokorelacji można zastosować test Durbina. Test ten można stosować również wówczas, gdy w modelu występują inne opóźnienia zmiennej objaśnianej.

Stawiamy hipotezy:
H_0: $\rho(\varepsilon_t, \varepsilon_{t-1}) = 0$,
H_1: $\rho(\varepsilon_t, \varepsilon_{t-1}) \neq 0$.

---

[5] Za estymator współczynnika autokorekcji reszt r_1 można przyjąć jedną ze statystyk:

$$1 - \frac{d}{2} \text{ lub } \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2} \text{ albo } \frac{n-k}{n-1}\frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2} \text{ lub } \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sqrt{\sum_{t=1}^{n} e_t^2 \sum_{t=2}^{n} e_{t-1}^2}}.$$

Sprawdzianem zespołu hipotez jest statystyka

$$h = \left(1 - \frac{1}{2}d\right)\sqrt{\frac{n}{1 - nS^2_{\alpha_{y(-1)}}}},$$

przy czym[6]:

$$1 - nS^2_{a_{y(-1)}} > 0,$$

gdzie: d – wartość statystyki Durbina–Watsona,

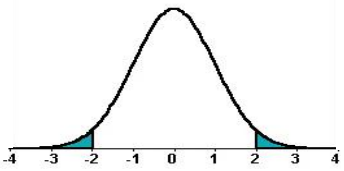
$S^2_{\alpha_{y(-1)}}$ – wariancja estymatora współczynnika regresji przy zmiennej opóźnionej.

Jeżeli $1 - nS^2_{\alpha_{y(-1)}} > 0$, to statystyka ta, przy prawdziwości hipotezy H_0, ma rozkład graniczny normalny $N(0, 1)$.

Obszar krytyczny testu jest dwustronny

$$\Theta = \{u : P(|u| \geq u_\alpha) = \alpha\},$$

przy czym U – zmienna losowa o rozkładzie normalnym $N(0, 1)$.



Rys. 2.4. Obszar krytyczny testu

Jeżeli zatem wyznaczona wartość empiryczna statystyki $|h|$ jest mniejsza co do modułu od wartości krytycznej $|h| < u_a$, to nie ma podstaw do odrzucenia hipotezy H_0 o braku autokorelacji składników losowych na korzyść hipotezy H_1.

AUTOKORELACJA DOWOLNEGO RZĘDU

Test 10 – istotności autokorelacji rzędu τ składników losowych. Stawiamy hipotezy:

H_0: $\rho(\varepsilon_t, \varepsilon_{t-\tau}) = 0$,

H_1: $\rho(\varepsilon_t, \varepsilon_{t-\tau}) \neq 0$ lub H_1: $\rho(\varepsilon_t, \varepsilon_{t-\tau}) > 0$, lub H_1: $\rho(\varepsilon_t, \varepsilon_{t-\tau}) < 0$.

Sprawdzianem zespołu hipotez jest statystyka

$$t = \frac{r_\tau\sqrt{n - \tau - 2}}{\sqrt{1 - r_\tau^2}},$$

---

[6] Jeżeli $1 - nS^2_{\alpha_{y(-1)}} \leq 0$, występowanie autokorelacji można zweryfikować, budując model ekonometryczny zależności ε_t od ε_{t-1}, y_{t-1}, x_1, x_2, ..., x_k, a następnie zweryfikować istotność współczynnika przy ε_{t-1}.

gdzie:

$$r_\tau = \frac{\sum_{t=\tau+1}^{n}(e_t - \bar{e}_1)(e_{t-\tau} - \bar{e}_2)}{\sqrt{\sum_{t=\tau+1}^{n}(e_t - \bar{e}_1)\sum_{t=1}^{n-\tau}(e_t - \bar{e}_2)}},$$

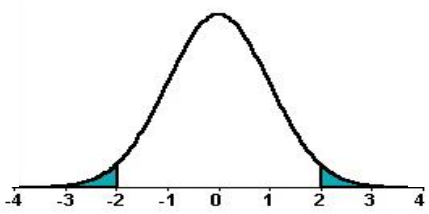
przy czym:

$$\bar{e}_1 = \frac{1}{n-\tau}\sum_{t=\tau+1}^{n} e_t \text{ oraz } \bar{e}_2 = \frac{1}{n-\tau}\sum_{t=1}^{n-\tau} e_t .$$

Statystyka ta, przy prawdziwości hipotezy zerowej, ma rozkład *t* Studenta o ($n - \tau - 2$) stopniach swobody.

Obszar krytyczny testu w przypadku hipotezy alternatywnej postaci: H_1: $\rho(\varepsilon_t, \varepsilon_{t-\tau}) \neq 0$ jest dwustronny

$$\Theta = \{t : P(|t| \geq t_\alpha) = \alpha\}.$$



Rys. 2.5. Obszar krytyczny testu

Jeżeli zatem wyznaczona wartość empiryczna statystyki jest mniejsza od wartości krytycznej $t < t_\alpha$, to nie ma podstaw do odrzucenia hipotezy H_0 o braku autokorelacji składników losowych rzędu τ na korzyść hipotezy H_1.

W przypadku hipotez H_1: $\rho(\varepsilon_t, \varepsilon_{t-\tau}) > 0$ oraz H_1: $\rho(\varepsilon_t, \varepsilon_{t-\tau}) < 0$ obszar krytyczny jest odpowiednio prawo- i lewostronny.

Test 11 – istotności autokorelacji dowolnego rzędu. Stawiamy hipotezy:

H_0: brak autokorelacji,

$H_1 : \varepsilon_t = AR(r) = \gamma_1\varepsilon_{t-1} + \gamma_2\varepsilon_{t-2} + \ldots + \gamma_r\varepsilon_{t-r} + u_t$.

Sprawdzianem zespołu hipotez jest statystyka

$$\chi^2 = \frac{e^T\mathbf{E}\left(\mathbf{E}^T\mathbf{E} - \mathbf{E}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{E}\right)^{-1}\mathbf{E}^T e}{S_\varepsilon^2},$$

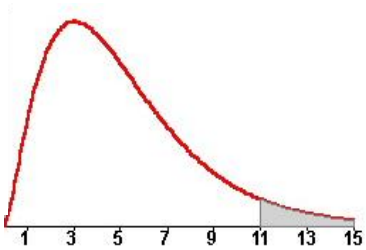
gdzie $e = (e_1, e_2, ..., e_n)$ – reszty modelu ekonometrycznego,

$$\mathbf{E}=\begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ e_1 & 0 & \cdots & 0 & 0 \\ e_2 & e_1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ e_{n-1} & e_{n-2} & \dots & e_{n-r-1} & e_{n-r} \end{bmatrix},$$

$$\mathbf{X}=\begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}.$$

Statystyka ta, przy prawdziwości hipotezy H_0, ma rozkład χ^2 o r stopniach swobody. Obszar krytyczny testu jest prawostronny

$$\Theta=\left\{\chi^2 : P\left(\chi^2 \geq \chi_a^2\right)=a\right\}.$$



Rys. 2.6. Obszar krytyczny testu

Jeżeli zatem wyznaczona wartość empiryczna statystyki χ^2 jest mniejsza od wartości krytycznej $\chi_\alpha^2\left(\chi^2 < \chi_\alpha^2\right)$, to nie ma podstaw do odrzucenia hipotezy H_0 o braku autokorelacji składników losowych na korzyść hipotezy H_1.

Test ten można również stosować w przypadku modeli autoregresyjnych ze średnią ruchomą $MA(r)$.

2.4.3. Symetria

Składniki losowe powinny mieć rozkład normalny, który jest rozkładem symetrycznym. Test poniższy sprawdza, czy frakcja reszt dodatnich p_+ i ujemnych p_- równa się 0,5.

Niech m oznacza liczbę reszt *in plus* (dodatnie reszty modelu).

Test 12 – symetrii składników losowych. Stawiamy hipotezy:

$H_0 : p_+ = \frac{1}{2}$,

$H_1 : p_+ \neq \frac{1}{2}$.

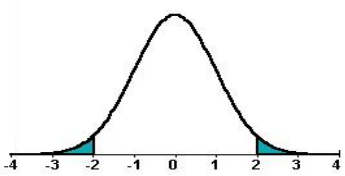
Sprawdzianem zespołu hipotez jest statystyka

$$t = \frac{\frac{m}{n} - \frac{1}{2}}{\sqrt{\frac{\frac{m}{n}\left(1 - \frac{m}{n}\right)}{n-1}}}.$$

Statystyka ta, przy prawdziwości hipotezy zerowej, ma rozkład *t* Studenta o $(n - 1)$ stopniach swobody.

Obszar krytyczny testu jest dwustronny

$$\Theta = \left\{t : P\left(|t| \geq t_a\right) = a\right\}.$$



Rys. 2.7. Obszar krytyczny testu

Jeżeli zatem wyznaczona wartość empiryczna statystyki jest mniejsza w module od wartości krytycznej $|t| < t_\alpha$, to nie ma podstaw do odrzucenia hipotezy H_0 na korzyść hipotezy H_1, tzn., że składniki losowe modelu są symetryczne.

2.4.4. Losowość

Na tym etapie weryfikujemy losowość rozkładu reszt modelu. Brak losowości może oznaczać:

- cykliczność badanej zmiennej zależnej *y*,
- niewłaściwe dobranie postaci analitycznej modelu ekonometrycznego.

Przedstawimy dwa testy losowości.

Test 13 – liczby serii. Stawiamy hipotezę:

H_0: błąd modelu jest losowy.

• Porządkujemy reszty chronologicznie lub zgodnie z rosnącymi wartościami jednej ze zmiennych objaśniających.

• Wyznaczamy liczbę serii L reszt tych samych znaków.

Przy prawdziwości hipotezy H_0 zmienna losowa L podlega rozkładowi liczby serii dla m elementów jednego rodzaju (reszty dodatnie) oraz $(n - m)$ elementów drugiego rodzaju (reszty ujemne)[7].

Obszar krytyczny testu jest dwustronny

$$\Theta = \left\{ L : P(L \le L_1) = \frac{\alpha}{2} \right\} \cup \left\{ L : P(L \ge L_2) = \frac{\alpha}{2} \right\}.$$

Jeżeli zatem wyznaczona wartość empiryczna statystyki nie wpada do obszaru krytycznego $L \in (L_1, L_2)$, to nie ma podstaw do odrzucenia hipotezy H_0 o losowości reszt modelu.

Uwaga. Dla dużej próby, gdy m, $(n - m) \to \infty$, rozkład liczby serii ma rozkład normalny:

$$N\left(\frac{2m(n-m)}{n} + 1, \sqrt{\frac{2m(n-m)(2m(n-m)-n)}{n^2(n-1)}} \right).$$

Test 14 – maksymalnej długości serii. Stawiamy hipotezę:

H_0: błąd modelu jest losowy.

• Porządkujemy reszty chronologicznie lub zgodnie z rosnącymi wartościami jednej ze zmiennych objaśniających.

• Wyznaczamy maksymalną długość serii $L_{\max}$ reszt tych samych znaków.

Obszar krytyczny testu jest prawostronny. Tablice statystyczne podają wartość minimalnej wielkości próby statystycznej, dla której dana długość serii $L_{\max}$ jest dopuszczalna dla zadanego poziomu istotności α.

2.4.5. Homoskedastyczność

Równość wariancji w podpróbach homogenicznych ze względu na wariancję składników losowych można przeprowadzić na podstawie testu Goldfelda–Quandta lub badając istotność współczynnika korelacji modułów składników losowych i czasu.

---

[7] Mediana rozkładu normalnego unormowanego równa się zeru.

Test 15 – Goldfelda–Quandta. Dla podprób o najmniejszej i największej wariancji (o liczebnościach odpowiednio n_1, n_2) budujemy równania regresji, a następnie stawiamy hipotezy:

$H_0 : \delta_{\varepsilon_1}^2 = \delta_{\varepsilon_2}^2$,

$H_1 : \delta_{\varepsilon_1}^2 > \delta_{\varepsilon_2}^2$ lub $H_1 : \delta_{\varepsilon_1}^2 < \delta_{\varepsilon_2}^2$.

Sprawdzianem zespołu hipotez jest statystyka

$$F = \frac{\max\left(S_{\varepsilon_1}^2, S_{\varepsilon_2}^2\right)}{\min\left(S_{\varepsilon_1}^2, S_{\varepsilon_2}^2\right)},$$

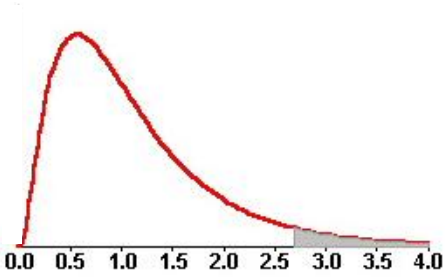
gdzie: $S_{\varepsilon_1}^2$ – estymator wariancji składników losowych modelu regresji dla pierwszej podpróby,

$S_{\varepsilon_2}^2$ – estymator wariancji składników losowych modelu regresji dla drugiej podpróby.

Przy prawdziwości hipotezy zerowej statystyka F ma rozkład F Snedecora o ($n_2 - k - 1$) stopniach swobody licznika i o ($n_1 - k - 1$) stopniach swobody mianownika.

Obszar krytyczny testu jest prawostronny

$$\Theta = \left\{F : P\left(F \geq F_\alpha\right) = \alpha\right\}.$$



Rys. 2.8. Obszar krytyczny testu

Jeżeli zatem wyznaczona wartość empiryczna statystyki F jest mniejsza od wartości krytyczne F_α: ($F < F_\alpha$), to nie ma podstaw do odrzucenia hipotezy H_0 o homoskedastyczności składników losowych modelu.

Test 16 – korelacji modułów składników losowych i czasu. Stałość wariacji składników losowych w czasie można również zbadać testem istotności współczynnika korelacji modułów reszt modelu i czasu (lub pewnej zmiennej objaśniającej zgodnie ze zjawiskiem autokorelacji).

Stawiamy hipotezy:

$H_0 : \rho(|\varepsilon_t|, t) = 0$,

$H_1 : \rho(|\varepsilon_t|, t) \neq 0$.

Sprawdzianem zespołu hipotez jest statystyka

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}\,,$$

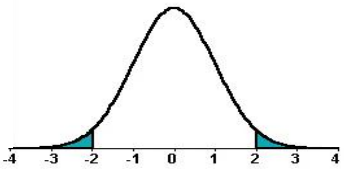
gdzie

$$r\left(|\varepsilon|,t\right) = \frac{\sum\left(|e_t|-|\bar{e}|\right)(t-\bar{t})}{\sqrt{\sum\left(|e_t|-|\bar{e}|\right)^2\sum(t-\bar{t})^2}}$$

Statystyka ta, przy prawdziwości hipotezy H_0, ma rozkład t Studenta o $(n-2)$ stopniach swobody.

Obszar krytyczny testu jest dwustronny

$$\Theta = \left\{t : P\left(|t| \geq t_\alpha\right) = \alpha\right\}.$$



Rys. 2.9. Obszar krytyczny testu

Jeżeli zatem wyznaczona wartość empiryczna statystyki jest mniejsza w module od wartości krytycznej $|t| < t_\alpha$, to nie ma podstaw do odrzucenia hipotezy H_0 o homoskedastyczności składników losowych modelu, na korzyść hipotezy H_1, że wariancja składników losowych zmienia się w czasie lub wraz ze wzrostem (spadkiem) pewnej zmiennej objaśniającej.

Test 17 – korelacji rangowej Spearmana. Test ten pozwala sprawdzić, czy wariancja składników losowych rośnie (maleje) wraz ze wzrostem wartości zmiennej objaśniającej x.

Stawiamy hipotezy:

$H_0 : \rho(|\varepsilon_x|, x) = 0$,

$H_1 : \rho(|\varepsilon_x|, x) \neq 0$.

Sprawdzianem zespołu hipotez jest statystyka korelacji rangowej Spearmana

$$r = r\left(|\varepsilon|, x\right) = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n\left(n^2-1\right)},$$

gdzie D_i – różnica rang zmiennej x oraz modułu reszt modelu dla i-tej obserwacji.

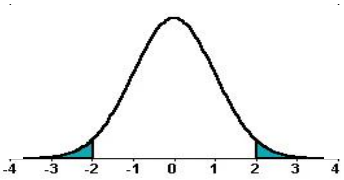
Rangę (1, 2, ..., *n*) przypisujemy kolejno wartościom zmiennej *x* (reszt *e*) uporządkowanym w ciąg niemalejący. Jeżeli wystąpią takie same wartości zmiennej *x* (reszt *e*), to przypisujemy im rangę równą średniej arytmetycznej odpowiadających im pozycji w ciągu.

Statystyka *r*, przy prawdziwości hipotezy H_0, ma rozkład asymptotycznie normalny $N\left(0, \frac{1}{\sqrt{n-1}}\right)$ (w praktyce dla $n > 10$).

Obszar krytyczny testu jest dwustronny:

$$\Theta = \left\{u : P\left(|u| \geq u_\alpha\right) = \alpha\right\},$$

przy czym *U* to zmienna losowa o rozkładzie normalnym *N*(0, 1).



Rys. 2.10. Obszar krytyczny testu

Jeżeli zatem dla wyznaczonej wartości empirycznej statystyki zachodzi $\left|r\sqrt{n-1}\right| < u_\alpha$, to nie ma podstaw do odrzucenia hipotezy H_0 o homoskedastyczności składników losowych modelu na korzyść hipotezy H_1.

2.4.6. Nieobciążoność składników losowych modeli nieliniowych

Dla modeli nieliniowych dodatkowo należy zbadać, czy składniki losowe modelu są nieobciążone. Wyznaczamy w tym celu reszty E_i modelu nieliniowego.

Test 18 – nieobciążoności składników losowych.

Stawiamy hipotezy

$H_0 : E(\tilde{\varepsilon}) = 0$,

$H_1 : E(\tilde{\varepsilon}) \neq 0$.

Sprawdzianem zespołu hipotez jest statystyka

$$t = \frac{\overline{E}}{S_E}\sqrt{n-1},$$

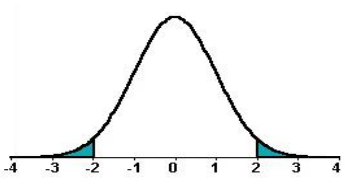
gdzie: $\overline{E}$ – średnia arytmetyczna reszty modelu nieliniowego.

S_E^2 – estymator wariancji składnika losowego modelu nieliniowego.

Statystyka ta, przy prawdziwości hipotezy H_0, ma rozkład t Studenta o $(n - 1)$ stopniach swobody.

Obszar krytyczny testu jest dwustronny

$$\Theta = \{t : P(|t| \geq t_\alpha) = \alpha\}.$$



Rys. 2.11. Obszar krytyczny testu

Jeżeli zatem wyznaczona wartość empiryczna statystyki jest mniejsza w module od wartości krytycznej ($|t| < t_\alpha$), to nie ma podstaw do odrzucenia hipotezy H_0 o nieobciążoności składników losowych na korzyść hipotezy H_1.

ROZDZIAŁ 3

MODELE EKONOMETRYCZNE
STUDIUM PRZYPADKU

W rozdziale przedstawiono kolejne kroki budowania i weryfikacji modeli ekonometrycznych dla rzeczywistych zagadnień. Aby zaprezentować Czytelnikowi szerokie możliwości stosowania ekonometrii, dobrano modele z różnych klas i z różnych dziedzin:

- *Czas podróży samochodem* w zależności od długości trasy – model liniowy z jedną zmienną objaśniającą.
- *Cena mieszkań* jako funkcja jego powierzchni – model nieliniowy (krzywa Tőrquista) z jedną zmienną objaśniającą.
- *Wzrost dzieci* jako funkcja wieku i płci – model liniowy z dwiema zmiennymi objaśniającymi (ilościową i jakościową).
- *Średnia temperatura we Wrocławiu* – model dwurównaniowy.
- *Podaż pieniądza w Polsce* – model autoregresyjny.
- *Bezrobocie* jako funkcja bezrobocia – model nieliniowy, autoregresyjny, okresowy ze zmienną opóźnioną w czasie i funkcją harmoniczną.

Modele te różnią się ponadto strukturą danych: niektóre dane analizowane są w układzie przekrojowym, podczas gdy inne występują jako szeregi czasowe.

Przykłady modeli starano się tak dobrać, aby zaprezentować różne możliwe warianty postępowania podczas konstrukcji modeli ekonometrycznych. Szczególny nacisk położono na to, w jaki sposób można wykorzystać niepomyślny dla weryfikowanego modelu ekonometrycznego wynik testu statystycznego w celu jego poprawy. Występowanie autokorelacji implikuje często konieczność uwzględnienia w modelu zmiennych opóźnionych w czasie. Brak losowości lub symetrii reszt modelu może wynikać z cykliczności badanej zmiennej lub nieliniowej zależności między zmienną objaśnianą a zmiennymi objaśniającymi. Heteroskedastyczność może być skutkiem nieliniowej zależności zmiennych lub różnej postaci analitycznej modeli ekonometrycznych dla podgrup o różnej wariancji składników losowych. Brak istotności stałej modelu może implikować brak liniowej zależności lub sugerować występowanie współzależności liniowej zmiennych objaśniających. Brak koincydencji zwykle świadczy o współliniowości zmiennych objaśniających.

W trakcie przedstawiania poszczególnych modeli przyjęto następującą konwencję:

- model pierwszy (czas podróży samochodem) został przedstawiony w całości, krok po kroku, zgodnie z wcześniejszą metodologią i z prezentacją koniecznych wzorów opisujących poszczególne statystyki,
- modele następne przedstawiono także w całości, jednakże tam, gdzie poszczególne etapy i kroki postępowania nie różnią się co do postaci od użytych w modelach wcześniejszych podano jedynie wartości obliczeń i otrzymany wniosek.

Każdy model zaprezentowany w tym rozdziale przeszedł pozytywnie wszystkie etapy weryfikacji statystycznej. Skonstruowane modele zastosowano do predykcji ekonometrycznej.

3.1. Czas podróży samochodem

Model opisujący zależność czasu podróży samochodem od długości trasy jest przykładem modelu liniowego z jedną zmienną objaśniającą. Struktura danych jest przekrojowa. Predykcja czasu podróży wyznaczonego na podstawie skonstruowanego modelu jest obarczona błędem względnym rzędu 3%.

Krok I. Określenie celu badań modelowych

Firma z siedzibą w Warszawie ma swoje przedstawicielstwo we Wrocławiu oraz w wielu miastach europejskich. Naszym celem jest określenie zależności czasu przejazdu od długości trasy z Warszawy do tych miejscowości.

Z wykładów fizyki wiemy, że czas przejazdu jest wprost proporcjonalny do przebytej drogi, jeżeli ruch jest jednostajny:

$$s = vt\ .$$

Jeżeli ruch odbywa się ze stałym przyspieszeniem, to zachodzi relacja:

$$s = a\frac{t^2}{2}\ .$$

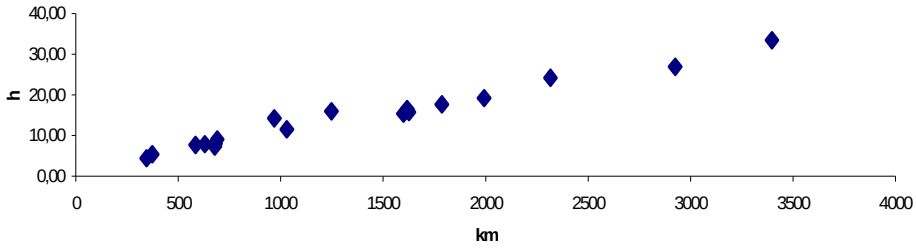
Nie mamy prostego wzoru, jeżeli ruch odbywa się z prędkością zmienną, a z taką przecież jeździmy samochodem – musielibyśmy wprowadzić pojęcie prędkości chwilowej, a przebytą drogę szacować jako całkę po niej. Rzecz sprowadza się nie tylko do tego, że jest to trudne matematycznie, ale i chyba niewykonalne w rzeczywistości. Spróbujemy więc zbudować model ekonometryczny, który pozwoli oszacować czas podróży w zależności od długości trasy i będzie uwzględniał wszystkie „nieregularności", z jakimi możemy spotkać się po drodze.

Krok II. Specyfikacja zmiennych wraz z gromadzeniem danych

Dane o odległości i czasie podróży podano w tabeli 3.1 i naniesiono na rysunku 3.1.

Tabela 3.1. Odległość i czas przejazdu. Opracowanie własne

Miejscowość docelowa	Odległość km	Czas h
Ateny	2317,1	24,28
Berlin	585,8	7,63
Bratysława	679,0	7,35
Budapeszt	691,5	9,05
Genewa	1598,1	15,42
Helsinki	968,8	14,30
Lizbona	3398,9	33,52
Londyn	1617,2	16,58
Lwów	373,2	5,43
Madryt	2925,8	27,02
Moskwa	1247,0	15,98
Neapol	1992,5	19,28
Paryż	1626,6	15,83
Praga	630,3	7,93
Rzym	1788,0	17,63
Wiedeń	682,2	8,12
Wrocław	344,6	4,40
Zagrzeb	1030,7	11,57



Rys. 3.1. Zależność czasu podróży od odległości

Krok III. Wybór klasy modelu

Naszym celem jest wyznaczenie czasu jazdy jako funkcji odległości, zatem za zmienną objaśnianą przyjmiemy czas, a za zmienną objaśniającą odległość. Podany wykres (rys. 3.1) wskazuje na liniowy kształt badanej zależności. Będziemy zatem wyznaczać zależność liniową postaci: $czas = \alpha_0 + \alpha_1 droga + \varepsilon$.

Krok IV. Estymacja parametrów strukturalnych

Wyniki estymacji modelu liniowego $czas = \alpha_0 + \alpha_1 droga + \varepsilon$ zależności czasu jazdy od odległości przedstawiono w postaci często spotykanej w programach statystycznych lub arkuszach kalkulacyjnych:

Statystyki regresji	
Wielokrotność *R*	0,986784
R kwadrat	0,973743
Dopasowany *R* kwadrat	0,972102
Błąd standardowy	1,319274
Obserwacje	18

ANALIZA WARIANCJI

	df	*SS*	*MS*	*F*	*Istotność F*
Regresja	1	1032,72	1032,72	593,3524	4,49E-14
Resztkowy	16	27,84773	1,740483		
Razem	17	1060,568			

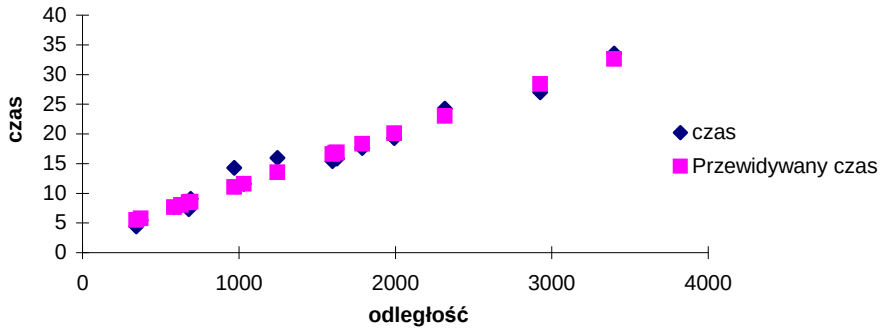
	Współczynniki	*Błąd standardowy*	*Statystyka t Studenta*	*Wartość p*	*Dolne* 95%	*Górne* 95,0%
Przecięcie	2,426929	0,585748	4,143296	0,000764	1,185198	3,66866
Odległość	0,008885	0,000365	24,35883	4,49E-14	0,008111	0,009658

Opisy na wydrukach: wielokrotność *R* – współczynnik korelacji wielorakiej,
R kwadrat – współczynnik determinacji,
Dopasowany *R* kwadrat – skorygowany współczynnik determinacji,
Błąd standardowy – dyspersja składnika losowego modelu,
Obserwacje – liczba obserwacji,
Regresja – regresja jako źródło zmienności,
Resztkowy – składnik losowy jako źródło zmienności,
Razem – zmienność całkowita,
Przecięcie – stała modelu.

Równanie regresji przyjmuje zatem postać:

$$\hat{czas} = 2{,}426929 + 0{,}008885\, droga\,.$$

Na rysunku 3.2 widzimy zaś, że różnice pomiędzy czasem przewidywanym a rzeczywistym nie wydają się zbyt duże. W następnym kroku postępowania pokażemy, że tak jest istotnie.



Rys. 3.2. Równanie regresji czasu podróży od odległości

Krok V. Weryfikacja modelu

Zbudowany model ekonometryczny $\hat{czas} = 2{,}426929 + 0{,}008885\, droga$ zweryfikujemy na poziomie istotności 0,05.

Dopasowanie modelu do danych empirycznych. Współczynnik determinacji modelu wynosi R^2 = 0,973743 (współczynnik zbieżności φ^2 = 2,6%).

Wniosek. Model wyjaśnia 97,4% zmienności badanej cechy. Świadczy to o dobrym dopasowaniu modelu do danych empirycznych.

Istotność układu współczynników regresji. Stawiamy hipotezy (test 1):

$$H_0 : \sum_{j=0}^{n} \alpha_j^2 = 0\,,$$

$$H_1 : \sum_{j=0}^{n} \alpha_j^2 \neq 0\,.$$

Sprawdzianem zespołu hipotez jest statystyka

$$F = \frac{R^2}{1-R^2}\frac{n-k-1}{k}$$

Statystyka ta, przy prawdziwości hipotezy zerowej, ma rozkład *F* Snedecora o 1 stopniu swobody licznika i 16 stopniach swobody mianownika.

Wyznaczona wartość empiryczna statystyki wynosi *F* = 593,3524, a odpowiadający jej krytyczny poziom istotności (istotność *F*) wynosi 4,49E-14 jest mniejszy od przyjętego poziomu istotności $\alpha = 0{,}05$. Odrzucamy zatem hipotezę H_0 na korzyść H_1.

Wniosek. Nie ma podstaw do odrzucenia hipotezy o zależności czasu podróży od odległości.

Istotność poszczególnych współczynników regresji: Dla każdego współczynnika modelu regresji (j = 0,1) stawiamy hipotezy (test 2):

H_0: $\alpha_j = 0$,

H_1: $\alpha_j \neq 0$.

Sprawdzianem zespołu hipotez jest statystyka:

$$t(a_j) = \frac{a_j}{S(a_j)}.$$

Statystyka ta, przy prawdziwości hipotez zerowych, ma rozkład *t* Studenta o 16 stopniach swobody.

Wyznaczone empiryczne wartości statystyk *t* Studenta wynoszą odpowiednio:

$t(\alpha_0) = 4{,}14$,

$t(\alpha_1) = 24{,}36$.

Odpowiadające im wartości krytycznego poziomu istotności (*wartość-p*)[8] 0,000764 oraz 4,491E-14 są mniejsze od przyjętego poziomu istotności $\alpha = 0{,}05$.

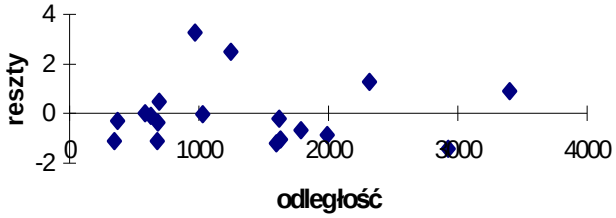
Wniosek. Nie ma podstaw do odrzucenia hipotezy o istotności obu współczynników modelu.

Analiza składników losowych modelu. Reszty modelu ekonometrycznego (rys. 3.3) uporządkowano według rosnącej wartości odległości.

Obserwacja	*Przewidywany czas*	*Składniki resztowe*	*Std. składniki resztowe*
17	5,488561	–1,08856	–0,85052
9	5,74266	–0,30933	–0,24168
2	7,631525	0,001808	0,001413
14	8,02689	–0,09356	–0,0731
3	8,45957	–1,10957	–0,86693
16	8,488	–0,37133	–0,29013
4	8,570627	0,479373	0,374544
6	11,03433	3,265675	2,55154
18	11,58428	–0,01762	–0,01376

[8] W wielu programach statystycznych wartość ta jest zwana *p-value*.

11	13,50602	2,477313	1,935577
5	16,6254	–1,20874	–0,94441
8	16,7951	–0,21176	–0,16546
13	16,87861	–1,04528	–0,8167
15	18,31259	–0,67925	–0,53071
12	20,12949	–0,84615	–0,66112
1	23,01343	1,269907	0,992205
10	28,42148	–1,40481	–1,09761
7	32,62478	0,891884	0,696848



Rys. 3.3. Rozkład reszt modelu liniowego czasu podróży od odległości

NORMALNOŚĆ

Stawiamy hipotezę H_0 składniki losowe mają rozkład $N(0; 1{,}319274)$. Zweryfikujemy ją za pomocą testu Dawida–Hellwiga (test 6).

Cele w tym przypadku to 18 odcinków o długości 1/18 pokazane w tabeli 3.2.

Tabela 3.2. Cele

Nr celi	Początek	Koniec
1	0,000	0,056
2	0,056	0,111
3	0,111	0,167
4	0,167	0,222
5	0,222	0,278
6	0,278	0,333
7	0,333	0,389
8	0,389	0,444
9	0,444	0,500
10	0,500	0,556
11	0,556	0,611
12	0,611	0,667
13	0,667	0,722
14	0,722	0,778
15	0,778	0,833
16	0,833	0,889
17	0,889	0,944
18	0,944	1,000

Reszty modelu, standaryzowane reszty, wartość dystrybuanty oraz nr celi, do której „wpada” dystrybuanta przedstawiono w tabeli 3.3.

Tabela 3.3. Reszty i dystrybuanta reszty modelu

Składniki resztowe	Std. składniki resztowe	Dystrybuanta	Cela
–1,404813223	–1,097610197	0,136187409	3
–1,208735081	–0,94441021	0,172480017	4
–1,109569542	–0,866930083	0,192990114	4
–1,088560631	–0,850515378	0,197519256	4
–1,045279401	–0,81669884	0,207050217	4
–0,846153087	–0,661117251	0,254268492	5
–0,6792532	–0,53071485	0,297808175	6
–0,371333583	–0,290130759	0,385858166	7
–0,309326744	–0,241683509	0,404512753	8
–0,211764199	–0,165455834	0,434292599	8
–0,093556382	–0,073097574	0,470864167	9
–0,017615297	–0,013763203	0,49450942	9
0,001808141	0,001412739	0,500563604	10
0,479373008	0,37454424	0,646000221	12
0,891884398	0,696848089	0,75705114	14
1,269907044	0,992205154	0,839451268	16
2,47731325	1,935577086	0,97354031	18
3,26567453	2,55154038	0,99463758	18

Puste cele to cele o numerach: 1, 2, 11, 13, 15, 17. Liczba pustych cel $K = 6$. Krytyczne liczby pustych cel dla 18 obserwacji dla przyjętego poziomu istotności $\alpha = 0{,}05$ wynoszą $K_1 = 3$ oraz $K_2 = 9$. Nie ma zatem podstaw do odrzucenia hipotezy H_0.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że składniki losowe mają rozkład normalny $N(0; 1{,}319274)$.

AUTOKORELACJA

Zbadamy, czy wraz ze wzrostem długości trasy występuje autokorelacja składników losowych rzędu pierwszego. W tym celu sortujemy dane niemalejąco względem odległości poszczególnych miejscowości od Warszawy. Następnie stawiamy hipotezy (test 7):

$H_0 : \rho_1 = 0$,

$H_1 : \rho_1 < 0$,

gdzie ρ_1 – współczynnik autokorelacji składników losowych rzędu pierwszego.

Wyznaczamy empiryczną wartość statystyki Durbina–Watsona

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

Empiryczna wartość statystyki d = 2,15911. Wartości krytyczne d_L = 4 – 1,39 = 2,61 oraz d_U = 4 – 1,16 = 2,84. Nie ma zatem podstaw do odrzucenia hipotezy $H_0 : \rho_1 = 0$.

Wniosek. Nie ma podstaw do odrzucenia hipotezy o braku autokorelacji składników losowych rzędu pierwszego.

SYMETRIA

Do sprawdzenia symetrii składnika losowego zastosujemy test 12.
Stawiamy hipotezy:

$$H_0 : p_+ = \frac{1}{2},$$

$$H_1 : p_+ \neq \frac{1}{2}.$$

Sprawdzianem zespołu hipotez jest statystyka:

$$t = \frac{\frac{m}{n} - \frac{1}{2}}{\sqrt{\frac{\frac{m}{n}\left(1 - \frac{m}{n}\right)}{n-1}}}$$

Statystyka ta, przy prawdziwości hipotezy H_0, ma rozkład t Studenta o 17 stopniach swobody. Empiryczna wartość statystyki wynosi –1,45774. Wartość krytyczna 2,11. Nie ma zatem podstaw do odrzucenia hipotezy H_0.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że rozkład składników losowych jest symetryczny.

LOSOWOŚĆ

Stawiamy hipotezę: H_0: reszty modelu są losowe. Zweryfikujemy ją testem liczby serii (test 13), zliczamy liczbę serii L tych samych znaków reszt w modelu. Porządkujemy reszty względem rosnących wartości długości tras i zliczamy liczbę serii, która w tym przypadku wynosi L = 10.

Krytyczne wartości liczby serii dla 6 reszt dodatnich i 12 reszt ujemnych, na przyjętym poziomie istotności α = 0,05 wynoszą 4 i 12. Empiryczna wartość statystyki nie wpada w obszar krytyczny – $4 < L = 10 < 12$.

Wniosek. Nie ma podstaw do odrzucenia hipotezy o losowości reszt modelu.

HOMOSKEDASTYCZNOŚĆ

Stałość wariancji składnika losowego zbadamy testem Spearmana (test 17). Testem tym można sprawdzić, czy wariancja składników losowych rośnie (maleje) wraz ze wzrostem wartości zmiennej objaśniającej *x*.

Stawiamy hipotezy:

$H_0 : \rho(|\varepsilon_x|, x) = 0$,

$H_1 : \rho(|\varepsilon_x|, x) \neq 0$.

Sprawdzianem zespołu hipotez jest statystyka

$$r = r\left(|\varepsilon|, x\right) = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n\left(n^2 - 1\right)},$$

gdzie D_i – różnica rang zmiennej i modułu reszty *e* dla *i*-tej obserwacji.

Tabela 3.4. Obliczenia do testu korelacji rang Spearmana

Miejscowość docelowa	Odległość *x*	Ranga *x*	Składniki resztowe	Moduł *e*	Ranga $\lvert e \rvert$	*D*	D^2
Wrocław	344,60	1	–1,08856	1,08856	12	–11	121
Lwów	373,20	2	–0,30933	0,30933	5	–3	9
Berlin	585,80	3	0,00181	0,00181	1	2	4
Praga	630,30	4	–0,09356	0,09356	3	1	1
Bratysława	679,00	5	–1,10957	1,10957	13	–8	64
Wiedeń	682,20	6	–0,37133	0,37133	6	0	0
Budapeszt	691,50	7	0,47937	0,47937	7	0	0
Helsinki	968,80	8	3,26567	3,26567	18	–10	100
Zagrzeb	1030,70	9	–0,01762	0,01762	2	7	49
Moskwa	1247,00	10	2,47731	2,47731	17	–7	49
Genewa	1598,10	11	–1,20874	1,20874	14	–3	9
Londyn	1617,20	12	–0,21176	0,21176	4	8	64
Paryż	1626,60	13	–1,04528	1,04528	11	2	4
Rzym	1788,00	14	–0,67925	0,67925	8	6	36
Neapol	1992,50	15	–0,84615	0,84615	9	6	36
Ateny	2317,10	16	1,26991	1,26991	15	1	1
Madryt	2925,80	17	–1,40481	1,40481	16	1	1
Lizbona	3398,90	18	0,89188	0,89188	10	8	64
						SUMA	**612**

Rangi (1, 2, ..., *n*) przypisujemy kolejno wartościom zmiennej *X* (reszt *e*) uporządkowanym w ciąg niemalejący. Jeżeli wystąpią takie same wartości zmiennej *X* (reszt *e*), to przypisujemy im rangę równą średniej arytmetycznej odpowiadających im pozycji w ciągu.

Na podstawie obliczeń z tabeli 3.4 wyznaczamy wartość empiryczną statystyki równą $r = 0{,}1$. Obszar krytyczny testu jest dwustronny. Na poziomie istotności $\alpha = 0{,}05$ wartość krytyczna statystyki Spearmana wynosi 0,399.

Wniosek. Nie ma podstaw do odrzucenia hipotezy H_0 o homoskedastyczności składników losowych.

Podsumowanie. Możemy zatem uznać model ekonometryczny

$$\hat{czas} = 2{,}426929 + 0{,}008885\, droga$$

za poprawny.

Krok VI. Wnioskowanie na podstawie modelu

Spróbujmy teraz na podstawie modelu wyznaczyć czas przejazdu samochodem z Warszawy do Amsterdamu, Brukseli i Pragi.

Tabela 3.5. Wartości prognoz punktowych i przedziałowych

Miejscowość docelowa	Odległość km	Czas h	Predykcja czasu h	Przedział ufności h	Względny błąd prognozy, %
Amsterdam	1204,2	12,72	13,13	10,25–16,00	3,2
Bruksela	1309,5	13,65	14,06	11,19–16,94	3,0
Praga	630,2	7,93	8,03	5,10–10,95	1,2

Maksymalny błąd względny prognozy wynosi 3,2%. Czas podróży do każdej miejscowości mieści się w odpowiednich przedziałach ufności.

Zauważmy, że poprawność modelu umożliwia wyciągnięcie różnego rodzaju wniosków praktycznych. Ustanowienie 95% przedziału ufności dla danej trasy przejazdu umożliwia 95% pewną kontrolę przejazdu. Czas przejazdu kierowcy, który okaże się poza tym przedziałem, powinien skłonić nas do szczegółowego przyjrzenia się temu: jeśli jest za krótki, to kierowca w sposób ewidentny łamał przepisy narażając siebie, pojazd i ładunek na niepotrzebne ryzyko; jeśli zaś jest zbyt długi, może wskazywać na jakieś nieprawidłowości w pracy kierowcy. Oczywiście, wyjaśnienia złożone przez kierowcę zbyt długo jadącego mogą być wiarygodne i w pełni akceptowalne, jednak złożenie ich jest konieczne.

3.2. Wzrost dzieci

Tabele norm wzrostu dzieci podają przedziałową normę wzrostu dla danej grupy wiekowej w zależności od płci. Skonstruowany liniowy model regresji opisujący zależność wzrostu dzieci od wieku i płci potwierdza poprawność norm.

Krok I. Określenie celu badań modelowych

Celem prowadzonych badań jest sprawdzenie, czy wzrost dziewczynek i chłopców zależy od płci i wieku dzieci. Literatura o rozwoju i żywieniu dzieci podaje normy wzrostu odrębne dla chłopców i dziewczynek. Przykład takich norm podano w tabeli 3.6.

Tabela 3.6. Normy wieku i wzrostu dla dzieci

Chłopcy		Dziewczynki	
Wiek miesiąc	Wzrost cm	Wiek miesiąc	Wzrost cm
15	76,5–82,1	15	75,2–81,5
18	79,1–84,9	18	78,5–84,1
21	81,7–87,7	21	80,4–86,0
24	84,0–90,1	24	81,4–87,3
27	85,3–93,3	27	84,1–92,5
30	88,1–94,3	30	86,5–92,5
33	89,6–96,4	33	87,2–96,0
36	91,6–99,0	36	89,9–97,3

Źródło: *Małe dziecko* (praca zbiorowa).

Chcemy te normy zweryfikować.

Krok II. Specyfikacja zmiennych wraz z gromadzeniem danych

Zbadano wzrost 8 dziewczynek i 8 chłopców w wieku od 15 do 36 miesięcy. Zgromadzone dane przedstawiono w tabeli 3.7.

Tabela 3.7. Wzrost dzieci

Chłopcy		Dziewczynki	
Wiek miesiąc	Wzrost cm	Wiek miesiąc	Wzrost cm
15	79	15	75
18	80	18	79
21	84	21	84
24	85	24	84
27	90	27	92
30	94	30	88
33	93	33	86
36	99	36	90

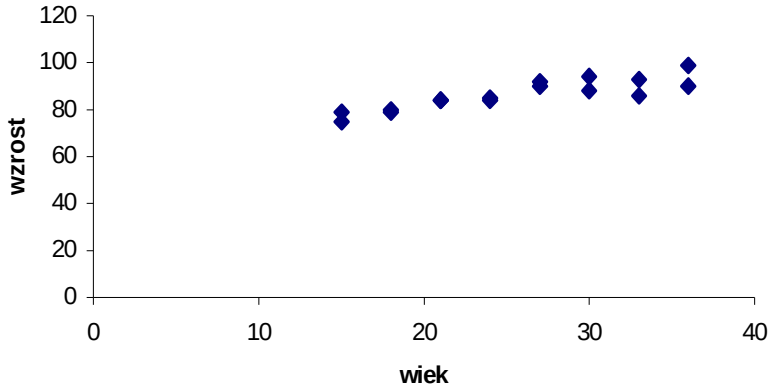
Opracowania własne

Krok III. Wybór klasy modelu

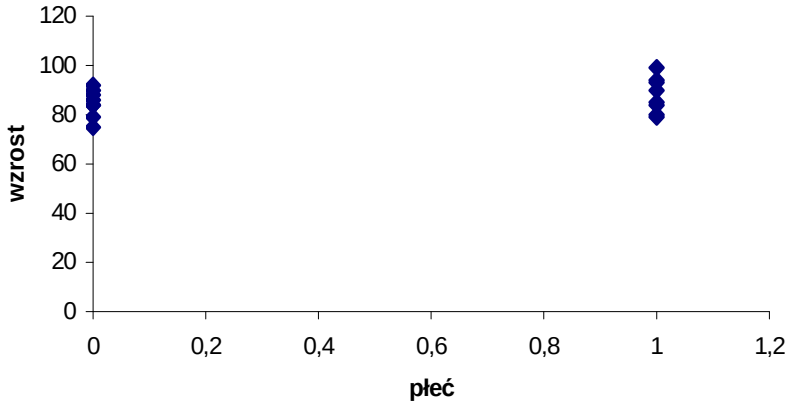
Skonstruujemy liniowy model ekonometryczny z dwiema zmiennymi objaśniającymi: zmienną ilościową x_1, opisującą wzrost oraz zmienną jakościową x_2, opisującą płeć (0 = dziewczynka, 1 = chłopiec) (rys. 3.4 i 3.5). Model przyjmie postać $y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \varepsilon$.

Tabela 3.8. Wartości zmiennej objaśnianej i zmiennych objaśniających

Wzrost [cm] y	Wiek [miesiąc] x_1	Płeć x_2
75	15	0
79	18	0
84	21	0
84	24	0
92	27	0
88	30	0
86	33	0
90	36	0
79	15	1
80	18	1
84	21	1
85	24	1
90	27	1
94	30	1
93	33	1
99	36	1



Rys. 3.4. Zależność wzrostu dzieci od wieku



Rys. 3.5. Zależność wzrostu dzieci od płci

Krok IV. Estymacja parametrów strukturalnych

Wyniki estymacji współczynników modelu liniowego są następujące:

Statystyki regresji	
Wielokrotność R	0,914621
R kwadrat	0,836532
Dopasowany R kwadrat	0,811383
Błąd standardowy	2,791602
Obserwacje	16

ANALIZA WARIANCJI

	df	*SS*	*MS*	*F*	*Istotność F*
Regresja	2	518,4405	259,2202	33,26304	7,71E-06
Resztkowy	13	101,3095	7,79304		

Razem	15	619,75

	Współczynniki	*Błąd standardowy*	*t Stat*	*Wartość-p*	*Dolne 95%*	*Górne 95%*
Przecięcie	64,5119	2,770753	23,28317	5,54E-12	58,52606	70,49775
Wiek	0,793651	0,10153	7,816942	2,88E-06	0,57431	1,012992
Płeć	3,25	1,395801	2,328412	0,036671	0,234556	6,265444

Wyznaczony model ekonometryczny ma postać:

$$\hat{wzrost} = 0{,}793651 \cdot wiek + 3{,}25 \cdot płeć + 64{,}5119.$$

Krok V. Weryfikacja modelu

Zbudowany model ekonometryczny $\hat{wzrost} = 0{,}793651 \cdot wiek + 3{,}25 \cdot płeć + 64{,}5119$ zweryfikujemy na poziomie istotności 0,05.

Dopasowanie modelu do danych empirycznych. Współczynnik dopasowania modelu $R^2 = 0{,}811383$ (współczynnik zbieżności $\varphi^2 = 18{,}9\%$).

Wniosek. Model wyjaśnia 81,1% zmienności badanej cechy.

Istotność układu współczynników regresji. Stawiamy hipotezę, że wzrost dzieci nie zależy ani od wieku, ani płci, wobec hipotezy alternatywnej o występowaniu przynajmniej jednej z tych zależności (test 1). Statystyka testowa, przy prawdziwości hipotezy zerowej, ma rozkład *F* Snedecora o 2 stopniach swobody licznika i 13 stopniach swobody mianownika.

Wartość empiryczna statystyki wynosi $F = 3326{,}304$, a odpowiadający jej krytyczny poziom istotności (istotność *F*) wynosi 7,71E-6 i jest mniejszy od przyjętego poziomu istotności $\alpha = 0{,}05$. Odrzucamy zatem hipotezę o nieistotności układu współczynników.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że wzrost dzieci zależy przynajmniej od jednej z cech: wiek, płeć.

Istotność poszczególnych współczynników regresji. Zweryfikujemy istotność każdego z trzech współczynników równania regresji (test 2). Przy prawdziwości hipotezy zerowej statystyka testowa ma rozkład *t* Studenta o 13 stopniach swobody. Empiryczne wartości statystyki wynoszą:

$t(\alpha_0) = 23{,}28317$,

$t(\alpha_1) = 7{,}816942$,

$t(\alpha_2) = 2{,}328412$.

Odpowiadające im wartości krytycznego poziomu istotności (*wartość-p*): 5,54E-12; 2,88E-06 oraz 0,036671 są mniejsze od przyjętego poziomu istotności $\alpha = 0{,}05$.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że wzrost dzieci zależy istotnie zarówno od wieku, jak i od płci.

Analiza składników losowych modelu. Reszty modelu ekonometrycznego uporządkowane według rosnącej wartości wieku dzieci oraz płci przedstawia tabela 3.9.

Tabela 3.9. Wartości reszt modelu i dystrybuanta

Obserwacja	*Przewidywane Y*	*Składniki resztowe*	*Std. składniki resztowe*
1	76,41667	–1,41667	–0,54512
9	79,66667	–0,66667	–0,25652
2	78,79762	0,202381	0,077874
10	82,04762	–2,04762	–0,7879
3	81,17857	2,821429	1,085649
11	84,42857	–0,42857	–0,16491
4	83,55952	0,440476	0,16949
12	86,80952	–1,80952	–0,69628
5	85,94048	6,059524	2,331627
13	89,19048	0,809524	0,311494
6	88,32143	–0,32143	–0,12368
14	91,57143	2,428571	0,934483
7	90,70238	–4,70238	–1,80942
15	93,95238	–0,95238	–0,36646
8	93,08333	–3,08333	–1,18643
16	96,33333	2,666667	1,026099

NORMALNOŚĆ

Stawiamy hipotezę H_0: składniki losowe mają rozkład $N(0; 2{,}791602)$. Zweryfikujemy ją testem Dawida–Hellwiga (test 6). Cele, w tym przypadku, to 16 odcinków o długości 1/16.

Tabela 3.10. Cele użyte w teście Hellwiga

Nr celi	Początek	Koniec
1	0,000	0,063
2	0,063	0,125
3	0,125	0,188
4	0,188	0,250
5	0,250	0,313
6	0,313	0,375
7	0,375	0,438
8	0,438	0,500

cd. tabeli 3.10

9	0,500	0,563
10	0,563	0,625
11	0,625	0,688
12	0,688	0,750
13	0,750	0,813
14	0,813	0,875
15	0,875	0,938
16	0,938	1,000

Reszty modelu, standaryzowane reszty, wartość dystrybuanty oraz nr celi, do której wpada dystrybuanta przedstawiono w tabeli 3.11.

Tabela 3.11. Wartości reszt modelu i dystrybuanta

Składniki resztowe	*Std. składniki resztowe*	*Dystrybuanta*	*Cela*
–1,416666667	–0,545115044	0,292837169	1
–4,702380952	–1,809415482	0,035193185	2
–3,083333333	–1,186426861	0,117726944	4
–2,047619048	–0,787897375	0,215378301	4
–1,80952381	–0,696281401	0,243126236	5
–0,952380952	–0,366463895	0,357009532	5
–0,666666667	–0,256524727	0,3987729	7
–0,428571429	–0,164908753	0,434507895	7
–0,321428571	–0,123681565	0,450783663	8
0,202380952	0,077873578	0,53103576	9
0,44047619	0,169489552	0,567294207	10
0,80952381	0,311494311	0,622287501	10
2,428571429	0,934482933	0,824972602	14
2,666666667	1,026098907	0,847577501	14
2,821428571	1,085649289	0,861182872	14
6,05952381	2,331626533	0,990139849	16

Puste cele to cele o numerach: 3, 6, 11, 12, 13, 15. Liczba pustych cel $K = 6$. Krytyczne liczby pustych cel dla 16 obserwacji, dla przyjętego poziomu istotności $\alpha = 0{,}05$, wynoszą $K_1 = 3$ oraz $K_2 = 8$. Nie ma zatem podstaw do odrzucenia hipotezy H_0.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że składniki losowe mają rozkład normalny $N(0; 2{,}791602)$.

AUTOKORELACJA

Hipotezę o braku autokorelacji rzędu pierwszego, wobec hipotezy alternatywnej o istnieniu autokorelacji dodatniej, zweryfikujemy testem Durbina–Watsona (test 7). Empiryczna wartość wynosi $d = 1{,}64222$. Wartości krytyczne $d_L = 1{,}10$ oraz $d_U = 1{,}37$. Nie ma zatem podstaw do odrzucenia hipotezy H_0: $\rho_1 = 0$.

Wniosek. Nie ma podstaw do odrzucenia hipotezy o braku autokorelacji składników losowych rzędu pierwszego.

SYMETRIA

Stawiamy hipotezę H_0 o symetrii składników losowych (test 12). Statystyka testowa ma rozkład t Studenta o 15 stopniach swobody. Empiryczna wartość statystyki t dla 7 reszt dodatnich wynosi $t = -0{,}48795$. Wartość krytyczna 2,131. Nie ma zatem podstaw do odrzucenia hipotezy H_0.

Wniosek. Składniki losowe są symetryczne.

LOSOWOŚĆ

Stawiamy hipotezę zerową H_0: reszty modelu są losowe.

Zweryfikujemy tę hipotezę testem liczby serii (test 13), zliczamy liczbę serii L tych samych znaków reszt w modelu, która w tym przypadku wynosi $L = 12$.

Krytyczne wartości liczby serii dla 7 reszt dodatnich i 9 reszt ujemnych, na przyjętym poziomie istotności $\alpha = 0{,}05$, wynoszą 4 i 13. Empiryczna wartość statystyki nie wpada w obszar krytyczny $-4 < L = 12 < 13$. Nie ma zatem podstaw do odrzucenia hipotezy H_0.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że rozkład składników losowych jest symetryczny.

HOMOSKEDASTYCZNOŚĆ

Równość wariancji w podpróbach homogenicznych ze względu na wariancję składników losowych można przeprowadzić testem Goldfelda–Quandta (test 15).

W tym celu zbudujemy dwa modele ekonometryczne (patrz wydruki):

Pierwszy model dla dziewczynek:

$$\hat{wzrost}_d = 68{,}55952 + 0{,}634921 \cdot wiek$$

Statystyki regresji	
Wielokrotność R	0,829428
R kwadrat	0,68795
Dopasowany R kwadrat	0,635942
Błąd standardowy	3,394089
Obserwacje	8

ANALIZA WARIANCJI

	df	*SS*	*MS*	*F*	*Istotność F*
Regresja	1	152,381	152,381	13,2277	0,010874
Resztkowy	6	69,11905	11,51984		
Razem	7	221,5			

	Współczynniki	*Błąd standardowy*	*t Stat*	*Wartość-p*	*Dolne 95%*	*Górne 95%*
Przecięcie	68,55952	4,610515	14,87025	5,82E-06	57,27799	79,84106
Wiek	0,634921	0,174573	3,63699	0,010874	0,207755	1,062086

Drugi model ekonometryczny dla chłopców:

$$\hat{wzrost}_{ch} = 6371429 + 0{,}952381 \cdot wiek\,.$$

Statystyki regresji	
Wielokrotność R	0,981367
R kwadrat	0,963082
Dopasowany R kwadrat	0,956929
Błąd standardowy	1,480026
Obserwacje	8

ANALIZA WARIANCJI

	df	*SS*	*MS*	*F*	*Istotność F*
Regresja	1	342,8571	342,8571	156,5217	1,59E-05
Resztkowy	6	13,14286	2,190476		
Razem	7	356			

	Współczynniki	*Błąd standardowy*	*t Stat*	*Wartość-p*	*Dolne 95%*	*Górne 95%*
Przecięcie	63,71429	2,01046	31,69139	6,56E-08	58,79486	68,63371
Wiek	0,952381	0,076124	12,51086	1,59E-05	0,766111	1,138651

Stawiamy następnie hipotezy:

$H_0 : \delta_{e_1}^2 = \delta_{e_2}^2$,

$H_1 : \delta_{e_1}^2 > \delta_{e_2}^2$.

Zespół hipotez weryfikujemy statystyką *F*, która, przy prawdziwości hipotezy zerowej, ma rozkład *F* Snedecora o 6 stopniach swobody licznika i o 6 stopniach swobody mianownika. Wyznaczona z próby wartość statystyki $F = 0{,}48795$, podczas gdy wartość krytyczna dla przyjętego poziomu istotności $\alpha = 0{,}05$ wynosi $F_\alpha = 3{,}79$.

Wniosek. Nie ma podstaw do odrzucenia hipotezy o równości wariancji składników losowych w obu podpróbach (dziewczynek i chłopców).

Podsumowanie. Możemy uznać model ekonometryczny

$$\hat{wzrost} = 0{,}793651 \cdot wiek + 3{,}25 \cdot płeć + 64{,}5119$$

za poprawny.

Krok VI. Wnioskowanie na podstawie modelu

Przeprowadzona weryfikacja świadczy o poprawności modelu:

$$\hat{wzrost} = 0{,}793651 \cdot wiek + 3{,}25 \cdot płeć + 64{,}5119.$$

Możemy zatem stwierdzić, że w badanej grupie wiekowej (15–36 miesięcy) wzrost dzieci jest proporcjonalny do wieku (dziecko rośnie średnio 0,793651 cm w ciągu miesiąca), przy czym w danej grupie wiekowej chłopcy są średnio wyżsi od dziewczynek o 3,25 cm.

Określimy teraz przedziały ufności dla wzrostu dziecka. Przyjmiemy poziom ufności równy 0,95.

Tabela 3.12. Przedział ufności dla wzrostu

Chłopcy			Dziewczynki		
Wiek miesiąc	Wzrost cm		Wiek miesiąc	Wzrost cm	
15	72,87	86,47	15	69,62	83,22
18	75,44	88,65	18	72,19	85,40
21	77,96	90,90	21	74,71	87,65
24	80,40	93,21	24	77,15	89,96
27	82,79	95,60	27	79,54	92,35
30	85,10	98,04	30	81,85	94,79
33	87,35	100,56	33	84,10	97,31
36	89,53	103,13	36	86,28	99,88

Z porównania przedziałów ufności z normami otrzymamy względne błędy podane w tabeli 3.13.

Tabela 3.13. Błąd predykcji przedziałowych

Chłopcy			Dziewczynki		
Wiek miesiąc	Błąd %		Wiek miesiąc	Błąd %	
15	3,1	6,1	15	9,0	1,4
18	3,9	5,4	18	8,7	0,6
21	3,0	5,7	21	8,6	0,1
24	1,2	6,8	24	8,1	0,2
27	1,6	3,3	27	6,8	1,0
30	1,6	6,0	30	7,1	0,5
33	0,2	4,7	33	6,1	0,9
36	0,4	6,0	36	5,8	0,9

Są to błędy rzędu kilku procent, co potwierdza poprawność norm.

3.3. Ceny mieszkań

Model ekonometryczny zależności ceny mieszkań od metrażu należy do klasy modeli nieliniowych. Zastosowano go do predykcji przedziałowej ceny 52-metrowych mieszkań.

Krok I. Cel badań

Pośrednik biura nieruchomości *Twój Dom* z siedzibą we Wrocławiu codziennie przyjmuje oferty mieszkań do sprzedaży. Klienci pytają: za jaką cenę mogą wystawić swoje mieszkanie na sprzedaż. Naszym celem jest zbudowanie modelu ekonometrycznego opisującego zależność ceny mieszkań od metrażu, który pomoże odpowiedzieć na postawione pytanie.

Krok II. Specyfikacja zmiennych wraz z gromadzeniem danych

Biuro ma system komputerowy, w którym ewidencjonowane są aktualnie zgłoszone oferty. Niecodziennie jednak w ofercie dnia można znaleźć mieszkanie, o które pytają klienci. Kierownik biura polecił więc zgromadzić informacje o innych ofertach sprzedaży mieszkań we Wrocławiu (tabela 3.14 i rys. 3.6).

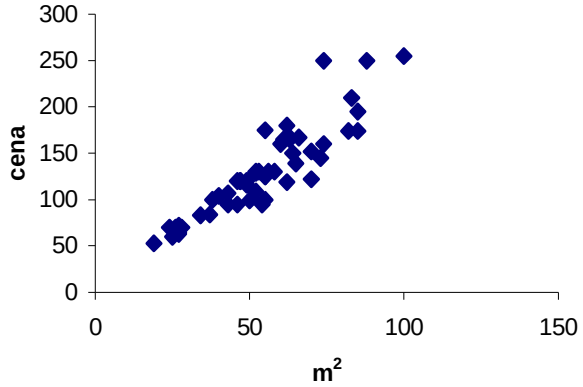
Tabela 3.14. Cena, metraż mieszkań i liczba pokoi

Metraż m^2	Cena tys. zł	Liczba pokoi	Metraż m^2	Cena tys. zł	Liczba pokoi
43	107	1	61	165	3
25	60	1	54	95	3
27	63	1	63	165	3
27	72	1	60	160	3
26	70	1	52	130	3
28	70	1	64	150	3
19	53	1	56	130	3
37	84	1	74	250	3
24	70	1	49	120	3
40	104	2	62	180	3
38	100	2	70	122	3
27	65	2	63	167	3
46	95	2	55	125	3
47	120	2	58	130	3
52	129	2	55	175	3
46	120	2	66	167	3
47	120	2	55	125	3

cd. tabeli 3.14

53	130	2	83	210	4
49	116	2	100	255	4
55	100	2	85	174	4
34	83	2	70	152	4
50	99	2	88	250	4
43	95	2	82	174	4
52	109	2	73	145	4
65	139	3	74	160	4
62	119	3	85	195	4

Źródło: Ogłoszenia w Gazecie Wyborczej



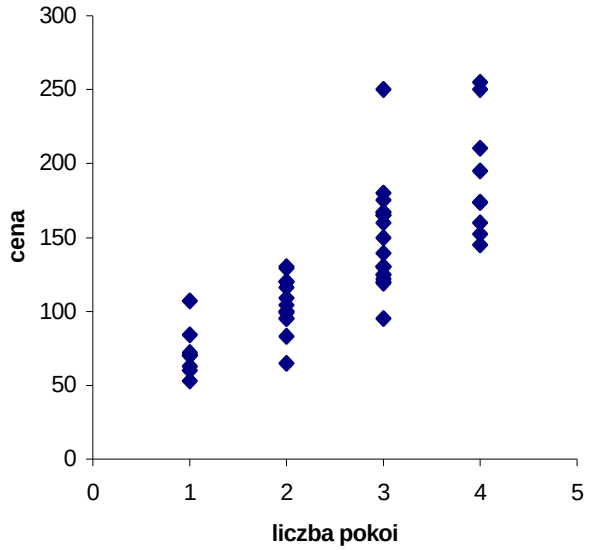
Rys. 3.6. Zależność ceny mieszkań od metrażu

Sporządzono wykres zależności ceny mieszkań od metrażu.

Krok III. Wybór klasy modelu

Cena mieszkań zwiększa się oczywiście wraz z metrażem. Z analizy wykresu zależności ceny mieszkania od metrażu wynika, że ceny mieszkań o małej powierzchni są mniej zróżnicowane niż ceny mieszkań o dużej powierzchni. Gdyby to przypuszczenie okazało się prawdziwe, oznaczałoby to, że model liniowy nie jest w tym przypadku modelem właściwym.

Aby sprawdzić to przypuszczenie, sporządzono wykres zależności cen mieszkań od liczby pokoi.



Rys. 3.7. Zależność ceny mieszkań od liczby pokoi

Z analizy wykresu na rys. 3.7 wysunięto przypuszczenie, że ceny mieszkań 1 i 2 pokojowych są mniej zróżnicowane niż ceny mieszkań 4 pokojowych, co w konsekwencji przenosi się na brak homoskedastyczności składników losowych modelu liniowego dla całej populacji.

Weryfikację tej hipotezy przeprowadzimy testem Goldfelda–Quandta (test 15). W teście tym wymagana jest jedynie znajomość postaci analitycznej modelu ekonometrycznego, nie jest natomiast konieczna znajomość parametrów strukturalnych modelu dla pełnego zbioru danych. Bazuje on na parametrach modeli ekonometrycznych podgrup podejrzanych o zróżnicowane wariancje. Zbudowano zatem dwa modele regresji liniowej zależności ceny mieszkań od metrażu (patrz wydruki):

Pierwszy model ekonometryczny dla mieszkań 1 i 2 pokojowych:

$$\hat{cena}_{1-2} = 17{,}98729 + 1{,}927599 \cdot metraż$$

Statystyki regresji	
Wielokrotność R	0,912808
R kwadrat	0,833218
Dopasowany R kwadrat	0,825637
Błąd standardowy	9,803596
Obserwacje	24

ANALIZA WARIANCJI

	df	*SS*	*MS*	*F*	*Istotność F*
Regresja	1	10563,4	10563,4	109,9089	5,08E-10
Resztkowy	22	2114,431	96,11049		

Razem	23	12677,83

	Współczynniki	*Błąd standardowy*	*t Stat*	*Wartość-p*	*Dolne 95%*	*Górne 95%*
Przecięcie	17,98729	7,437371	2,418502	0,024308	2,563113	33,41147
metraż	1,927599	0,183865	10,48375	5,08E-10	1,546285	2,308913

Drugi model ekonometryczny dla mieszkań 4-pokojowych:

$$\hat{cena}_4 = -131{,}764 + 3{,}920107 \cdot metraż$$

Statystyki regresji	
Wielokrotność R	0,883972
R kwadrat	0,781407
Dopasowany R kwadrat	0,750179
Błąd standardowy	20,24754
Obserwacje	9

ANALIZA WARIANCJI

	df	*SS*	*MS*	*F*	*Istotność F*
Regresja	1	10258,48	10258,48	25,02297	0,001561
Resztkowy	7	2869,739	409,9627		
Razem	8	13128,22			

	Współczynniki	*Błąd standardowy*	*t Stat*	*Wartość-p*	*Dolne 95%*	*Górne 95%*
Przecięcie	–131,764	64,78689	–2,03381	0,081453	–284,961	21,43223
metraż	3,920107	0,783661	5,002296	0,001561	2,067043	5,77317

Dla wyróżnionych podprób o liczebnościach odpowiednio $n_1 = 24$, $n_2 = 9$ stawiamy hipotezy:

$$H_0 : \delta_{e_1}^2 = \delta_{e_2}^2 ,$$

$$H_1 : \delta_{e_1}^2 < \delta_{e_2}^2 .$$

Zespół hipotez weryfikujemy statystyką o rozkładzie F Snedecora o 7 stopniach swobody licznika i o 22 stopniach swobody mianownika. Obliczona z próby wartość statystyki $F = 4{,}265536$, a wartość krytyczna wynosi $F_\alpha = 2{,}46$. Odrzucamy zatem hipotezę H_0 na korzyść H_1.

Wniosek. Odrzucamy hipotezę o równości wariancji składników losowych modeli liniowych w obu podpróbach (mieszkań 1- i 2-pokojowych oraz mieszkań 4 pokojowych).

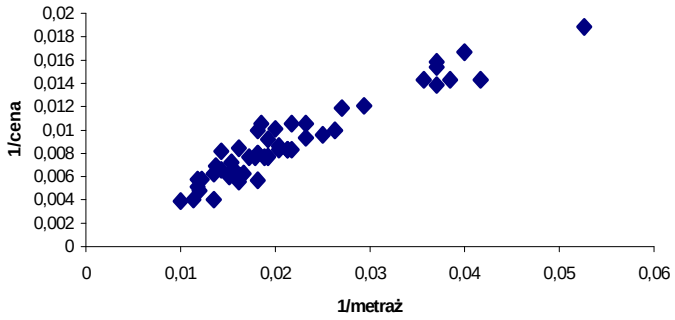
W celu wyrównania wariancji dokonamy transformacji danych, przyjmując za zmienną objaśnianą odwrotność ceny mieszkań:

$$y' = \frac{1}{cena}$$

a za zmienną objaśniającą odwrotność metrażu

$$x' = \frac{1}{metraż}$$

Z przedstawionego wykresu zależności odwrotności ceny mieszkań od metrażu (rys. 3.8) widać, że w nowej skali rozrzut obserwacji jest mniej zróżnicowany.



Rys. 3.8. Zależność odwrotności ceny mieszkań od odwrotności metrażu

Wydaje się zatem, że model

$$\frac{1}{cena} = \alpha_0 + \alpha_1 \cdot \frac{1}{metraż} + \varepsilon$$

będzie właściwym odwzorowaniem rzeczywistości.

Krok IV. Estymacja parametrów strukturalnych

Na podstawie przeskalowanych danych wyznaczamy wartości parametrów modelu:

$$\frac{1}{cena} = \alpha_0 + \alpha_1 \frac{1}{metraż} + \varepsilon .$$

Statystyki regresji	
Wielokrotność R	0,952546
R kwadrat	0,907345
Dopasowany R kwadrat	0,905492

Błąd standardowy	0,001067
Obserwacje	52

ANALIZA WARIANCJI

	df	*SS*	*MS*	*F*	*Istotność F*
Regresja	1	0,000557	0,000557	489,634	1,75E-27
Resztkowy	50	5,69E-05	1,14E-06		
Razem	51	0,000614			

	Współczynniki	*Błąd standardowy*	*t Stat*	*Wartość-p*	*Dolne 95%*	*Górne 95%*
Przecięcie	0,001218	0,000374	3,254715	0,002039	0,000466	0,00197
1/metraż	0,35788	0,016173	22,12768	1,75E-27	0,325395	0,390365

Model regresji przyjmuje zatem postać

$$\hat{\frac{1}{cena}} = 0{,}001218074 + 0{,}357888172\frac{1}{metraż}.$$

Jest to model nieliniowy. Wyznaczając interesującą nas zależność ceny mieszkań od metrażu, otrzymujemy model w postaci krzywej Tőrquista

$$\hat{cena} = \frac{metraż}{0{,}001218074 + 0{,}357888172 \cdot metraż}.$$

Krok V. Weryfikacja modelu

Zbudowany model ekonometryczny zweryfikujemy na poziomie istotności $\alpha = 0{,}05$.

Dopasowanie modelu do danych empirycznych. Współczynnik determinacji modelu wynosi $R^2 = 0{,}905492$ (współczynnik zbieżności $\varphi^2 = 9{,}5\%$). Model wyjaśnia 90,5% zmienności badanej cechy.

Dla modeli nieliniowych należy zbadać wskaźnik średniego względnego dopasowania modelu:

$$\Psi = \frac{1}{n}\sum_{t=1}^{n}\frac{|E_t|}{|\hat{y}_t|}$$

gdzie E_i – reszty modelu nieliniowego.

Wyznaczamy reszty modelu nieliniowego E_t (tab. 3.15).

W naszym modelu $\Psi = 10{,}4\%$.

Tabela 3.15. Predykcja ceny mieszkań i błąd modelu nieliniowego

Metraż m^2	Cena tys. zł	Liczba pokoi	Prognoza ceny tys. zł	Reszty modelu nieliniowego (E_i)	Metraż m^2	Cena tys. zł	Liczba pokoi	Prognoza ceny tys. zł	Reszty modelu nieliniowego (E_i)
43	107	1	104,8123	2,18774	61	165	3	141,144	23,85598
25	60	1	64,3779	–4,3779	54	95	3	127,4619	–32,4619
27	63	1	69,09468	–6,09468	63	165	3	144,9546	20,0454
27	72	1	69,09468	2,905324	60	160	3	139,2226	20,77743
26	70	1	66,74366	3,256337	52	130	3	123,4509	6,549076
28	70	1	71,43108	–1,43108	64	150	3	146,8439	3,156077
19	53	1	49,86568	3,134319	56	130	3	131,4269	–1,42692
37	84	1	91,82306	–7,82306	74	250	3	165,1721	84,82793
24	70	1	61,99725	8,00275	49	120	3	117,3468	2,653223
40	104	2	98,37603	5,623973	62	180	3	143,0547	36,94533
38	100	2	94,02055	5,979454	70	122	3	157,9617	–35,9617
27	65	2	69,09468	–4,09468	63	167	3	144,9546	22,0454
46	95	2	111,1348	–16,1348	55	125	3	129,4501	–4,45007
47	120	2	113,2176	6,782357	58	130	3	135,3469	–5,3469
52	129	2	123,4509	5,549076	55	175	3	129,4501	45,54993
46	120	2	111,1348	8,865151	66	167	3	150,5911	16,40895
47	120	2	113,2176	6,782357	55	125	3	129,4501	–4,45007
53	130	2	125,4622	4,537828	83	210	4	180,8356	29,16435
49	116	2	117,3468	–1,34678	100	255	4	208,469	46,53097
55	100	2	129,4501	–29,4501	85	174	4	184,2154	–10,2154
34	83	2	85,15014	–2,15014	70	152	4	157,9617	–5,96172
50	99	2	119,3933	–20,3933	88	250	4	189,2186	60,78145
43	95	2	104,8123	–9,81226	82	174	4	179,1323	–5,1323
52	109	2	123,4509	–14,4509	73	145	4	163,3842	–18,3842
65	139	3	148,7227	–9,72271	74	160	4	165,1721	–5,17207
62	119	3	143,0547	–24,0547	85	195	4	184,2154	10,78464

Wniosek. Świadczy to o dobrym dopasowaniu modelu do danych empirycznych.

Zbudowany model ekonometryczny zweryfikujemy na poziomie istotności $\alpha = 0{,}05$.

Istotność układu współczynników regresji. Stawiamy hipotezę H_0 o braku zależności liniowej odwrotności ceny mieszkań od odwrotności metrażu, wobec hipotezy alternatywnej, że zależność ta występuje (test 1). Zweryfikujemy ją statystyką $F = \frac{R^2}{1-R^2}\frac{n-k-1}{k}$, która przy prawdziwości hipotezy zerowej ma rozkład F Snedecora o 1 stopniu swobody licznika i 50 stopniach swobody mianownika.

Wartość empiryczna statystyki F = 489,634. Odpowiadający jej krytyczny poziom istotności (istotność F) wynosi 1,74E-27 i jest mniejszy od przyjętego poziomu istotności α = 0,05. Odrzucamy zatem hipotezę H_0 na korzyść H_1.

Wniosek. Nie ma podstaw do odrzucenia hipotezy o zależności odwrotności ceny mieszkań od odwrotności metrażu.

Istotność poszczególnych współczynników regresji. Dla każdego współczynnika modelu regresji (j = 0,1) stawiamy hipotezy (test 2)

H_0: $\alpha_j = 0$,

H_1: $\alpha_j \neq 0$.

Zespół hipotez weryfikujemy statystyką t Studenta o 50 stopniach swobody. Empiryczne wartości statystyk t Studenta wynoszą:

$t(\alpha_0)$ = 3,254715,

$t(\alpha_1)$ = 22,12768.

Odpowiadające im wartości krytycznego poziomu istotności (wartość-p) 0,002039 i 1,75E-27 są mniejsze od przyjętego poziomu istotności $\alpha = 0{,}05$.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że oba współczynniki modelu są istotnie różne od zera.

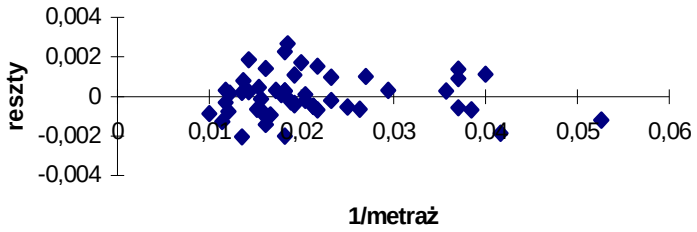
Analiza składników losowych modelu. Wartości reszt modelu regresji uporządkowane według rosnących wartości metrażu mieszkania przedstawiono w tabeli 3.16 i na rysunku 3.9.

Tabela 3.16. Reszty modelu uporządkowane względem metrażu mieszkań

Metraż	*Obserwacja*	*Przewidywane 1/cena*	*Składniki resztowe*	*Std. składniki resztowe*
19	7	0,020053872	–0,001185948	–1,122725686
24	9	0,016129748	–0,001844033	–1,745729093
25	2	0,015533281	0,001133386	1,072965797
26	5	0,014982696	–0,000696982	–0,659825908
27	3	0,014472895	0,001400121	1,325481219
27	4	0,014472895	–0,000584006	–0,552873143
27	12	0,014472895	0,000911720	0,863117068
28	6	0,013999509	0,000286206	0,270948250
34	21	0,011743961	0,000304231	0,288013091
37	8	0,010890511	0,001014251	0,960181842
38	11	0,010635973	–0,000635973	–0,602069787
40	10	0,010165078	–0,000549694	–0,520389721
43	1	0,009540869	–0,000195074	–0,184674890
43	23	0,009540869	0,000985447	0,932913618
46	13	0,008998078	0,001528238	1,446768715
46	16	0,008998078	–0,000664744	–0,629307159
47	14	0,008832546	–0,000499212	–0,472599576
47	17	0,008832546	–0,000499212	–0,472599576
49	19	0,008521751	9,89388E-05	0,093664434
49	35	0,008521751	–0,000188418	–0,178373095
50	22	0,008375677	0,001725333	1,633356329
52	15	0,008100385	–0,000348447	–0,329871394

cd. tabeli 3.16

52	24	0,008100385	0,001073927	1,016676644
52	31	0,008100385	–0,000408077	–0,386322831
53	18	0,00797053	–0,000278222	–0,263390395
54	28	0,007845484	0,002680831	2,537917809
55	20	0,007724986	0,002275014	2,153734277
55	39	0,007724986	0,000275014	0,260353080
55	41	0,007724986	–0,002010700	–1,903511144
55	43	0,007724986	0,000275014	0,260353080
56	33	0,007608791	8,35165E-05	0,079064248
58	40	0,007388422	0,000303886	0,287686042
60	30	0,007182743	–0,000932743	–0,883019416
61	27	0,007084962	–0,001024356	–0,969748081
62	26	0,006990335	0,001413027	1,337699026
62	36	0,006990335	–0,001434779	–1,358291940
63	29	0,006898712	–0,000838105	–0,793426565
63	38	0,006898712	–0,000910688	–0,862139365
64	32	0,006809952	–0,000143285	–0,135646464
65	25	0,006723923	0,000470322	0,445249348
66	42	0,006640501	–0,000652477	–0,617693626
70	37	0,006330648	0,001866074	1,766594285
70	47	0,006330648	0,000248300	0,235062903
73	50	0,006120542	0,000776010	0,734641123
74	34	0,006054292	–0,002054292	–1,944779318
74	51	0,006054292	0,000195708	0,185274529
82	49	0,005582466	0,000164660	0,155882280
83	44	0,005529883	–0,000767978	–0,727037947
85	46	0,005428429	0,000318698	0,301708016
85	52	0,005428429	–0,000300224	–0,284218967
88	48	0,005284894	–0,001284894	–1,216397083
100	45	0,004796876	–0,000875307	–0,828644880



Rys. 3.9. Reszty modelu liniowego odwrotności ceny od odwrotności metrażu

NORMALNOŚĆ

Hipotezę o normalności składników losowych modelu zweryfikujemy testem χ^2 (test 3).

Stawiamy hipotezę: H_0: składnik losowy ma rozkład $N(0; S_\varepsilon = 0{,}001067)$.

Zweryfikujemy ją statystyką:

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i},$$

gdzie: $r = 4$ – liczba klas,

n_i – liczba obserwacji w i-tej klasie,

p_i – prawdopodobieństwo hipotetyczne wartości błędu losowego w i-tej klasie.

Statystyka ta, przy prawdziwości hipotezy H_0, ma rozkład χ^2 o 2 stopniach swobody.

Tabela 3.17. Obliczenia statystyki χ^2

Klasa		n_i	$F(x)$	p_i	np_i	$\frac{(n_i - np_i)^2}{np_i}$
od	do					
(–∞)	–0,54903	16	0,290175	0,290175	15,08911	0,054989
–0,54903	0,254428	15	0,59292	0,302745	15,74273	0,035042
0,254428	1,05788	12	0,845346	0,252426	13,12617	0,096621
1,05788	(+∞)	7	0,994424	0,154654	8,04199	0,135009
					SUMA=	0,32166

Empiryczna wartość statystyki wynosi $\chi^2 = 0{,}32166$, a wartość krytyczna $\chi^2_\alpha = 5{,}991$. Nie ma zatem podstaw do odrzucenia hipotezy o normalności składników losowych.

Wniosek. Nie podstaw do odrzucenia hipotezy, że składniki losowe mają rozkład normalny $N(0; 0{,}001067)$.

AUTOKORELACJA

Stawiamy hipotezy (test 7):

$H_0 : \rho_1 = 0$,

$H_1 : \rho_1 < 0$,

gdzie ρ_1 – współczynnik autokorelacji składników losowych rzędu pierwszego.

Wyznaczamy empiryczną wartość statystyki Durbina–Watsona dla reszt modelu uporządkowanych względem rosnących wartości odwrotności metrażu mieszkań.

Empiryczna wartość statystyki wynosi $d = 2{,}13272$, $d' = 1{,}86728$. Wartości krytyczne $d_L = 1{,}50$ oraz $d_U = 1{,}59$. Nie ma zatem podstaw do odrzucenia hipotezy H_0: $\rho_1 = 0$.

Wniosek. Nie ma podstaw do odrzucenia hipotezy o braku autokorelacji składników losowych rzędu pierwszego.

SYMETRIA

Stawiamy hipotezę H_0 o jednakowej frakcji dodatnich i ujemnych błędów modelu (test 12). Weryfikujemy ją statystyką *t* Studenta o 51 stopniach swobody.

Empiryczna wartość statystyki wynosi –0,27487. Wartość krytyczna 2,01. Nie ma zatem podstaw do odrzucenia hipotezy H_0.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że rozkład składników losowych jest symetryczny.

LOSOWOŚĆ

Stawiamy hipotezę zerową H_0: Reszty modelu są losowe. Zweryfikujemy tę hipotezę testem serii (test 13), zliczając liczbę serii *L* tych samych znaków reszt w modelu. Empiryczna liczba serii wynosi L =25.

Wartości krytyczne testu serii dla 25 reszt dodatnich i 27 ujemnych, na przyjętym poziomie istotności $\alpha = 0{,}05$, aproksymujemy rozkładem normalnym N(26,96; 3,56), obliczając granice obszaru dopuszczalnego:

$-1{,}96 \cdot 3{,}56 + 26{,}96 = 21{,}08 \approx 21$,

$1{,}96 \cdot 3{,}36 + 26{,}96 = 33{,}96 \approx 34$.

Empiryczna wartość statystyki nie wpada w obszar krytyczny – $21 < L = 25 < 34$. Nie ma zatem podstaw do odrzucenia hipotezy zerowej.

Wniosek. Nie ma podstaw do odrzucenia hipotezy o losowości reszt modelu.

HOMOSKEDASTYCZNOŚĆ

Badanie równości wariancji składników losowych dla modelu liniowego przeprowadzimy testem Goldfelda–Quandta (test 15). W tym celu, podobnie jak poprzednio, zbudujemy 2 modele regresji liniowej zależności odwrotności ceny mieszkań od odwrotności metrażu.

Pierwszy model ekonometryczny dla mieszkań 1 i 2 pokojowych (patrz wydruki):

$$\frac{\hat{1}}{cena_{1-2}} = 0{,}002413 + 0{,}32248\frac{1}{metraż}$$

Statystyki regresji	
Wielokrotność R	0,951894
R kwadrat	0,906102
Dopasowany R kwadrat	0,901833
Błąd standardowy	0,000998
Obserwacje	24

ANALIZA WARIANCJI

	Df	*SS*	*MS*	*F*	*Istotność F*
Regresja	1	0,000212	0,000212	212,2956	8,8E-13
Resztkowy	22	2,19E-05	9,97E-07		
Razem	23	0,000234			

	Współczynniki	*Błąd standardowy*	*t Stat*	*Wartość-p*	*Dolne 95%*	*Górne 95%*
Przecięcie	0,002413	0,000656	3,680107	0,001312	0,001053	0,003773
1/metraż	0,32248	0,022133	14,57037	8,8E-13	0,27658	0,368381

Drugi model ekonometryczny dla mieszkań 4-pokojowych:

$$\frac{\hat{1}}{cena_4} = -0{,}00345 + 0{,}723554\frac{1}{metraż}$$

Statystyki regresji	
Wielokrotność R	0,898504
R kwadrat	0,80731
Dopasowany R kwadrat	0,779783
Błąd standardowy	0,000503
Obserwacje	9

ANALIZA WARIANCJI

	df	*SS*	*MS*	*F*	*Istotność F*
Regresja	1	7,43E-06	7,43E-06	29,32779	0,000992
Resztkowy	7	1,77E-06	2,53E-07		
Razem	8	9,2E-06			

	Współczynniki	*Błąd standardowy*	*t Stat*	*Wartość-p*	*Dolne 95%*	*Górne 95%*
Przecięcie	–0,00345	0,001651	–2,08759	0,075237	–0,00735	0,000457
1/metraż	0,723554	0,133608	5,415514	0,000992	0,407622	1,039485

Dla wyróżnionych podprób o liczebnościach odpowiednio $n_1 = 24$, $n_2 = 9$ stawiamy hipotezy:

$H_o : \delta_{e_1}^2 = \delta_{e_2}^2$,

$H_1 : \delta_{e_1}^2 > \delta_{e_2}^2$.

Zespół hipotez weryfikujemy statystyką:

$$F = \frac{S_{e_1}^2}{S_{e_2}^2}$$

gdzie: $S_{e_1}^2$ – estymator wariancji składników losowych modelu regresji dla podpróby o większej wariancji (mieszkania 1- i 2-pokojowe),

$S_{e_2}^2$ – estymator wariancji składników losowych modelu regresji dla podpróby o mniejszej wariancji (mieszkania 4-pokojowe).

Przy prawdziwości hipotezy zerowej statystyka F ma rozkład F Snedecora o 22 stopniach swobody licznika i o 7 stopniach swobody mianownika. Obliczona z próby wartość statystyki wynosi $F = 1{,}983921$, wartość krytyczna wynosi $F_\alpha = 3{,}43$. Nie ma zatem podstaw do odrzucenia hipotezy H_0.

Wniosek. Nie ma podstaw do odrzucenia hipotezy o równości wariancji składników losowych w badanych podgrupach mieszkań.

NIEOBCIĄŻONOŚĆ

Skonstruowany model ekonometryczny jest nieliniowy. Należy zatem zbadać nieobciążoność składników losowych modelu

$$\hat{cena} = \frac{metraż}{0{,}001218074\,metraż + 0{,}357888172}.$$

W tym celu wyznaczamy reszty modelu nieliniowego E_i (test 18 nieobciążoności składników losowych). Stawiamy hipotezy

$H_0 : E(\tilde{\varepsilon}) = 0$,

$H_1 : E(\tilde{\varepsilon}) \neq 0$.

Hipotezę tę weryfikujemy statystyką

$$t = \frac{\overline{E}}{S_E}\sqrt{n-1}.$$

Statystyka t, przy prawdziwości hipotezy H_0, ma rozkład t Studenta o 51 stopniach swobody.

Obliczona z próby wartość statystyki $t = 2{,}5793E - 14$, a wartość krytyczna $t_\alpha = 2{,}008$. Nie ma zatem podstaw do odrzucenia hipotezy H_0.

Wniosek. Nie ma zatem podstaw do odrzucenia hipotezy o nieobciążoności składników losowych modelu nieliniowego.

Podsumowanie. Możemy zatem uznać model ekonometryczny

$$\hat{cena} = \frac{metraż}{0{,}001218074 + 0{,}357888172\,metraż}$$

za poprawny.

Krok VI. Wnioskowanie na podstawie modelu

Spróbujemy teraz wyznaczyć cenę, za jaką są wystawiane na sprzedaż mieszkania 52-metrowe.

Ocena punktowa ceny mieszkań o powierzchni 52 m^2 wynosi:

$$\hat{cena} = \frac{52}{0{,}001218074 \cdot 52 + 0{,}357888172} = 123451\,\text{zł},$$

a przedział ufności dla ceny 52 metrowych mieszkań (na poziomie ufności 0,95) to:

(97 421 zł, 168 461zł).

W ofercie biura „Twój Dom" znajdowały się 3 mieszkania 52-metrowe za 109 000 zł, 129 000 zł oraz 130 000 zł. Wyznaczony przedział ufności obejmuje wszystkie trzy ceny.

Model sprawdził się w predykcji ekonometrycznej.

3.4. Temperatura we Wrocławiu

Model ekonometryczny opisujący średnią miesięczną temperaturę we Wrocławiu jest modelem dwurównaniowym. Dla miesięcy styczeń–sierpień przyjęto model kwadratowy, dla okresu wrzesień–grudzień model liniowy. Dane do budowy modelu pochodzą z lat 1997–1999. Na podstawie danych z lat 1997–2001 wykonano predykcję przedziałową średniej miesięcznej temperatury dla poszczególnych miesięcy oraz okresów kilkuletnich.

Krok I. Określenie celu badań

Celem badań jest budowa modelu ekonometrycznego umożliwiającego określenie średniej miesięcznej temperatury powietrza we Wrocławiu.

Krok II. Specyfikacja zmiennych wraz z gromadzeniem danych

Średnie miesięczne temperatury powietrza we Wrocławiu z lat 1997–2001 przedstawia tabela 3.18 i wykres na rysunku 3.10.

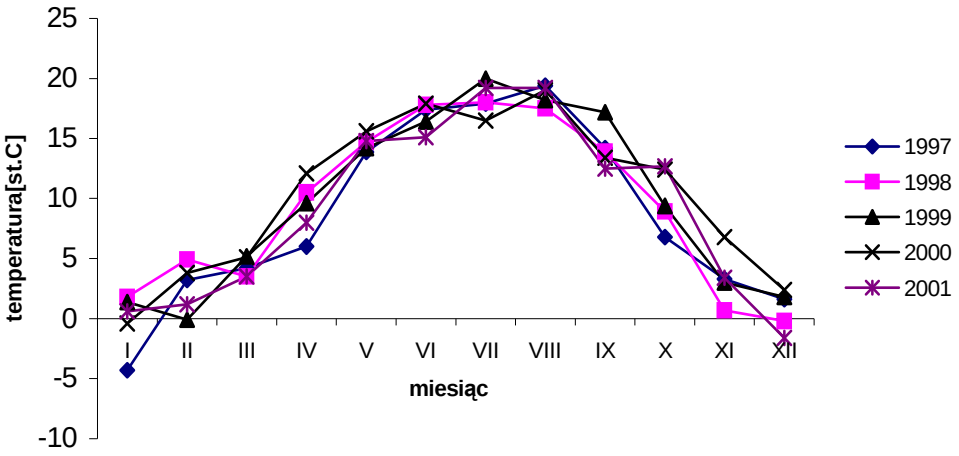
Tabela 3.18. Średnie miesięczne temperatury powietrza we Wrocławiu [°C]

Miesiąc	1997	1998	1999	2000	2001
I	–4,3	1,8	1,4	–0,4	0,6
II	3,2	4,9	-0,1	3,8	1,2
III	4,2	3,5	5,2	5,1	3,5

cd. tabeli 3.18

IV	6,0	10,5	9,6	12,1	8,0
V	13,9	14,7	14,2	15,6	14,8
VI	17,4	17,8	16,4	17,9	15,1
VII	17,9	18,0	20,0	16,5	19,2
VIII	19,4	17,5	18,2	19,0	19,2
IX	14,2	13,9	17,2	13,4	12,5
X	6,8	8,9	9,4	12,4	12,7
XI	3,3	0,7	3,0	6,8	3,4
XII	1,6	–0,2	1,8	2,4	–1,6

Źródło: Dolnośląski Rocznik Statystyczny



Rys. 3.10. Średnia miesięczna temperatura we Wrocławiu

Model ekonometryczny zbudujemy na podstawie danych temperaturowych z lat 1997–1999. Przebieg temperatury w latach 2000–2001 wykorzystamy do oceny trafności predykcji ekonometrycznej. Zmienna objaśniająca przyjmie wartości 1, 2,…,12 zgodnie z numeracją miesięcy.

Krok III. Wybór klasy modelu

Z analizy rocznego przebiegu temperatury można wnioskować, że najchłodniejszym miesiącem roku jest styczeń (średnia temperatura za okres 1997–2001 jest ujemna). Od lutego zaczyna się regularny i dość szybki wzrost temperatury do czerwca, kiedy wzrost temperatury staje się już powolny, aby w lipcu ociągnąć wartość

maksymalną. Również sierpień należy do miesięcy bardzo ciepłych. W następnych miesiącach szybki i systematyczny spadek doprowadza w grudniu do spadku temperatury w okolice zera.Taki przebieg temperatury w skali roku sugeruje, że model ekonometryczny powinien być dwurównaniowy. Z wykresu temperaturowego wynika, że dla okresu styczeń–sierpień można konstruować model wielomianowy, a dla miesięcy wrzesień–grudzień liniowy.

Dla okresu styczeń–sierpień przyjmujemy model wielomianowy trzeciego stopnia (stopień wielomianu wynika z przeprowadzonej wcześniej analizy wzrostu temperatury), a dla miesięcy wrzesień–grudzień model liniowy:

$$\hat{y} = \begin{cases} \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \varepsilon_1 & \text{dla } x = 1, 2, ..., 8 \\ \beta_0 + \beta_1 x + \varepsilon_2 & \text{dla } x = 9, 10, 11, 12 \end{cases}$$

gdzie: y – średnia miesięczna temperatura,
x – miesiąc.

Wartości zmiennych przedstawiono w tabeli 3.19.

Tabela 3.19. Wartości zmiennych objaśnianej i objaśniającej

Temperatura y	Miesiąc x	Temperatura y	Miesiąc x	Temperatura y	Miesiąc x
–4,3	1	1,8	1	1,4	1
3,2	2	4,9	2	–0,1	2
4,2	3	3,5	3	5,2	3
6,0	4	10,5	4	9,6	4
13,9	5	14,7	5	14,2	5
17,4	6	17,8	6	16,4	6
17,9	7	18.0	7	20,0	7
19,4	8	17,5	8	18,2	8
14,2	9	13,9	9	17,2	9
6,8	10	8,9	10	9,4	10
3,3	11	0,7	11	3.0	11
1,6	12	–0,2	12	1,8	12

Krok IVa. Estymacja parametrów strukturalnych modelu ekonometrycznego dla okresu styczeń–sierpień

Po wprowadzeniu następujących podstawień:
$x_1 = x$
$x_2 = x^2$
$x_3 = x^3$

otrzymujemy model ekonometryczny średniej miesięcznej temperatury postaci:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \varepsilon_1 .$$

Dane do wyznaczenia współczynników modelu metodą najmniejszych kwadratów przyjmą postać (tabela 3.20):

Tabela 3.20. Dane do modelu wielomianowego

Temperatura y	Miesiąc x_1	Miesiąc^2 x_2	Miesiąc^3 x_3
–4,3	1	1	1
3,2	2	4	8
4,2	3	9	27
6	4	16	64
13,9	5	25	125
17,4	6	36	216
17,9	7	49	343
19,4	8	64	512
1,8	1	1	1
4,9	2	4	8
3,5	3	9	27
10,5	4	16	64
14,7	5	25	125
17,8	6	36	216
18	7	49	343
17,5	8	64	512
1,4	1	1	1
–0,1	2	4	8
5,2	3	9	27
9,6	4	16	64
14,2	5	25	125
16,4	6	36	216
20	7	49	343
18,2	8	64	512

Wyznaczone estymatory współczynników modelu liniowego są następujące:

Statystyki regresji	
Wielokrotność R	0,975129
R kwadrat	0,950878
Dopasowany R kwadrat	0,943509
Błąd standardowy	1,7729
Obserwacje	24

ANALIZA WARIANCJI

	Df	*SS*	*MS*	*F*	*Istotność F*
Regresja	3	1216,866	405,622	129,0485	2,96E-13
Resztkowy	20	62,8635	3,143175		
Razem	23	1279,73			

	Współczynniki	*Błąd standardowy*	*t Stat*	*Wartość-p*	*Dolne 95%*	*Górne 95%*
Przecięcie	0,9	2,522139	0,35684	0,724949	–4,36109	6,161088
x1	–2,48925	2,282162	–1,09074	0,288353	–7,24975	2,271254
x2	1,708189	0,572449	2,984004	0,007335	0,514082	2,902296
x3	–0,14066	0,041998	–3,34911	0,003195	–0,22826	–0,05305

Model ekonometryczny ma postać:

$$\hat{temperatura} = 0{,}9 - 2{,}48925x + 1{,}708189x^2 - 0{,}14066x^3.$$

Krok Va. Weryfikacja modelu

Zbudowany model ekonometryczny zweryfikujemy na poziomie istotności $\alpha = 0{,}05$.

Dopasowanie modelu do danych empirycznych. Współczynnik determinacji modelu wynosi $R^2 = 0{,}950878$ (współczynnik zbieżności $\varphi^2 = 4{,}9\%$).

Wniosek. Model wyjaśnia 95,1% zmienności badanej cechy, świadczy to o dobrym dopasowaniu modelu do danych empirycznych.

Istotność układu współczynników regresji. Stawiamy hipotezy (test 1):

$$H_0 : \sum_{j=0}^{n} \alpha_j^2 = 0,$$

$$H_1 : \sum_{j=0}^{n} \alpha_j^2 \neq 0.$$

Zespół hipotez weryfikujemy statystyką

$$F = \frac{R^2}{1-R^2} \frac{n-k-1}{k}.$$

Statystyka F, przy prawdziwości hipotezy zerowej, ma rozkład F Snedecora o 3 stopniach swobody licznika i 20 stopniach swobody mianownika. Wartość empiryczna statystyki wynosi $F = 129{,}0485$, a odpowiadający jej krytyczny poziom istot-

ności (istotność *F*) wynosi 2,96E-13 i jest mniejszy od przyjętego poziomu istotności $\alpha = 0{,}05$. Odrzucamy zatem hipotezę H_0 na korzyść H_1.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że średnia miesięczna temperatura w okresie styczeń–sierpień zależy przynajmniej od jednej ze zmiennych x, x^2, x^3.

Istotność poszczególnych współczynników regresji. Dla każdego współczynnika modelu regresji (j = 0, 1, 2, 3) stawiamy hipotezy (test 2):

$H_0 : \alpha_j = 0$,

$H_1 : \alpha_j \neq 0$.

Zespół hipotez weryfikujemy statystyką

$$t(a_j) = \frac{a_j}{S(a_j)}.$$

Statystyka ta, przy prawdziwości hipotez zerowych, ma rozkład *t* Studenta o 20 stopniach swobody.

Empiryczne wartości statystyk *t* Studenta wynoszą:

$t(\alpha_0) = 0{,}35684$,

$t(\alpha_1) = -1{,}09074$,

$t(\alpha_2) = 2{,}984004$,

$t(\alpha_3) = -3{,}34911$.

Odpowiadające im wartości krytycznego poziomu istotności (wartość-*p*) wynoszą odpowiednio 0,724949; 0,288353; 0,007335 oraz 0,003195. Nie ma zatem podstaw do odrzucenia hipotezy, że stała modelu α_0 oraz współczynnik α_1 jest nieistotny, czyli są równe zeru (wartości krytycznego poziomu istotności dla tych współczynników są większe od przyjętego poziomu istotności $\alpha = 0{,}05$).

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że zmienna $x_1 = x$ jest nieistotna. Analizowany model nie jest poprawny.

Krok IIIa′. Ponowny wybór klasy modelu dla okresu styczeń–sierpień

Usuwamy z modelu zmienną x_1, która okazała się zmienną nieistotną i sprawdzamy, czy model

$$y = \alpha'_0 + \alpha'_2 x^2 + \alpha'_3 x^3 + \varepsilon$$

dobrze opisuje rzeczywistość.

Krok IVa′. Estymacja parametrów strukturalnych modelu ekonometrycznego dla okresu styczeń–sierpień

Wyniki estymacji parametrów modelu liniowego $y = \alpha'_0 + \alpha'_2 x_2 + \alpha'_3 x_3 + \varepsilon$ są następujące:

Statystyki regresji	
Wielokrotność R	0,97363
R kwadrat	0,947955
Dopasowany R kwadrat	0,942999
Błąd standardowy	1,780891
Obserwacje	24

ANALIZA WARIANCJI

	df	*SS*	*MS*	*F*	*Istotność F*
Regresja	2	1213,127	606,5633	191,2501	3,33E-14
Resztkowy	21	66,60299	3,171571		
Razem	23	1279,73			

	Współczynniki	*Błąd standardowy*	*t Stat*	*Wartość-p*	*Dolne 95%*	*Górne 95%*
Przecięcie	–1,7051	0,814108	–2,09444	0,048529	–3,39814	–0,01207
x2	1,095255	0,109665	9,987316	1,98E-09	0,867195	1,323315
x3	–0,09724	0,013459	–7,22495	4,05E-07	–0,12523	–0,06925

Wyznaczony model ekonometryczny ma postać:

$$\hat{temperatura} = -1{,}7051 + 1{,}095255\, x^2 - 0{,}09724\, x^3 .$$

Krok Va′. Weryfikacja modelu

Weryfikację modelu przeprowadzamy na poziomie istotności 0,05.

Dopasowanie modelu do danych empirycznych. Współczynnik modelu wynosi $R^2 = 0{,}947955$ (współczynnik zbieżności $\varphi^2 = 5{,}5\%$).

Wniosek. Model wyjaśnia 94,3% zmienności badanej cechy. Świadczy to o dobrym dopasowaniu modelu do danych empirycznych.

Istotność układu współczynników regresji. Stawiamy hipotezę o nieistotności układu współczynników modelu regresji (test 1). Statystyka *F*, przy prawdziwości hipotezy zerowej, ma rozkład *F* Snedecora o 3 stopniach swobody licznika i 21 stopniach swobody mianownika.

Wartość empiryczna statystyki wynosi F = 191,2501, a odpowiadający jej krytyczny poziom istotności (istotność F) wynosi 3,33E-14 jest mniejszy od przyjętego poziomu istotności $\alpha = 0{,}05$. Odrzucamy zatem hipotezę H_0 na korzyść H_1.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że średnia miesięczna temperatura w okresie styczeń–sierpień zależy przynajmniej od jednej ze zmiennych x^2, x^3.

Istotność poszczególnych współczynników regresji. Dla każdego współczynnika modelu regresji α_j (j = 0, 2, 3) testujemy hipotezę o jego istotności (test 2). Weryfikujemy je statystyką o rozkładzie t Studenta o 21 stopniach swobody.

Empiryczne wartości statystyk t Studenta wynoszą:

$t(\alpha_0') = -2{,}0944$,

$t(\alpha_2') = 9{,}987316$,

$t(\alpha_3') = -7{,}22495$.

Odpowiadające im wartości krytycznego poziomu istotności (wartość-p) wynoszą odpowiednio 0,048529; 1,98E-09 oraz 4,05E-07. Wszystkie zatem współczynniki modelu są istotnie różne od zera (wartości krytycznego poziomu istotności są mniejsze od przyjętego poziomu istotności $\alpha = 0{,}05$).

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że wszystkie współczynniki badanego modelu są istotnie różne od zera.

Analiza składników losowych modelu. Reszty modelu ekonometrycznego przedstawiono w tabeli 3.21.

Tabela 3.21. Reszty modelu wielomianowego

Obserwacja	*Przewidywane Y*	*Składniki resztowe*	*Std. składniki resztowe*
1	–0,70709	–3,59291	–2,11137
2	1,897987	1,302013	0,765125
3	5,526681	–1,32668	–0,77962
4	9,595544	–3,59554	–2,11291
5	13,52113	0,378871	0,222643
6	16,71999	0,68001	0,399607
7	18,60868	–0,70868	–0,41645
8	18,60375	0,796252	0,467916
9	–0,70709	2,50709	1,473286
10	1,897987	3,002013	1,764126
11	5,526681	–2,02668	–1,19097
12	9,595544	0,904456	0,531502
13	13,52113	1,178871	0,692761
14	16,71999	1,08001	0,634666
15	18,60868	–0,60868	–0,35769
16	18,60375	–1,10375	–0,64862
17	–0,70709	2,10709	1,238227
18	1,897987	–1,99799	–1,17411

cd. tabeli 3.21

19	5,526681	–0,32668	–0,19197
20	9,595544	0,004456	0,002619
21	13,52113	0,678871	0,398937
22	16,71999	–0,31999	–0,18804
23	18,60868	1,391322	0,817607
24	18,60375	–0,40375	–0,23726

NORMALNOŚĆ

Stawiamy hipotezę H_0 składniki losowe mają rozkład $N(0, 1{,}780891)$.

Weryfikację hipotezy przeprowadzimy testem Shapiro–Wilka (test 5). Wyznaczamy wartość statystyki testowej:

$$W = \frac{\left[\sum_{i=1}^{\left[\frac{n}{2}\right]} a_{n,i}\left(e_{(n-i+1)} - e_{(i)}\right)\right]^2}{\sum_{i=1}^{n}\left(e_i - \bar{e}\right)^2}$$

gdzie: $a_{n,i}$ – współczynniki Shapiro–Wilka,

$e_{(1)}, e_{(2)}, \ldots, e_{(n)}$ – wartości reszt uporządkowane niemalejąco.

Empiryczna wartość statystyki W wynosi 0,963931. Wartość krytyczna $W_\alpha = 0{,}916$. Ponieważ $W > W_\alpha$, nie ma więc podstaw do odrzucenia hipotezy H_0.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że składniki losowe modelu mają rozkład normalny $N(0; 1{,}780891)$.

SYMETRIA

Stawiamy hipotezę H_0 o symetrii składników losowych (test 12). Statystyka testowa, przy prawdziwości hipotezy H_0, ma rozkład t Studenta o 23 stopniach swobody. Empiryczna wartość statystyki dla 13 reszt dodatnich wynosi $t = 0{,}401048$. Wartość krytyczna 2,069. Nie ma zatem podstaw do odrzucenia hipotezy H_0.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że rozkład składników losowych jest symetryczny.

LOSOWOŚĆ

Stawiamy hipotezę zerową H_0: reszty modelu są losowe. Zweryfikujemy tę hipotezę testem maksymalnej długości serii (test 14). Jednoznaczne uporządkowanie reszt w naszym przypadku nie jest możliwe ze względu na trzykrotne obserwacje temperatury dla każdego miesiąca (tej samej wartości zmiennej objaśniającej). Uporządkujemy zatem reszty tak, aby otrzymać najgorszy przypadek dla tego testu, to jest najdłuższą serię z możliwych w ramach dopuszczalnych uszeregowań. Najdłuższą serię o długości $L_{\max} = 7$ otrzymamy z reszt dodatnich dla miesięcy kwiecień, maj, czerwiec.

Tabela 3.22. Uporządkowane reszty modelu wielomianowego

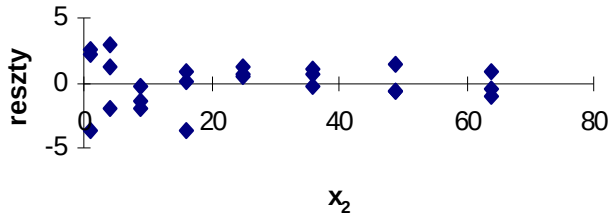
Miesiąc	Reszty	Miesiąc	Reszty
1	–3,59291	5	0,378871
1	2,50709	5	1,178871
1	2,10709	5	0,678871
2	1,302013	6	0,68001
2	3,002013	6	1,08001
2	–1,99799	6	–0,31999
3	–1,32668	7	–0,70868
3	–2,02668	7	–0,60868
3	–0,32668	7	1,391322
4	–3,59554	8	0,796252
4	0,904456	8	–1,10375
4	0,004456	8	–0,40375

Dla maksymalnej długości serii $L_{\max} = 7$ minimalna liczba obserwacji na przyjętym poziomie istotności $\alpha = 0{,}05$ wynosi $n_\alpha = 22$. Ponieważ w naszym przykładzie $n = 24$, zatem $n_\alpha < n$, a więc nie mamy podstaw do odrzucenia hipotezy zerowej.

Wniosek. Nie ma podstaw do odrzucenia hipotezy o losowości reszt modelu.

HOMOSKEDASTYCZNOŚĆ

Stałość wariancji składnika losowego zbadamy testem Spearmana (test 17).



Rys. 3.11. Reszty modelu wielomianowego

Na wykresie (rys. 3.11) reszt można zaobserwować, że reszty modelu są większe w miesiącach zimowych i maleją wraz ze wzrostem wartości zmiennej objaśniającej. Testem Spearmana sprawdzimy, czy wariancja składników losowych maleje liniowo wraz ze wzrostem wartości zmiennej objaśniającej x_2. W naszym przypadku taki sam wynik otrzymamy, gdy w miejsce x_2 przyjmiemy zmienną x.

Stawiamy hipotezy

$H_0 : \rho(|\varepsilon_x|, x) = 0$,

$H_1 : \rho(|\varepsilon_x|, x) \neq 0$.

Sprawdzianem tego zespołu hipotez jest statystyka korelacji rangowej Spearmana

$$r = r\left(|\varepsilon|, x_2\right) = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n\left(n^2 - 1\right)},$$

gdzie D_i – różnica rang zmiennej x oraz modułu reszt modelu dla i-tej obserwacji.

Rangi (1, 2,...,n) przypisujemy kolejno wartościom zmiennej x (reszt modelu e) uporządkowanym w ciąg niemalejący. Zmienna x każdą wartość przyjmuje trzykrotnie, zatem tym samym wartościom przypisujemy rangę równą średniej arytmetycznej odpowiadających im pozycji w ciągu.

Tabela 3.23. Obliczenia do testu Spearmana

Moduł reszt	Ranga reszt	x	Ranga x	D	D^2
0,004456	1	4	11	10	100
0,319990	2	6	17	15	225
0,326681	3	3	8	5	25
0,378871	4	5	14	10	100
0,403748	5	8	23	18	324
0,608678	6	7	20	14	196
0,678871	7	5	14	7	49
0,680010	8	6	17	9	81
0,708678	9	7	20	11	121
0,796252	10	8	23	13	169
0,904456	11	4	11	0	0
1,080010	12	6	17	5	25
1,103748	13	8	23	10	100
1,178871	14	5	14	0	0
1,302013	15	2	5	–10	100
1,326681	16	3	8	–8	64
1,391322	17	7	20	3	9
1,997987	18	2	5	–13	169
2,026681	19	3	8	–11	121
2,107090	20	1	2	–18	324
2,507090	21	1	2	–19	361
3,002013	22	2	5	–17	289
3,592910	23	1	2	–21	441
3,595544	24	4	11	–13	169
				SUMA	3562

Na podstawie obliczeń z tabeli 3.23 wyznaczamy wartość empiryczną statystyki $r = -0{,}5487$. Statystyka r, przy prawdziwości hipotezy H_0, ma rozkład asymptotycznie normalny $N\left(0, \frac{1}{\sqrt{23}}\right)$.

W wyniku standaryzacji

$$\frac{r}{\frac{1}{\sqrt{23}}} = -2{,}63145$$

otrzymujemy wartość empiryczną statystyki, która, przy prawdziwości hipotezy H_0, ma rozkład $N(0, 1)$.

Obszar krytyczny testu jest dwustronny. Na poziomie istotności $\alpha = 0{,}05$ wartość krytyczna wynosi 1,96 i hipotezę H_0 o stałości wariancji składników losowych należałoby odrzucić. Nie mamy natomiast podstaw do odrzucenia hipotezy H_0 na poziomie istotności $\alpha = 0{,}005$, któremu odpowiada wartość krytyczna 2,81.

Wniosek. Nie ma podstaw do odrzucenia hipotezy o homoskedastyczności składników losowych (należy zwrócić uwagę na zmianę poziomu istotności).

Podsumowanie. Możemy uznać, że dla okresu styczeń–sierpień skonstruowany model ekonometryczny:

$$\hat{temperatura} = -1{,}7051 + 1{,}095255x^2 - 0{,}09724x^3$$

jest poprawny.

Krok IVb. Estymacja parametrów strukturalnych modelu ekonometrycznego dla okresu wrzesień–grudzień

Zgodnie z wcześniejszymi założeniami, dla miesięcy wrzesień-grudzień, skonstruujemy liniowy model regresji postaci:

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

gdzie: y – średnia miesięczna temperatura,
x – miesiąc.

Wyniki estymacji parametrów tego modelu liniowego są następujące:

Statystyki regresji	
Wielokrotność R	0,944752
R kwadrat	0,892556
Dopasowany R kwadrat	0,881812
Błąd standardowy	2,045336
Obserwacje	12

ANALIZA WARIANCJI

	df	*SS*	*MS*	*F*	*Istotność F*
Regresja	1	347,5227	347,5227	83,07182	3,69E-06
Resztkowy	10	41,834	4,1834		
Razem	11	389,3567			

	Współczynniki	*Błąd standardowy*	*t Stat*	*Wartość-p*	*Dolne 95%*	*Górne 95%*
Przecięcie	57,25667	5,576433	10,26761	1,25E-06	44,8316	69,68174
X	–4,81333	0,528104	–9,11437	3,69E-06	–5,99002	–3,63665

Wyznaczony model ma postać:

$$\hat{temperatura} = 57{,}25667 - 4{,}81333\,x\,.$$

Krok Vb. Weryfikacja modelu

Badany model ekonometryczny zweryfikujemy na poziomie istotności $\alpha = 0{,}05$.

Dopasowanie modelu do danych empirycznych. Współczynnik dopasowania modelu wynosi $R^2 = 0{,}892556$.

Wniosek. Model wyjaśnia 89,3% zmienności badanej cechy, świadczy to o dobrym dopasowaniu modelu do danych empirycznych.

Istotność układu współczynników regresji. Stawiamy hipotezę o braku istotności układu współczynników (test 1) i weryfikujemy ją statystyką o rozkładzie *F* Snedecora o 1 stopniu swobody licznika i 10 stopniach swobody mianownika.

Wartość empiryczna statystyki wynosi $F = 83{,}07182$, a odpowiadający jej krytyczny poziom istotności (istotność *F*) wynosi 3,69E-6 i jest mniejszy od przyjętego poziomu istotności $\alpha = 0{,}05$. Odrzucamy zatem hipotezę H_0 na korzyść H_1.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że średnia miesięczna temperatura w okresie wrzesień–grudzień zależy od zmiennej *x*.

Istotność poszczególnych współczynników regresji. Istotność poszczególnych współczynników regresji zbadamy w klasyczny sposób (test 1) statystyką o rozkładzie *t* Studenta o 10 stopniach swobody.

Obliczone wartości statystyk *t* Studenta wynoszą:

$t(b_0) = 10{,}26761$,

$t(b_1) = -9{,}11437$.

Odpowiadające im wartości krytycznego poziomu istotności (wartość-p) wynoszą odpowiednio 1,25E-6; 3,69E-6. Wszystkie zatem współczynniki modelu są istotnie

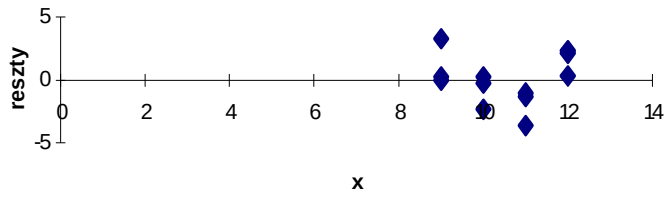
różne od zera (wartości krytycznego poziomu istotności są mniejsze od przyjętego poziomu istotności $\alpha = 0{,}05$).

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że oba współczynniki modelu są istotne.

Analiza składników losowych modelu. Reszty modelu ekonometrycznego przedstawiono w tabeli 3.24 i na rysunku 3.12.

Tabela 3.24. Reszty modelu liniowego

Obserwacja	Przewidywane Y	Składniki resztowe	Std. składniki resztowe
1	13,93667	0,263333	0,135032
2	9,123333	–2,32333	–1,19136
3	4,31	–1,01	–0,51791
4	–0,50333	2,103333	1,078549
5	13,93667	–0,03667	–0,0188
6	9,123333	–0,22333	–0,11452
7	4,31	–3,61	–1,85114
8	–0,50333	0,303333	0,155543
9	13,93667	3,263333	1,673374
10	9,123333	0,276667	0,141869
11	4,31	–1,31	–0,67174
12	–0,50333	2,303333	1,181105



Rys. 3.12. Reszty modelu liniowego

NORMALNOŚĆ

Stawiamy hipotezę H_0: składniki losowe mają rozkład $N(0, 2{,}045336)$.
Zweryfikujemy tę hipotezę testem Shapiro–Wilka (test 5) statystyką:

$$W = \frac{\left[\sum_{i=1}^{\left[\frac{n}{2}\right]} a_{n,i}\left(e_{(n-i+1)} - e_{(i)}\right)\right]^2}{\sum_{i=1}^{n}\left(e_i - \bar{e}\right)^2}$$

gdzie: $a_{n,i}$ – współczynniki Shapiro–Wilka,

$e_{(1)}, e_{(2)}, ..., e_{(n)}$ – wartości reszt uporządkowane niemalejąco.

Empiryczna wartość statystyki W wynosi 0,96873. Wartość krytyczna $W_\alpha = 0{,}859$. Ponieważ $W > W_\alpha$, nie ma więc podstaw do odrzucenia hipotezy H_0.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że składniki losowe mają rozkład normalny $N(0, S_\varepsilon = 2{,}045336)$.

SYMETRIA

Stawiamy hipotezę o symetryczności reszt (test 12) i weryfikujemy ją statystyką o rozkładzie t Studenta o 11 stopniach swobody. Empiryczna wartość statystyki dla 6 reszt dodatnich wynosi $t = 0$. Wartość krytyczna 2,201. Nie ma zatem podstaw do odrzucenia hipotezy H_0.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że rozkład składników losowych jest symetryczny.

LOSOWOŚĆ

Stawiamy hipotezę zerową o tym, że reszty modelu są losowe i weryfikujemy ją testem maksymalnej długości serii (test 14).

Jednoznaczne uporządkowanie reszt w naszym przypadku nie jest możliwe ze względu na trzykrotne obserwacje temperatury dla każdego miesiąca (tej samej wartości zmiennej objaśniającej). Uporządkujemy zatem reszty, konstruując najgorszy przypadek dla tego testu, to jest konstruując najdłuższą z możliwych serii w ramach dopuszczalnych uszeregowań. Najdłuższą serię o długości $L_{\max} = 5$ możemy otrzymać z reszt ujemnych dla miesięcy październik i listopad (tabela 3.25).

Tabela 3.25. Uporządkowane reszty modelu liniowego

Miesiąc	Reszty
9	–0,03667
9	0,263333
9	3,263333
10	0,276667
10	**–2,32333**
10	**–0,22333**
11	**–1,01**
11	**–3,61**
11	**–1,31**
12	2,103333
12	0,303333
12	2,303333

Dla maksymalnej długości serii $L_{max} = 5$ minimalna liczba obserwacji na przyjętym poziomie istotności $\alpha = 0{,}05$ wynosi $n_\alpha = 10$. Ponieważ w naszym przykładzie $n = 12$, zatem $n_\alpha < n$, nie ma podstaw do odrzucenia hipotezy zerowej.

Wniosek. Nie ma podstaw do odrzucenia hipotezy o losowości reszt modelu.

HOMOSKEDASTYCZNOŚĆ

Podobnie jak dla okresu styczeń–sierpień, stałość wariancji składnika losowego zbadamy testem Spearmana (test 17).

Rangi (1, 2,..., *n*) przypisujemy kolejno wartościom zmiennej *x* jako średnie arytmetyczne, odpowiadających im pozycjom w ciągu uporządkowanym niemalejąco.

Tabela 3.26. Obliczenia do testu Spearmana

Reszty	Moduł reszt	Ranga reszt	x_1	Ranga x_2	D	D^2
–0,03667	0,036667	1	9	2	1	1
0,263333	0,263333	2	10	5	3	9
3,263333	3,263333	3	9	2	–1	1
0,276667	0,276667	4	10	5	1	1
–2,32333	2,323333	5	12	11	6	36
–0,22333	0,223333	6	11	8	2	4
–1,01	1,01	7	11	8	1	1
–3,61	3,61	8	12	11	3	9
–1,31	1,31	9	12	11	2	4
2,103333	2,103333	10	10	5	–5	25
0,303333	0,303333	11	9	2	–9	81
2,303333	2,303333	12	11	8	–4	16
					SUMA	188

Na podstawie obliczeń (tabela 3.26) wyznaczamy wartość empiryczną statystyki wynoszącą $r = 0{,}343$. Obszar krytyczny testu jest dwustronny. Na poziomie istotności $\alpha = 0{,}05$ wartość krytyczna testu wynosi 0,497 i jest większa od wartości empirycznej 0,343.

Wniosek. Nie ma zatem podstaw do odrzucenia hipotezy o stałości wariancji składników losowych.

Podsumowanie. Skonstruowany dla okresu styczeń–sierpień model ekonometryczny można uznać:

$$\hat{temperatura} = 57{,}25667 - 4{,}81333\,x$$

za poprawny.

Model ekonometryczny średniej miesięcznej temperatury:

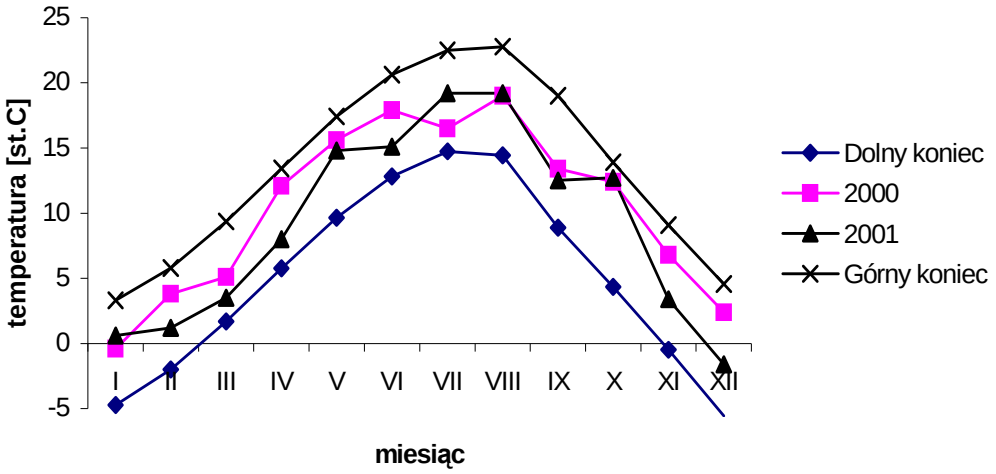
$$\hat{temp} = \begin{cases} -1{,}705 + 1{,}095x^2 - 0{,}097x^3 & \text{dla} \quad x = 1,\ 2, \ldots,\ 8 \\ 57{,}256 - 4{,}813x & \text{dla} \quad x = 9, 10, 11, 12 \end{cases}$$

uznajemy za poprawny.

Krok VI. Wnioskowanie na podstawie modelu

Na podstawie skonstruowanego modelu skonstruujemy przedziałową ocenę średniej temperatury oraz wartości oczekiwanej średniej temperatury we Wrocławiu.

Przedziałową ocenę średniej temperatury wyznaczają przedziały ufności dla poziomu ufności ($1 - \alpha = 0{,}95$) – (tabela 3.27).



Rys. 3.13. Średnie temperatury miesięczne we Wrocławiu oraz przedziały ufności średnich temperatur w latach 2000–2001

Tabela 3.27. Przedziałowa prognoza średniej miesięcznej temperatury

Miesiąc	Dolny koniec	Górny koniec
I	–4,71716	3,30298
II	–1,996	5,79198
III	1,70039	9,35297
IV	5,7527	13,4384
V	9,63009	17,4122
VI	12,8267	20,6133
VII	14,7297	22,4877
VIII	14,4262	22,7813
IX	8,87553	18,9978
X	4,34359	13,9031
XI	–0,46974	9,08974
XII	–5,56447	4,5578

Na wykresie (rys. 3.13) widać, że 100% obserwacji średniej miesięcznej temperatury z lat 2000 i 2001 mieści się w wyznaczonym przedziale ufności.

Przedziałową ocenę wartości oczekiwanej średniej temperatury wieloletniej wyznaczają przedziały ufności (tab. 3.28).

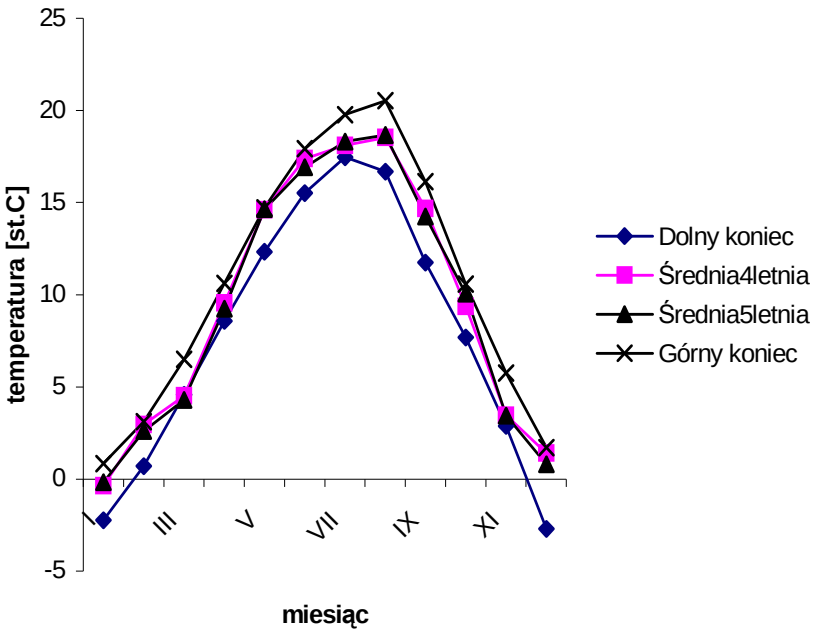
Tabela 3.28. Przedziałowa predykcja wartości oczekiwanej średniej miesięcznej temperatury

Miesiąc	Dolny koniec	Górny koniec
I	–2,2447	0,830516
II	0,6952	3,10077
III	4,5654	6,48796
IV	8,57035	10,6207
V	12,3279	14,7143
VI	15,5195	17,9205
VII	17,4553	19,7621
VIII	16,6711	20,5364
IX	11,7353	16,1381
X	7,68219	10,5645
XI	2,86885	5,75115
XII	–2,70472	1,69805

Sprawdzimy, czy średnie cztero- i pięcioletnie temperatury miesięczne w latach 1997–2000 i 1997–2001 (tab. 3.29) mieszczą się w tych przedziałach.

Tabela 3.29. Średnia miesięczna temperatura dla okresu cztero- i pięcioletniego

Średnia temperatura 4-letnia	Średnia temperatura 5-letnia
–0,38	–0,18
2,95	2,6
4,5	4,3
9,55	9,24
14,6	14,64
17,38	16,92
18,1	18,32
18,53	18,66
14,68	14,24
9,375	10,04
3,45	3,44
1,4	0,8



Rys. 3.14. Średnie miesięczne temperatury 4- i 5-letnie we Wrocławiu oraz przedziały ufności

Z przedstawionych na wykresie (rys. 3.14) i w tabeli 3.29 średnich miesięcznych temperatur wieloletnich oraz przedziałów ufności dla wartości oczekiwanej średniej temperatury można zaobserwować, że jedynie jedna prognoza (średnia 4-letnia temperatura kwietnia) nie mieści się w wyznaczonym przedziale ufności. Stanowi to $1/24 \cdot 100\% = 4{,}2\%$ wszystkich prognoz, co jest statystycznie dopuszczalne w przypadku przedziałów ufności wyznaczonych na poziomie ufności 0,95.

Model sprawdził się w prognozowaniu średniej miesięcznej temperatury dla dwóch kolejnych lat (2000, 2001) oraz w prognozowaniu średniej miesięcznej temperatury wieloletniej (1997–2000 i 1997–2001).

3.5. Podaż pieniądza

Model podaży pieniądza w Polsce zbudowano na podstawie danych z okresu od stycznia 1998 do marca 2001. Należy on do klasy modeli autoregresyjnych ze zmienną objaśnianą występującą w roli zmiennej objaśniającej z opóźnieniem miesięcznym i rocznym. Średni błąd względny predykcji dla okresu kwiecień 2001–marzec 2002 wyniósł 1,36% (dla maksymalnej wartości błędu na poziomie 3,69%).

Krok I. Cel badań

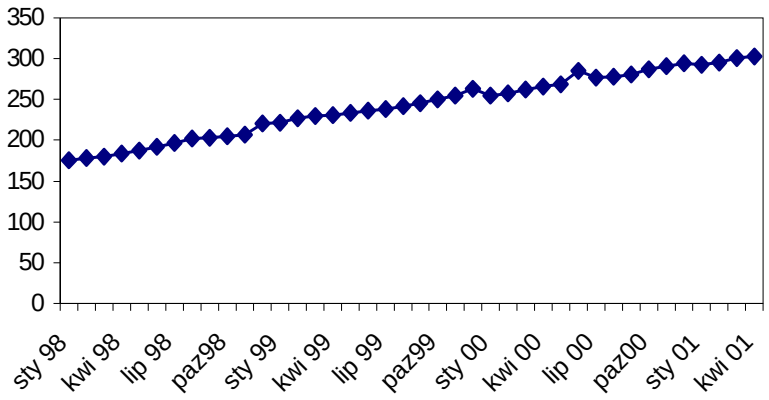
Celem badań jest budowa modelu ekonometrycznego, który umożliwiłby analizę struktury oraz prognozę podaży pieniądza w Polsce.

Podaż pieniądza obok PKB, stopy bezrobocia, czy stopy inflacji należy do podstawowych wskaźników makroekonomicznych.

Podaż pieniądza to całkowita wartość znajdujących się w obiegu zasobów pieniądza. Obejmuje ona:

- pieniądz gotówkowy w obiegu (poza kasami banków),
- zobowiązania gotówkowe złotówkowe wobec osób prywatnych i podmiotów gospodarczych, tj. złotowe depozyty bieżące, złotowe depozyty terminowe i zablokowane oraz złotową część kategorii: bony oszczędnościowe i certyfikaty depozytowe (niezbywalne),
- pożyczki otrzymane od funduszy i fundacji niefinansowych,
- kredyty i pożyczki otrzymane od niebankowych instytucji finansowych,
- zobowiązania z tytułu sprzedaży papierów wartościowych z udzielonym przyrzeczeniem odkupu,
- zobowiązania walutowe wobec osób prywatnych i podmiotów gospodarczych (walutowe depozyty bieżące, walutowe depozyty terminowe i zablokowane oraz walutową część kategorii: bony oszczędnościowe i certyfikaty depozytowe niezbywalne),
- pożyczki otrzymane od funduszy i fundacji niefinansowych,
- kredyty i pożyczki otrzymane od niebankowych instytucji finansowych,
- zobowiązania z tytułu sprzedaży papierów wartościowych z udzielonym przyrzeczeniem odkupu.

Krok II. Specyfikacja zmiennych wraz z gromadzeniem danych



Rys. 3.15. Podaż pieniądza w Polsce

Aktualne dane o wielkości podaży pieniądza w Polsce są dostępne w Internecie na stronie http://www.money.pl/gospodarka/wskazniki/pkb/.

Analizowane dane obejmujące okres od stycznia 1998 do marca 2001 przedstawiono w tabeli 3.30 i na rysunku 3.15.

Tabela 3.30. Podaż pieniądza w Polsce

Data	Czas	Podaż pieniądza [mld zł]	Data	Czas	Podaż pieniądza [mld zł]	Data	Czas	Podaż pieniądza [mld zł]	Data	Czas	Podaż pieniądza [mld zł]
I 98	1	175,7	I 99	13	221,8	I 00	25	255,3	I 01	37	292,6
II 98	2	178,2	II 99	14	226,8	II 00	26	257,8	II 01	38	295,5
III 98	3	180,4	III 99	15	230,3	III 00	27	262	III 01	39	301
IV 98	4	183,6	IV 99	16	230,8	IV 00	28	265,8			
V 98	5	187,4	V 99	17	233,3	V 00	29	268,7			
VI 98	6	192,3	VI 99	18	236,2	VI 00	30	284,9			
VII 98	7	196,9	VII 99	19	238,5	VII 00	31	277,1			
VIII 98	8	202,2	VIII 99	20	241,8	VIII 00	32	277,9			
IX 98	9	203,5	IX 99	21	246	IX 00	33	280,6			
X 98	10	204,8	X 99	22	250,7	X 00	34	287,4			
XI 98	11	207,1	XI 99	23	254,6	XI 00	35	291,2			
XII 98	12	220,8	XII 99	24	263,5	XII 00	36	294,4			

Źródło: http://www.money.pl/gospodarka/wskazniki/pkb/

Krok III. Wybór klasy modelu

Analiza danych wykazuje, iż podaż pieniądza ma tendencję wzrostową. Sprawdzimy, czy jest to trend o charakterze liniowym:

$$podaż\ pieniądza = \alpha_0 + \alpha_1 czas + \varepsilon$$

na co wydaje się wskazywać wykres (rys. 3.15). W tym celu za zmienną objaśnianą przyjmujemy wielkość podaży pieniądza w okresach y_t, a za zmienną objaśniającą czas w skali bezwzględnej $t = 1, 2, ..., 39$.

Krok IV. Estymacja współczynników regresji

Wyniki estymacji modelu liniowego zależności podaży pieniądza od czasu są następujące:

Statystyki regresji	
Wielokrotność R	0,994756
R kwadrat	0,989539
Dopasowany R kwadrat	0,989256
Błąd standardowy	3,94549
Obserwacje	39

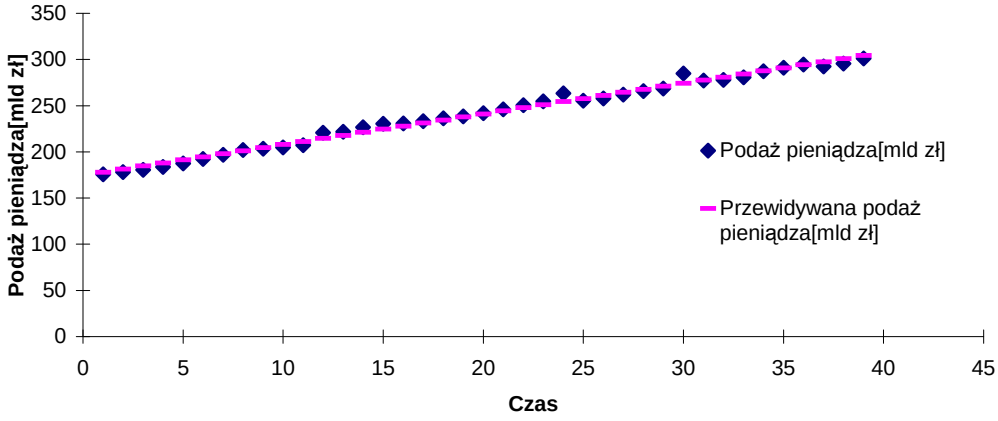
ANALIZA WARIANCJI

	df	*SS*	*MS*	*F*	*Istotność F*
Regresja	1	54483,86	54483,86	3499,984	3,01E-38
Resztkowy	37	575,9748	15,56689		
Razem	38	55059,84			

	Współczynniki	*Błąd standardowy*	*t Stat*	*Wartość-p*	*Dolne 95%*	*Górne 95%*
Przecięcie	174,59	1,288265	135,5233	1,68E-51	171,9797	177,2003
Czas	3,321012	0,056135	59,16066	3,01E-38	3,207271	3,434753

Równanie trendu ma zatem postać (rys. 3.16):

$$\hat{y}_t = 174{,}59 + 3{,}321012\,t\,.$$



Rys. 3.16. Model liniowy podaży pieniądza w Polsce

Krok V. Weryfikacja modelu

Skonstruowany model ekonometryczny zweryfikujemy na poziomie istotności $\alpha = 0{,}05$.

Dopasowanie modelu do danych empirycznych. Współczynnik dopasowania modelu wynosi R^2 = 0,989256 (współczynnik zbieżności φ^2 =1,1%).

Wniosek. Model wyjaśnia 98,9% zmienności badanej cechy, świadczy to o dobrym dopasowaniu modelu do danych empirycznych.

Istotność układu współczynników regresji. Stawiamy hipotezę o nieistotności układu współczynników regresji (test 1) i weryfikujemy ją statystyką o rozkładzie *F* Snedecora o 1 stopniu swobody licznika i 37 stopniach swobody mianownika.

Wartość empiryczna statystyki wynosi F = 3499,984, a odpowiadający jej krytyczny poziom istotności (istotność *F*) wynosi 3,01E-38 i jest mniejszy od przyjętego poziomu istotności $\alpha = 0{,}05$. Odrzucamy zatem hipotezę H_0 na korzyść H_1.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że podaż pieniądza zależy od czasu.

Istotność poszczególnych współczynników regresji. Podobnie jak w poprzednich modelach badamy istotność poszczególnych współczynników regresji (test 2). Weryfikację istotności dokonujemy na podstawie statystyki o rozkładzie *t* Studenta o 37 stopniach swobody.

Empiryczne wartości statystyk *t* Studenta wynoszą:

$t(\alpha_0)$ = 135,5233,

$t(\alpha_1)$ = 59,16066.

Odpowiadające im wartości krytycznego poziomu istotności (wartość-*p*) 1,68E-51 i 3,01E-38 są mniejsze od przyjętego poziomu istotności $\alpha = 0{,}05$.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że oba współczynniki modelu są istotnie różne od zera.

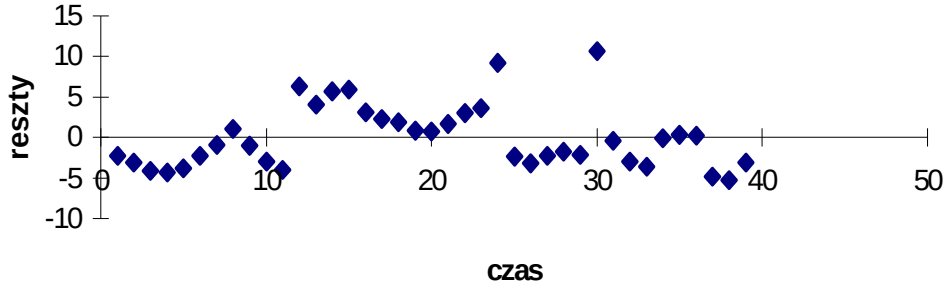
Analiza składników losowych modelu. Reszty modelu przedstawiono w tabeli 3.31 i na rysunku 3.17.

Tabela 3.31. Reszty modelu podaży pieniądza w Polsce

Obserwacja	Przewidywane podaż pieniądza[mld zł]	Składniki resztowe	Std. składniki resztowe
1	177,911	–2,21103	–0,56792
2	181,232	–3,03204	–0,7788
3	184,553	–4,15305	–1,06674
4	187,8741	–4,27406	–1,09782
5	191,1951	–3,79507	–0,97479
6	194,5161	–2,21609	–0,56922
7	197,8371	–0,9371	–0,2407
8	201,1581	1,041889	0,267616
9	204,4791	–0,97912	–0,25149
10	207,8001	–3,00013	–0,7706
11	211,1211	–4,02115	–1,03286
12	214,4422	6,357841	1,633051
13	217,7632	4,036829	1,036884
14	221,0842	5,715816	1,468143
15	224,4052	5,894804	1,514117
16	227,7262	3,073792	0,789523

cd. tabeli 3.31

17	231,0472	2,25278	0,578641
18	234,3682	1,831768	0,470501
19	237,6892	0,810756	0,208248
20	241,0103	0,789744	0,202851
21	244,3313	1,668731	0,428624
22	247,6523	3,047719	0,782826
23	250,9733	3,626707	0,931542
24	254,2943	9,205695	2,36454
25	257,6153	–2,31532	–0,5947
26	260,9363	–3,13633	–0,80559
27	264,2573	–2,25734	–0,57981
28	267,5784	–1,77835	–0,45678
29	270,8994	–2,19937	–0,56492
30	274,2204	10,67962	2,743127
31	277,5414	–0,44139	–0,11337
32	280,8624	–2,9624	–0,76091
33	284,1834	–3,58341	–0,92042
34	287,5044	–0,10443	–0,02682
35	290,8254	0,374561	0,096208
36	294,1465	0,253549	0,065126
37	297,4675	–4,86746	–1,25024
38	300,7885	–5,28848	–1,35838
39	304,1095	–3,10949	–0,79869



Rys. 3.17. Reszty trendu liniowego podaży pieniądza w Polsce

NORMALNOŚĆ

Stawiamy hipotezę H_0: składniki losowe ε mają rozkład $N(0, 3{,}94549)$. Zweryfikujemy tę hipotezę testem Shapiro–Wilka (test 5).

Empiryczna wartość statystyki W wynosi 0,9178. Dla poziomu istotności $\alpha = 0{,}05$ wartość krytyczna $W_\alpha = 0{,}917$. Ponieważ $W > W_\alpha$, nie ma więc podstaw do odrzucenia hipotezy H_0.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że składniki losowe mają rozkład normalny $N(0, S_\varepsilon = 3{,}94549)$ (należy zwrócić uwagę na zmianę poziomu istotności).

AUTOKORELACJA

Stawiamy hipotezę zerową o braku autokorelacji, wobec hipotezy alternatywnej o występowaniu autokorelacji dodatniej, i weryfikujemy ją testem Durbina–Watsona (test 7).

Empiryczna wartość statystyki $d = 1{,}135034$. Wartości krytyczne $d_L = 1{,}43$ oraz $d_U = 1{,}54$. Odrzucamy zatem hipotezę H_0: $\rho_1 = 0$ na korzyść hipotezy alternatywnej H_1: $\rho_1 > 0$.

Wniosek. Nie ma podstaw do odrzucenia hipotezy o istnieniu autokorelacji składników losowych rzędu pierwszego.

Podsumowanie. Analizowany model ekonometryczny nie jest poprawny statystycznie, gdyż reszty modelu wykazują autokorelację pierwszego rzędu.

Krok III′. Ponowny wybór klasy modelu

Jeżeli przyjrzymy się wykresowi reszt trendu liniowego (rys. 3.16), to zauważymy, że podaż pieniądza podlega wahaniom w cyklu rocznym. Autokorelacja oraz wahania reszt sugerują następujący model ekonometryczny ze zmiennymi opóźnionymi w czasie:

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-12} + \varepsilon_t ,$$

gdzie: t – czas,
y_t – podaż pieniądza w okresie t,
y_{t-1} – podaż pieniądza w okresie $t - 1$,
y_{t-12} – podaż pieniądza w okresie $t - 12$.

Model ten należy do klasy modeli dynamicznych autoregresyjnych. Zmienna objaśniająca y_{t-1} opisuje wpływ wielkości podaży pieniądza z poprzedniego okresu (miesiąca), a zmienna y_{t-12} wpływ podaży pieniądza rok wcześniej na jej aktualną wielkość.

Krok IV′. Estymacja parametrów strukturalnych

W tabeli 3.32 przedstawiono dane w nowym układzie.

Tabela 3.32. Dane do modelu ze zmiennymi opóźnionymi

Data (t)	Podaż pieniądza (t)	Podaż pieniądza ($t-1$)	Podaż pieniądza ($t-12$)
styczeń 99	221,8	220,8	175,7
luty 99	226,8	221,8	178,2
marzec 99	230,3	226,8	180,4
kwiecień 99	230,8	230,3	183,6
maj 99	233,3	230,8	187,4
czerwiec 99	236,2	233,3	192,3
lipiec 99	238,5	236,2	196,9
sierpień 99	241,8	238,5	202,2
wrzesień 99	246	241,8	203,5
październik 99	250,7	246	204,8
listopad 99	254,6	250,7	207,1
grudzień 99	263,5	254,6	220,8
styczeń 00	255,3	263,5	221,8
luty 00	257,8	255,3	226,8
marzec 00	262	257,8	230,3
kwiecień 00	265,8	262	230,8
maj 00	268,7	265,8	233,3
czerwiec 00	284,9	268,7	236,2
lipiec 00	277,1	284,9	238,5
sierpień 00	277,9	277,1	241,8
wrzesień 00	280,6	277,9	246
październik 00	287,4	280,6	250,7
listopad 00	291,2	287,4	254,6
grudzień 00	294,4	291,2	263,5
styczeń 01	292,6	294,4	255,3
luty 01	295,5	292,6	257,8
marzec 01	301	295,5	262

Wyniki estymacji parametrów strukturalnych modelu $y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-12} + \varepsilon_t$ przedstawiono na wydruku:

Statystyki regresji	
Wielokrotność R	0,988373
R kwadrat	0,976881
Dopasowany R kwadrat	0,974955
Błąd standardowy	3,84178
Obserwacje	27

ANALIZA WARIANCJI

	df	*SS*	*MS*	*F*
Regresja	2	14967,64	7483,822	507,0589
Resztkowy	24	354,2226	14,75928	
Razem	26	15321,87		

	Współczynniki	*Błąd standardowy*	*t Stat*	*Wartość-p*
Przecięcie	43,67912	13,71655	3,184409	0,003988
Podaż pieniądza ($t-1$)	0,414576	0,173198	2,393653	0,024853
Podaż pieniądza ($t-12$)	0,499943	0,148984	3,355673	0,002629

Liczba obserwacji zmniejszyła się z 39 do 27 ze względu na zmienną y_{t-12} opóźnioną w czasie o 12 jednostek (miesięcy).

Estymowany model ekonometryczny przyjmuje postać:

$$\hat{podaż} = 43{,}67912 + 0{,}144576\,podaż_{t-1} + 0{,}499943\,podaż_{t-12}.$$

Krok V′. Weryfikacja modelu

Zbudowany model ekonometryczny zweryfikujemy na poziomie istotności $\alpha = 0{,}05$.

Dopasowanie modelu do danych empirycznych. Współczynnik dopasowania modelu wynosi $R^2 = 0{,}974955$ (współczynnik zbieżności $\varphi^2 = 2{,}5\%$).

Wniosek. Model wyjaśnia 97,5% zmienności badanej cechy. Świadczy to o dobrym dopasowaniu modelu do danych empirycznych.

Istotność układu współczynników regresji. Stawiamy hipotezę o nieistotności układu współczynników regresji(test 1) i zweryfikujemy ją statystyką *F*, która przy prawdziwości hipotezy zerowej, ma rozkład *F* Snedecora o 2 stopniach swobody licznika i 24 stopniach swobody mianownika.

Wartość empiryczna statystyki wynosi $F = 507{,}0589$, a odpowiadający jej krytyczny poziom istotności (istotność *F*) wynosi 2,33E-20 i jest mniejszy od przyjętego poziomu istotności $\alpha = 0{,}05$. Odrzucamy zatem hipotezę H_0 na korzyść H_1.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że podaż pieniądza w bieżącym okresie zależy od podaży pieniądza w przeszłości.

Istotność poszczególnych współczynników regresji. Dla każdego współczynnika modelu regresji ($j = 0, 1, 2$) stawiamy hipotezy (test 2):

H_0: $\alpha_j = 0$,

H_1: $\alpha_j \neq 0$.

Hipotezy weryfikujemy statystyką mającą rozkład *t* Studenta o 24 stopniach swobody.

Empiryczne wartości statystyki *t* Studenta wynoszą:
$t(\alpha_0)$ = 3,184409,
$t(\alpha_1)$ = 2,393653,
$t(\alpha_2)$ = 3,355673.

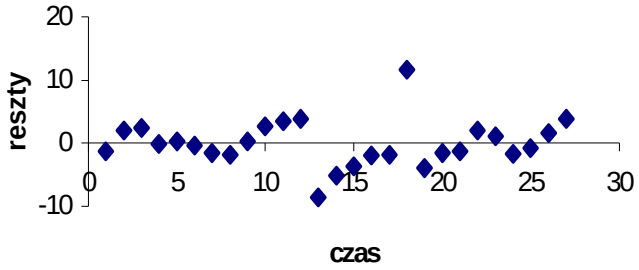
Odpowiadające im wartości krytycznego poziomu istotności (wartość-*p*) 0,003988, 0,024853, i 0,002629 są mniejsze od przyjętego poziomu istotności $\alpha = 0{,}05$.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że wszystkie trzy współczynniki modelu są istotnie różne od zera.

Analiza składników losowych modelu. Reszty modelu przedstawiono w tabeli 3.33 oraz na rysunku 3.18.

Tabela 3.33. Reszty modelu

Obserwacja	Przewidywane podaż pieniądza(t)	Składniki resztowe	Std. składniki resztowe
	223,0576	–1,25756	–0,34071
2	224,722	2,078002	0,562982
3	227,8948	2,405245	0,65164
4	230,9456	–0,14559	–0,03944
5	233,0527	0,247339	0,06701
6	236,5388	–0,33882	–0,0918
7	240,0408	–1,54083	–0,41745
8	243,6441	–1,84405	–0,4996
9	245,6621	0,337918	0,09155
10	248,0532	2,646771	0,717076
11	251,1516	3,448393	0,934255
12	259,6177	3,882327	1,051818
13	263,8073	-8,50735	–2,30485
14	262,9075	–5,10753	–1,38376
15	265,6938	–3,69377	–1,00073
16	267,685	–1,88497	–0,51068
17	270,5102	–1,81021	–0,49043
18	273,1623	11,73768	3,180027
19	281,0283	–3,92833	–1,06428
20	279,4444	–1,54444	–0,41843
21	281,8759	–1,27586	–0,34566
22	285,345	2,055047	0,556763
23	290,1138	1,08615	0,294265
24	296,1387	–1,73873	–0,47107
25	293,3658	–0,76585	–0,20749
26	293,8695	1,630535	0,441752
27	297,1715	3,828504	1,037236



Rys. 3.18. Reszty modelu autoregresyjnego podaży pieniądza w Polsce

NORMALNOŚĆ

Stawiamy hipotezę H_0: składniki losowe mają rozkład $N(0, 3{,}84178)$. Zweryfikujemy tę hipotezę testem Shapiro–Wilka (test 5).

Empiryczna wartość statystyki W wynosi 0,92933. Wartość krytyczna $W_\alpha = 0{,}923$. Ponieważ $W > W_\alpha$, nie ma więc podstaw do odrzucenia hipotezy H_0.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że składniki losowe modelu mają rozkład normalny $N(0; 3{,}84178)$.

AUTOKORELACJA

Skonstruowany model podaży pieniądza w Polsce jest modelem autoregresyjnym, w którym opóźniona zmienna objaśniana y jest zmienną objaśniającą y_{t-1}. Do zweryfikowania hipotezy o autokorelacji składników losowych modelu stosujemy test Durbina (test 9).

Stawiamy hipotezy:

H_0: $\rho(\varepsilon_t, \varepsilon_{t-1}) = 0$,

H_1: $\rho(\varepsilon_t, \varepsilon_{t-1}) \neq 0$.

Empiryczna wartość statystyki

$$h = \left(1 - \frac{1}{2}d\right)\sqrt{\frac{n}{1 - nS^2_{\alpha_{y(-1)}}}} = 0{,}69232,$$

gdzie: $d = 1{,}883827$; $S_{\alpha_{y(-1)}} = 0{,}173198$.

Wartość krytyczna statystyki dla $\alpha = 0{,}05$ wynosi 1,96. Empiryczna wartość statystyki $|h|$ jest zatem mniejsza od wartości krytycznej $|h| < u_\alpha$, nie ma więc podstaw do odrzucenia hipotezy H_0 o braku autokorelacji pierwszego rzędu na korzyść hipotezy H_1.

Ponieważ w modelu występuje zmienna y_{t-12}, dla zbadania zjawiska autokorelacji zweryfikujemy ponadto hipotezy (test 11):

H_0: brak autokorelacji,

$H_1 : \varepsilon_t = AR(12)$ (lub równoważnie: $H_1 : \varepsilon_t = \sum_{t=1}^{12} \gamma_\tau \varepsilon_{t-\tau}$).

Empiryczna wartość statystyki wynosi:

$$\chi^2 = \frac{e^T \mathbf{E}\left(\mathbf{E}^T\mathbf{E} - \mathbf{E}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{E}\right)^{-1}\mathbf{E}^T e}{S_\varepsilon^2} = 33{,}137\,.$$

Statystyka ta, przy prawdziwości hipotezy H_0, ma rozkład χ^2 o 12 stopniach swobody. Wyznaczona wartość empiryczna statystyki χ^2 = 33,137 jest mniejsza od wartości krytycznej χ_α^2 dla poziomu istotności α = 0,0009.

Wniosek. Na poziomie istotności α = 0,0009 nie ma podstaw do odrzucenia hipotezy o autokorelacji składników losowych.

SYMETRIA

Stawiamy hipotezę o symetrii reszt i weryfikujemy ją statystyką o rozkładzie *t* Studenta o 26 stopniach swobody (test 12). Empiryczna wartość statystyki *t* wynosi –0,58. Wartość krytyczna 2,056. Nie ma zatem podstaw do odrzucenia hipotezy H_0.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że rozkład składników losowych jest symetryczny.

LOSOWOŚĆ

Stawiamy hipotezę zerową H_0: reszty modelu są losowe.

Zweryfikujemy tę hipotezę testem serii (test 13), zliczając liczbę serii *L* tych samych znaków reszt modelu, która w tym przypadku wynosi 25.

Krytyczne wartości liczby serii dla 25 reszt dodatnich i 27 reszt ujemnych na przyjętym poziomie istotności α = 0,05 wynoszą 8 i 19. Nie ma zatem podstaw do odrzucenia hipotezy zerowej.

Wniosek. Nie ma podstaw do odrzucenia hipotezy o losowości reszt modelu.

HOMOSKEDASTYCZNOŚĆ

Stałość wariancji składników losowych w czasie sprawdzamy testem istotności współczynnika korelacji modułów reszt modelu i czasu (test 16).

Stawiamy hipotezy:

H_0: $\rho(|\varepsilon|, t) = 0$,

H_1: $\rho(|\varepsilon|, t) \neq 0$.

Sprawdzianem zespołu hipotez jest statystyka

$$t = \frac{r\left(|\varepsilon|, t\right)}{\sqrt{1-r^2}}\sqrt{n-2}\,,$$

gdzie $r(|\varepsilon|, t) = \dfrac{\sum\left(|e_t| - |\bar{e}|\right)(t-\bar{t})}{\sqrt{\sum\left(|e_t| - |\bar{e}|\right)^2 \sum (t-\bar{t})^2}}$.

Statystyka ta, przy prawdziwości hipotezy H_0, ma rozkład *t* Studenta o (*n*–2)-stopniach swobody.

W naszym przykładzie $r(|\varepsilon|,t)=0{,}137$, zatem $t = 0{,}69$. Wartość krytyczna $|t_\alpha| = 2{,}0595$.

Wniosek. Nie ma zatem podstaw do odrzucenia hipotezy o jednorodności wariancji składników losowych w czasie.

Podsumowanie. Przeprowadzona weryfikacja statystyczna świadczy o poprawności modelu:

$$\hat{podaż} = 43{,}67912 + 0{,}144576\, podaż_{t-1} + 0{,}499943\, podaż_{t-12}$$

Należy pamiętać, że hipotezę o braku autokorelacji składników losowych modelu przyjęliśmy na poziomie istotności $\alpha = 0{,}0009$.

Krok VI. Wnioskowanie na podstawie modelu

Na podstawie modelu:

$$\hat{podaż} = 43{,}67912 + 0{,}144576\, podaż_{t-1} + 0{,}499943\, podaż_{t-12}$$

wykonamy prognozę podaży pieniądza na kolejny rok (tabela 3.34).

Średni względny błąd prognozy wynosi:

$$\psi = \frac{1}{12}\sum_{t=1}^{12}\frac{|y_t - \hat{y}_t|}{y_t}100\% = 1{,}36\%,$$

a maksymalny względny błąd 3,69%.

Tabela 3.34. Błędy predykcji

Data	Rzeczywista podaż pieniądza	Predykcja	Błąd względny
kwiecień 01	303,0	301,4	0,54%
maj 01	305,0	303,6	0,45%
czerwiec 01	307,6	312,6	1,61%
lipiec 01	314,6	309,7	1,55%
sierpień 01	318,5	313,0	1,71%
wrzesień 01	320,7	316,0	1,46%
październik 01	324,7	320,3	1,35%
listopad 01	326,3	323,9	0,74%
grudzień 01	334,7	326,1	2,56%
styczeń 02	328,5	328,7	0,07%
luty 02	329,5	327,6	0,58%
marzec 02	319,0	330,8	3,69%

Małe błędy prognoz świadczą o przydatności skonstruowanego modelu w prognozowaniu wielkości podaży pieniądza.

3.6. Stopa bezrobocia

Model ekonometryczny stopy bezrobocia w Polsce jest modelem nieliniowym autoregresyjnym. Podobnie jak model podaży pieniądza zbudowany został na podstawie danych z okresu od stycznia do marca 2001. Predykcja dla kolejnego roku charakteryzuje się średnim błędem względnym na poziomie 0,79%, wobec maksymalnego względnego błędu na poziomie 1,52%. W pierwszym etapie modelowania skonstruowano trend liniowy. Analiza składników losowych trendu wskazała na wyraźny sezonowy charakter badanego zjawiska. Końcowy model zawiera funkcję harmoniczną oraz zmienne opóźnione w czasie.

Krok I. Określenie celu badań modelowych

Bezrobocie – zjawisko gospodarcze polegające na tym, że pewna część ludzi zdolnych do pracy nie znajduje zatrudnienia. Jego miarą jest stopa bezrobocia: relacja liczby bezrobotnych do liczby ludności w wieku produkcyjnym. Stopa bezrobocia to jeden z podstawowych wskaźników makroekonomicznych.

W celu właściwej dystrybucji środków finansowych dla bezrobotnych postanowiono określić trend bezrobocia w kraju. Z bieżących doświadczeń wiadomo, że bezrobocie wzrasta.

Krok II. Specyfikacja zmiennych wraz z gromadzeniem danych

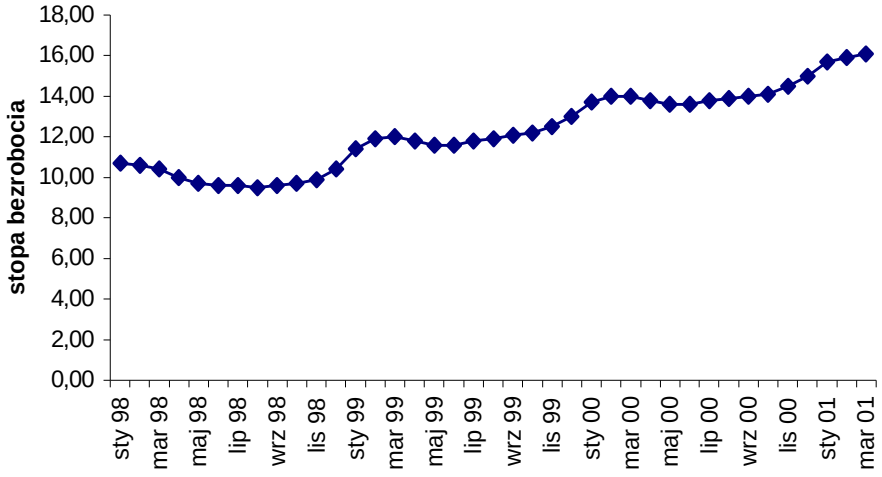
Dane o wielkości bezrobocia w Polsce są dostępne w Internecie na stronie http://www.money.pl/gospodarka/wskazniki/pkb/.

Zgromadzone dane o stopie bezrobocia w Polsce w okresie od stycznia 1998 do kwietnia 2001 przedstawiono w tabeli 3.35 i na rysunku 3.19.

Tabela 3.35. Stopa bezrobocia w Polsce

Czas	Stopa bezrobocia [%]	Czas	Stopa bezrobocia [%]
styczeń 98	10,70	styczeń 00	13,70
luty 98	10,60	luty 00	14,00
marzec 98	10,40	marzec 00	14,00
kwiecień 98	10,00	kwiecień 00	13,80
maj 98	9,70	maj 00	13,60
czerwiec 98	9,60	czerwiec 00	13,60
lipiec 98	9,60	lipiec 00	13,80
sierpień 98	9,50	sierpień 00	13,90
wrzesień 98	9,60	wrzesień 00	14,00
październik 98	9,70	październik 00	14,10
listopad 98	9,90	listopad 00	14,50
grudzień 98	10,40	grudzień 00	15,00
styczeń 99	11,40	styczeń 01	15,70
luty 99	11,90	luty 01	15,90
marzec 99	12,00	marzec 01	16,10
kwiecień 99	11,80		
maj 99	11,60		
czerwiec 99	11,60		
lipiec 99	11,80		
sierpień 99	11,90		
wrzesień 99	12,10		
październik 99	12,20		
listopad 99	12,50		
grudzień 99	13,00		

Źródło:http://www.money.pl/gospodarka/wskaźniki/pkb/



Rys. 3.19. Stopa bezrobocia w Polsce

Krok III. Wybór klasy modelu

Naszym celem jest wyznaczenie trendu bezrobocia. Za zmienną objaśnianą zatem przyjmiemy stopę bezrobocia, a za zmienną objaśniającą czas mierzony w skali bezwzględnej (kolumna 2 tabeli 3.35).

Wykres (rys. 3.18) wskazuje, że bezrobocie rośnie w czasie. Będziemy zatem wyznaczać model ekonometryczny postaci:

$$stopa\ \ bezrobocia = a_0 + a_1 czas + \varepsilon.$$

Krok IV. Estymacja parametrów strukturalnych

Wyniki estymacji liniowego modelu ekonometrycznego wartości stopy bezrobocia od czasu

$$stopa\ \ bezrobocia = \alpha_0 + \alpha_1 czas + \varepsilon$$

są następujące:

Statystyki regresji	
Wielokrotność R	0,951586
R kwadrat	0,905516
Dopasowany R kwadrat	0,902962
Błąd standardowy	0,60981
Obserwacje	39

ANALIZA WARIANCJI

	df	*SS*	*MS*	*F*	*Istotność F*
Regresja	1	131,8645	131,8645	354,5997	1,51E-20
Resztkowy	37	13,75913	0,371868		
Razem	38	145,6236			

	Współczynniki	*Błąd standardowy*	*t Stat*	*Wartość-p*	*Dolne 95%*	*Górne 95%*
Przecięcie	9,019568	0,199113	45,29879	5,15E-34	8,616128	9,423009
Czas	0,163381	0,008676	18,83082	1,51E-20	0,145801	0,18096

Równanie regresji przyjmuje zatem postać:

$$\widehat{stopa\ \ bezrobocia} = 9{,}019568 + 0{,}163381 czas.$$

Krok V. Weryfikacja modelu

Zbudowany model ekonometryczny zweryfikujemy na poziomie istotności $\alpha = 0{,}05$.

Dopasowanie modelu do danych empirycznych. Współczynnik dopasowania modelu wynosi $R^2 = 0{,}905516$ (współczynnik zbieżności $\varphi^2 = 9{,}5\%$).

Wniosek. Model wyjaśnia 90,5% zmienności badanej cechy. Świadczy to o dobrym dopasowaniu modelu do danych empirycznych.

Istotność układu współczynników regresji. Stawiamy hipotezę o istotności współczynników regresji (test 1) i weryfikujemy ją statystyką o rozkładzie *F* Snedecora o 1 stopniu swobody licznika i 37 stopniach swobody mianownika.

Wartość empiryczna statystyki wynosi $F = 354{,}5997$, a odpowiadający jej krytyczny poziom istotności (istotność *F*) wynosi 1,51E-20, jest mniejszy od przyjętego poziomu istotności $\alpha = 0{,}05$. Odrzucamy zatem hipotezę H_0 na korzyść H_1.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że stopa bezrobocia zależy od czasu.

Istotność poszczególnych współczynników regresji. Istotność współczynników regresji weryfikujemy statystyką o rozkładzie *t* Studenta o 37 stopniach swobody (test 2).

Empiryczne wartości statystyk *t* Studenta wynoszą:

$t(\alpha_0) = 45{,}29879$

$t(\alpha_1) = 18{,}83082$.

Odpowiadające im wartości krytycznego poziomu istotności (wartość-*p*) 5,15E-34 i 1,51E-20 są mniejsze od przyjętego poziomu istotności $\alpha = 0{,}05$.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że oba współczynniki modelu są istotnie różne od zera.

Analiza składników losowych modelu. Reszty modelu przedstawiono w tabeli 3.36.

Tabela 3.36. Reszty trendu liniowego stopy bezrobocia w Polsce

Obserwacja	Przewidywana stopa bezrobocia	Składniki resztowe	Std. składniki resztowe
1	9,182949	1,517051	2,521137
2	9,346329	1,253671	2,083434
3	9,50971	0,89029	1,479544
4	9,67309	0,32691	0,54328
5	9,836471	–0,13647	–0,2268
6	9,999852	–0,39985	–0,6645
7	10,16323	–0,56323	–0,93602
8	10,32661	–0,82661	–1,37372
9	10,48999	–0,88999	–1,47905
10	10,65337	–0,95337	–1,58438
11	10,81675	–0,91675	–1,52352
12	10,98013	–0,58013	–0,96411

cd. tabeli 3.36

13	11,14352	0,256484	0,426243
14	11,3069	0,593104	0,98566
15	11,47028	0,529723	0,88033
16	11,63366	0,166343	0,27644
17	11,79704	–0,19704	–0,32745
18	11,96042	–0,36042	–0,59897
19	12,1238	–0,3238	–0,53811
20	12,28718	–0,38718	–0,64344
21	12,45056	–0,35056	–0,58258
22	12,61394	–0,41394	–0,68791
23	12,77732	–0,27732	–0,46087
24	12,9407	0,059298	0,098546
25	13,10408	0,595918	0,990336
26	13,26746	0,732537	1,217379
27	13,43084	0,569157	0,945862
28	13,59422	0,205776	0,341972
29	13,7576	–0,1576	–0,26192
30	13,92099	–0,32099	–0,53343
31	14,08437	–0,28437	–0,47258
32	14,24775	–0,34775	–0,57791
33	14,41113	–0,41113	–0,68324
34	14,57451	–0,47451	–0,78857
35	14,73789	–0,23789	–0,39534
36	14,90127	0,098731	0,164079
37	15,06465	0,635351	1,055869
38	15,22803	0,67197	1,116725
39	15,39141	0,70859	1,177582

NORMALNOŚĆ

Stawiamy hipotezę H_0: składniki losowe mają rozkład $N(0, 0{,}60981)$. Zweryfikujemy tę hipotezę testem Shapiro–Wilka (test 5). Empiryczna wartość statystyki W wynosi 0,943908. Wartość krytyczna $W_\alpha = 0{,}939$. Ponieważ $W > W_\alpha$, nie ma więc podstaw do odrzucenia testowanej hipotezy.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że składniki losowe modelu mają rozkład normalny $N(0, S_\varepsilon = 0{,}60981)$.

AUTOKORELACJA

Stawiamy hipotezy (test 7):

H_0: $\rho_1 = 0$,

H_1: $\rho_1 > 0$,

gdzie ρ_1 – współczynnik autokorelacji rzędu pierwszego.

Wyznaczamy empiryczną wartość statystyki Durbina–Watsona. Empiryczna wartość statystyki d = 0,244916. Wartości krytyczne d_L = 1,43 oraz d_U = 1,54. Odrzucamy zatem hipotezę H_0: $\rho_1 = 0$ na korzyść hipotezy alternatywnej H_1: $\rho_1 > 0$.

Wniosek. Nie ma podstaw do odrzucenia hipotezy o istnieniu autokorelacji składników losowych rzędu pierwszego.

SYMETRIA

Stawiamy hipotezę o symetrii reszt modelu i testujemy ją statystyką, która, przy prawdziwości hipotezy H_0, ma rozkład t Studenta o 38 stopniach swobody (test 12). Empiryczna wartość statystyki wynosi –0,48779. Wartość krytyczna 2,02. Nie ma zatem podstaw do odrzucenia hipotezy H_0.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że rozkład składników losowych jest symetryczny.

LOSOWOŚĆ

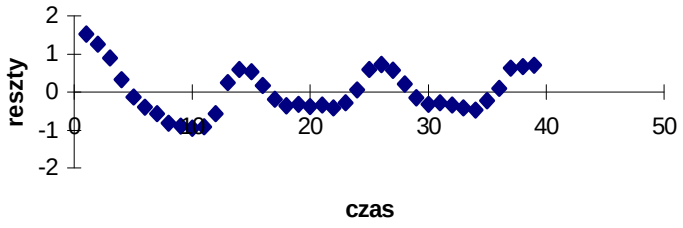
Stawiamy hipotezę zerową H_0: reszty modelu są losowe.

Zweryfikujemy tę hipotezę testem serii (test 13), zliczając liczbę serii L tych samych znaków reszt w modelu. Strukturę reszt dobrze obrazuje wykres (rys. 3.20).

Na wykresie widać wyraźnie, że reszty oscylują wokół zera, tworząc L = 7 serii. Wartości krytyczne testu serii dla 17 reszt dodatnich i 22 ujemnych na przyjętym poziomie istotności α = 0,05 aproksymujemy rozkładem normalnym N(20,17; 3,11), i otrzymujemy unormowaną liczbę serii:

$$L' = \frac{7 - 20{,}17}{3{,}11} = -4{,}24,$$

$$L_1 = -1{,}96 \cdot 3{,}11 + 20{,}17 = 14{,}1\,;\ L_2 = 1{,}96 \cdot 3{,}11 + 20{,}17 = 26{,}3\,.$$



Rys. 3.20. Reszty trendu liniowego stopy bezrobocia w Polsce

Empiryczna liczba serii $L = 7 < L_1 = 14{,}1$, a więc wpada do obszaru krytycznego.

Wniosek. Hipotezę o losowości reszt modelu należy odrzucić. W tym przypadku nielosowy rozkład reszt wynika z sezonowości stopy bezrobocia.

HOMOSKEDASTYCZNOŚĆ

Jednorodność wariancji składników losowych w czasie sprawdzimy testem istotności współczynnika korelacji modułów reszt modelu i czasu (test 16).

Hipotezę o homoskedastyczności składników losowych weryfikujemy statystyką, która, przy prawdziwości hipotezy H_0, ma rozkład t Studenta o $(n–2)$ stopniach swobody.

W naszym przykładzie $r(|\varepsilon|, t) = –0{,}42$, zatem $t = –2{,}83$. Wartość krytyczna $|t_\alpha| = 2{,}026$.

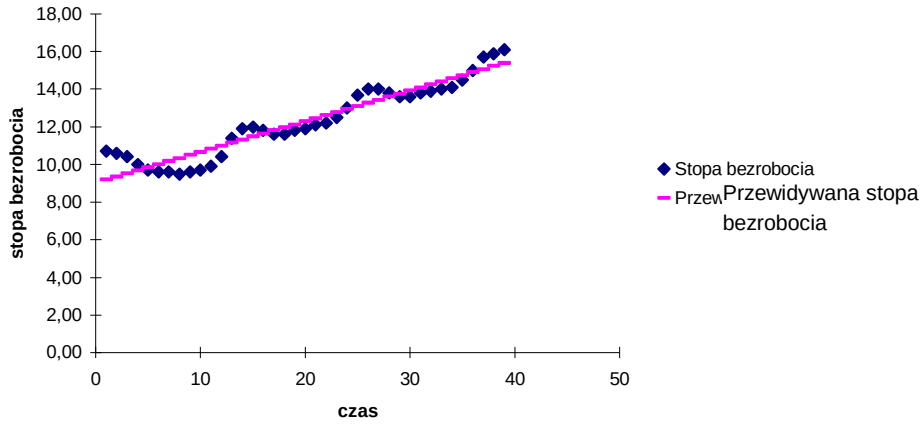
Wniosek. Odrzucamy hipotezę o równości wariancji składników losowych w czasie.

Krok VI. Ocena modelu

Przyjrzyjmy się wykresowi regresji. Widać na nim, że stopa bezrobocia w sezonie zimowym jest wyższa od trendu, a w okresach letnich niższa. W tym przypadku nielosowy rozkład reszt modelu jest związany z sezonowością stopy bezrobocia.

Na rysunku można zaobserwować także nierówność wariancji składników losowych w czasie, potwierdzoną testem na heteroskedastyczność składników losowych modelu. W roku 1998 wahania sezonowe były większe niż w latach następnych.

Badania nasze wykazały, że w badanym okresie stopa bezrobocia ma tendencję wzrostową (średni przyrost miesięczny to około 0,16% miesięcznie) z wahaniami sezonowymi. Należy zatem skonstruować model, który uwzględni sezonowość badanego zjawiska.



Rys. 3.2. Reszty trendu liniowego stopy bezrobocia w Polsce

Krok III′. Ponowny wybór klasy modelu

W sezonie letnim stopa bezrobocia wzrasta, w zimowym maleje. Długość cyklu wahań obejmuje 12 miesięcy (rys. 3.20, 3.21).

Za zmienne objaśniające przyjmiemy (tab. 3.37):

t – czas,

y_{t-1} – stopę bezrobocia w poprzednim miesiącu,

y_{t-12} – stopę bezrobocia w tym samym miesiącu rok wcześniej,

$\cos\left(\frac{2\pi}{12}t\right)$ – funkcję kosinus ze względu na harmoniczny charakter stopy bezrobocia.

Tabela 3.37. Dane do modelu nieliniowego

Data	y_t	t	y_{t-1}	y_{t-12}	$\cos\left(\frac{2\pi}{12}t\right)$
styczeń 99	11,40	1	10,40	10,7	0,866025
luty 99	11,90	2	11,40	10,6	0,5
marzec 99	12,00	3	11,90	10,40	0
kwiecień 99	11,80	4	12,00	10,00	–0,5
maj 99	11,60	5	11,80	9,70	–0,86603
czerwiec 99	11,60	6	11,60	9,60	–1
lipiec 99	11,80	7	11,60	9,60	–0,86603
sierpień 99	11,90	8	11,80	9,50	–0,5
wrzesień 99	12,10	9	11,90	9,60	0
październik 99	12,20	10	12,10	9,70	0,5
listopad 99	12,50	11	12,20	9,90	0,866025
grudzień 99	13,00	12	12,50	10,40	1
styczeń 00	13,70	13	13,00	11,40	0,866025
luty 00	14,00	14	13,70	11,90	0,5
marzec 00	14,00	15	14,00	12,00	0
kwiecień 00	13,80	16	14,00	11,80	–0,5
maj 00	13,60	17	13,80	11,60	–0,86603
czerwiec 00	13,60	18	13,60	11,60	–1
lipiec 00	13,80	19	13,60	11,80	–0,86603
sierpień 00	13,90	20	13,80	11,90	–0,5
wrzesień 00	14,00	21	13,90	12,10	0
październik 00	14,10	22	14,00	12,20	0,5
listopad 00	14,50	23	14,10	12,50	0,866025
grudzień 00	15,00	24	14,50	13,00	1
styczeń 01	15,70	25	15,00	13,70	0,866025
luty 01	15,90	26	15,70	14,00	0,5
marzec 01	16,10	27	15,90	14,00	0

Będziemy zatem estymować model liniowy postaci:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-12} + \alpha_3 \cos\left(\frac{\pi}{6}t\right) + \varepsilon$$

Etap IV′. Estymacja parametrów strukturalnych

Wyniki estymacji współczynników modelu liniowego

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-12} + \alpha_3 \cos\left(\frac{\pi}{6}t\right) + \varepsilon$$

są następujące:

Statystyki regresji	
Wielokrotność R	0,994951
R kwadrat	0,989928
Dopasowany R kwadrat	0,988615
Błąd standardowy	0,149257
Obserwacje	27

ANALIZA WARIANCJI

	df	*SS*	*MS*	*F*	*Istotność F*
Regresja	3	50,36169	16,78723	753,5475	4,27E-23
Resztkowy	23	0,512385	0,022278		
Razem	26	50,87407			

	Współczynniki	*Błąd standardowy*	*t Stat*	*Wartość-p*	*Dolne 95%*	*Górne 95%*
Przecięcie	0,886472	0,281668	3,147219	0,004512	0,303797	1,469146
Bezr ($t-1$)	0,726354	0,059303	12,24824	1,47E-11	0,603677	0,84903
bezr($t-12$)	0,256408	0,061142	4,193652	0,000347	0,129926	0,38289
$\cos\left(\frac{\pi}{6}t\right)$	0,238394	0,047085	5,063105	3,99E-05	0,140992	0,335795

Model ekonometryczny przyjmuje zatem postać:

$$\hat{y}_t = 0{,}886472 + 0{,}726354 y_{t-1} + 0{,}256408 y_{t-12} + 0{,}238394 \cos\left(\frac{\pi}{6}t\right).$$

Krok V′. Weryfikacja modelu

Zbudowany model ekonometryczny zweryfikujemy na poziomie istotności $\alpha = 0{,}05$.

Dopasowanie modelu do danych empirycznych. Odchylenie standardowe reszt S_e = 0,149257. Współczynnik dopasowania modelu R^2 = 0,988615 (współczynnik zbieżności φ^2 = 1,9%). Model wyjaśnia 98,1% zmienności badanej cechy. Ponieważ model jest nieliniowy, wyznaczymy ponadto wskaźnik średniego względnego dopasowania modelu:

$$\Psi = \frac{1}{n}\sum_{t=1}^{n}\frac{|E_t|}{|\hat{y}_t|}100\% = 0{,}9\%$$

Wniosek. Model jest dobrze dopasowany do danych empirycznych.

Istotność układu współczynników regresji. Stawiamy hipotezę o istotności układu współczynników regresji (test 1) i weryfikujemy ją statystyką, która, przy prawdziwości hipotezy zerowej, ma rozkład *F* Snedecora o 3 stopniach swobody licznika i 35 stopniach swobody mianownika.

Wartość empiryczna statystyki wynosi *F* = 753,5475, a odpowiadający jej krytyczny poziom istotności (istotność *F*) wynosi 4,274E-23 i jest mniejszy od przyjętego poziomu istotności $\alpha = 0{,}05$. Odrzucamy zatem hipotezę H_0 na korzyść H_1.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że stopa bezrobocia w bieżącym miesiącu *t* zależy przynajmniej od jednej ze zmiennych:

y_{t-1} – stopy bezrobocia w poprzednim miesiącu,
y_{t-12} – stopy bezrobocia w tym samym miesiącu rok wcześniej,
$\cos\left(\frac{2\pi}{12}t\right)$ – funkcja kosinusa.

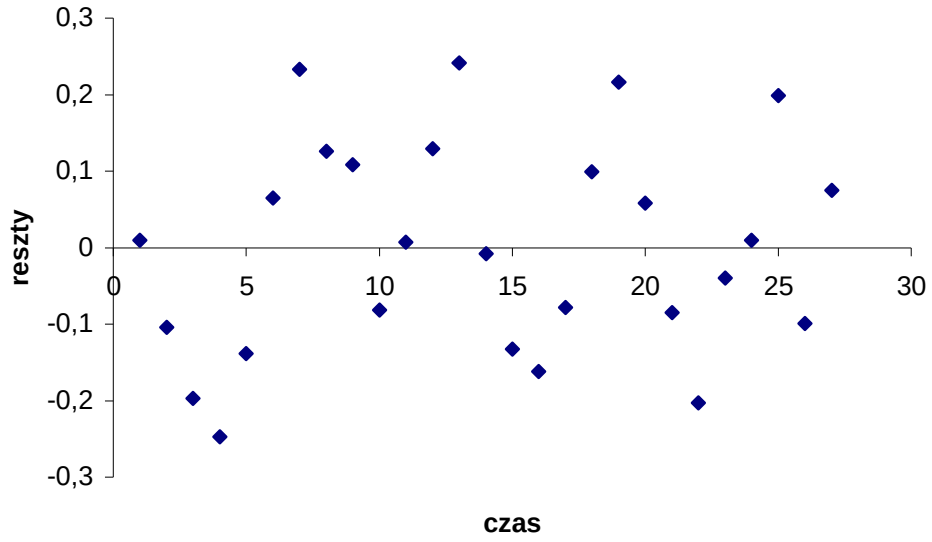
Istotność poszczególnych współczynników regresji. Dla każdego współczynnika modelu regresji stawiamy hipotezy dotyczące jego istotności (test 2) i weryfikujemy ją statystyką, która, przy prawdziwości hipotez zerowych, ma rozkład *t* Studenta o 37 stopniach swobody.

Empiryczne wartości statystyk *t* Studenta wynoszą:
$t(\alpha_0)$ = 3,147219,
$t(\alpha_1)$ = 12,24824,
$t(\alpha_2)$ = 4,193652,
$t(\alpha_3)$ = 5,063105.

Odpowiadające im wartości krytycznego poziomu istotności (wartość-*p*) wynoszą odpowiednio: 0,004512, 1,47E-35, 0,000347 oraz 3,99E-05 i są mniejsze od przyjętego poziomu istotności $\alpha = 0{,}05$.

Wniosek. Nie ma zatem podstaw do odrzucenia hipotezy, że wszystkie współczynniki modelu są istotnie różne od zera.

Analiza składników losowych modelu. Reszty modelu przedstawiono w tabeli 3.38 i na rysunku 3.22.



Rys. 3.22. Reszty zmodyfikowanego modelu bezrobocia

Tabela 3.38. Reszty modelu nieliniowego

Obserwacja	*Przewidywane bezrobocie(t)*	*Składniki resztowe*	*Std. składniki resztowe*
1	11,39057	0,009431	0,067178
2	12,00402	–0,10402	–0,74101
3	12,19672	–0,19672	–1,40134
4	12,0476	–0,2476	–1,76374
5	11,73815	–0,13815	–0,98407
6	11,5353	0,064704	0,460913
7	11,56723	0,232765	1,658085
8	11,77412	0,125877	0,896676
9	11,9916	0,108404	0,772208
10	12,2817	–0,0817	–0,58201
11	12,49288	0,007121	0,050723
12	12,87093	0,129072	0,919433
13	13,45857	0,241426	1,719777
14	14,00797	–0,00797	–0,05676
15	14,13232	–0,13232	–0,94255
16	13,96184	–0,16184	–1,15285
17	13,67803	–0,07803	–0,55583
18	13,50082	0,099181	0,706507
19	13,58404	0,215961	1,538377

cd. tabeli 3.38

20	13,84221	0,057791	0,411668
21	14,08532	–0,08532	–0,60779
22	14,3028	–0,2028	–1,4446
23	14,53961	–0,03961	–0,28217
24	14,9903	0,009704	0,069125
25	15,50102	0,19898	1,417419
26	15,99913	–0,09913	–0,70616
27	16,02521	0,074795	0,532794

NORMALNOŚĆ

Stawiamy hipotezę H_0: składniki losowe ma rozkład $N(0, 0{,}149257)$. Zweryfikujemy tę hipotezę testem Shapiro–Wilka (test 5).

Empiryczna wartość statystyki W wynosi 0,963803. Wartość krytyczna $W_\alpha = 0{,}923$. Ponieważ $W > W_\alpha$, nie ma więc podstaw do odrzucenia testowanej hipotezy.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że składniki losowe mają rozkład normalny $N(0; 0{,}149257)$.

AUTOKORELACJA

Model

$$\hat{y}_t = 0{,}886472 + 0{,}726354 y_{t-1} + 0{,}256408 y_{t-12} + 0{,}238394 \cos\left(\frac{\pi}{6} t\right)$$

jest modelem autoregresyjnym, w którym opóźniona zmienna objaśniana y jest zmienną objaśniającą. Dla zweryfikowania hipotezy o autokorelacji składników losowych modelu zastosujemy zatem test Durbina (test 9).

Stawiamy hipotezy

H_0: $\rho(\varepsilon_t, \varepsilon_{t-1}) = 0$,

H_1: $\rho(\varepsilon_t, \varepsilon_{t-1}) \neq 0$.

Empiryczna wartość statystyki:

$$h = \left(1 - \frac{1}{2} d\right) \sqrt{\frac{n}{1 - n S^2_{\alpha_{y(-1)}}}} = 2{,}434,$$

gdzie: $d = 1{,}10889$; $S_{a_{y(-1)}} = 0{,}059303$.

Wartość krytyczna statystyki dla $\alpha = 0{,}01$ wynosi 2,58. Empiryczna wartość statystyki $|h|$ jest zatem mniejsza od wartości krytycznej $|h| < u_\alpha$, więc nie ma podstaw do odrzucenia hipotezy H_0 o braku autokorelacji pierwszego rzędu na korzyść hipotezy H_1.

Ponieważ w modelu występuje zmienna y_{t-12}, dla zbadania zjawiska autokorelacji zweryfikujemy hipotezy (test 11):

H_0: brak autokorelacji,

$H_1 : \varepsilon_t = AR(12)$ (lub równoważnie: $H_1 : \varepsilon_t = \sum_{t=1}^{12} \gamma_\tau \varepsilon_{t-\tau}$).

Empiryczna wartość statystyki:

$$\chi^2 = \frac{e^T \mathbf{E}\left(\mathbf{E}^T\mathbf{E} - \mathbf{E}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{E}\right)^{-1}\mathbf{E}^T e}{s_e^2} = 18{,}1842 .$$

Statystyka ta, przy prawdziwości hipotezy H_0, ma rozkład χ^2 o 12 stopniach swobody.

Wartość krytyczna $\chi^2_\alpha = 21{,}026$. Wyznaczona wartość empiryczna statystyki χ^2 = 18,1842 jest mniejsza od wartości krytycznej $\chi^2_\alpha = 21{,}026$. Nie ma podstaw do odrzucenia hipotezy H_0 o braku autokorelacji na korzyść hipotezy H_1.

Wniosek. Nie ma podstaw do odrzucenia hipotezy o braku autokorelacji składników losowych.

SYMETRIA

Stawiamy hipotezę o symetrii składników losowych i weryfikujemy ją testem istotności (test 12), w którym statystyka testowa, przy prawdziwości hipotezy H_0, ma rozkład *t* Studenta o 26 stopniach swobody.

Empiryczna wartość statystyki wynosi 0,188982. Wartość krytyczna 2,052. Nie ma zatem podstaw do odrzucenia hipotezy H_0.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że rozkład składników losowych jest symetryczny.

LOSOWOŚĆ

Stawiamy hipotezę zerową H_0: reszty modelu są losowe. Zweryfikujemy tę hipotezę testem serii, zliczając liczbę serii *L* reszt tych samych znaków (test 13).

Empiryczna liczba serii wynosi L = 11.Wartości krytyczne testu serii dla 14 reszt dodatnich i 13 reszt ujemnych, na przyjętym poziomie istotności α = 0,05, wynoszą

$$L_1 = 8 \text{ oraz } L_2 = 19.$$

Spełniona zatem jest relacja, $L_1 = 8 < L = 11 < L_2 = 19$, a więc nie ma podstaw do odrzucenia hipotezy H_0.

Wniosek. Nie ma podstaw do odrzucenia hipotezy o losowości reszt modelu.

HOMOSKEDASTYCZNOŚĆ

Jednorodność wariancji składników losowych w czasie zweryfikujemy testem istotności współczynnika korelacji modułów reszt modelu i czasu (test 16). Hipotezę

tę weryfikujemy statystyką, która, przy prawdziwości hipotezy H_0, ma rozkład t Studenta o 25 stopniach swobody.

W naszym przykładzie $r(|\varepsilon|, t) = -0{,}13$. Zatem $t = -0{,}658$. Obszar krytyczny testu jest dwustronny. Wartość krytyczna statystyki wynosi $t_\alpha = 2{,}06$. Ponieważ $|t| < t_\alpha$, więc nie ma podstaw do odrzucenia hipotezy H_0 o stałości wariancji składników losowych modelu na korzyść hipotezy H_1.

Wniosek. Nie ma podstaw do odrzucenia hipotezy o homoskedastyczności składników losowych.

Podsumowanie. Model regresji

$$\hat{y}_t = 0{,}886472 + 0{,}726354 y_{t-1} + 0{,}256408 \cdot_{t-12} + 0{,}238394 \cos\left(\frac{\pi}{6} t\right)$$

możemy uznać za poprawny.

Krok VI. Wnioskowanie na podstawie modelu

Spróbujmy teraz na podstawie naszego modelu wyznaczyć prognozę stopy bezrobocia dla okresu od kwietnia 2001 do marca 2002 (tabela 3.39).

Średni względny błąd prognoz wynosi 0,79%, a maksymalny względny błąd 1,52%.

Małe błędy prognoz świadczą o przydatności skonstruowanego modelu w prognozowaniu wielkości stopy bezrobocia.

Tabela 3.39. Predykcja stopy bezrobocia i błędy predykcji

Data	Stopa bezrobocia	Predykcja	Reszta	Moduł reszty	Błąd względny, [%]
kwiecień 01	16,0	16,00	0,00	0,00	0,00
maj 01	15,9	15,79	0,11	0,11	0,70
czerwiec 01	15,9	15,68	0,22	0,22	1,36
lipiec 01	16,0	15,77	0,23	0,23	1,45
sierpień 01	16,2	15,95	0,25	0,25	1,52
wrzesień 01	16,3	16,24	0,06	0,06	0,35
październik 01	16,4	16,46	–0,06	0,06	0,37
listopad 01	16,8	16,72	0,08	0,08	0,46
grudzień 01	17,4	17,17	0,23	0,23	1,30
styczeń 02	18,0	17,76	0,24	0,24	1,35
luty 02	18,1	18,16	–0,06	0,06	0,31
marzec 02	18,1	18,16	–0,06	0,06	0,34

ROZDZIAŁ 4

MODELOWANIE EKONOMETRYCZNE W EXCELU

4.1. Studium przypadku: Frekwencja w czasie wyborów prezydenckich

Modelowanie zjawisk społecznych jest szczególnie trudne, choć jednocześnie, ze zrozumiałych względów wzbudza największe zainteresowanie. Przedstawimy model opisujący frekwencję w czasie wyborów prezydenckich jako jednorównaniowy model liniowy z wieloma zmiennymi objaśniającymi. W tym przykładzie chcemy pokazać jednocześnie jak można w tym celu wykorzystać arkusz kalkulacyjny Excel.

Krok I. Cel badań

Celem badań jest budowa modelu regresyjnego frekwencji w wyborach prezydenta RP umożliwiającego (w jakimś stopniu) prognozowanie frekwencji na podstawie danych socjoekonomicznych.

Na podstawie lektury artykułów, jakie ukazały się po wyborach w ogólnodostępnych publikacjach, takich jak *Gazeta Wyborcza*, *Wprost* czy *Polityka* wysuwa się przypuszczenie, iż na frekwencję w wyborach prezydenckich w poszczególnych miejscach w kraju mogły mieć wpływ następujące czynniki:

- czynniki osobiste wyborcy:
  - wiek,
  - wykształcenie,
  - stosunek do religii,
  - zawód,
  - zainteresowanie kulturą,
  - przedsiębiorczość,
  - zamożność.

- czynniki makroekonomiczne:
  - odsetek osób niezatrudnionych,
  - warunki pogodowe w dniu wyborów (zachmurzenie, opady, temperatura),
  - gęstość zaludnienia,
  - liczba dzieci w przeciętnej rodzinie.

Wybór tych czynników jest do pewnego stopnia arbitralny, a wśród kryteriów wyboru niebagatelną rolę gra dostępność danych, które w tym przypadku można uzyskać z danych Państwowej Komisji Wyborczej (frekwencja) oraz Banku Danych Lokalnych GUS i IMGW.

Zauważmy ponadto, że w żadnej mierze nie zajmujemy się takimi czynnikami, jak program wyborczy kandydatów, czy też ogólniej, ich osobiste walory, upatrując przyczyn takiej, a nie innej frekwencji jedynie w czynnikach na swój sposób „ubocznych".

Krok II. Specyfikacja zmiennych wraz z gromadzeniem danych

Przyjęto, że dane zbierane będą z 373 powiatów.

Po uzyskaniu z Instytutu Meteorologii i Gospodarki Wodnej (w formie Codziennego Biuletynu Meteorologicznego IMGW) poglądowych danych o warunkach pogodowych w dniu wyborów stwierdzono, iż warunki pogodowe nie różniły się istotnie w poszczególnych regionach kraju – zaniechano zatem uzyskiwania szczegółowych danych w tym zakresie..

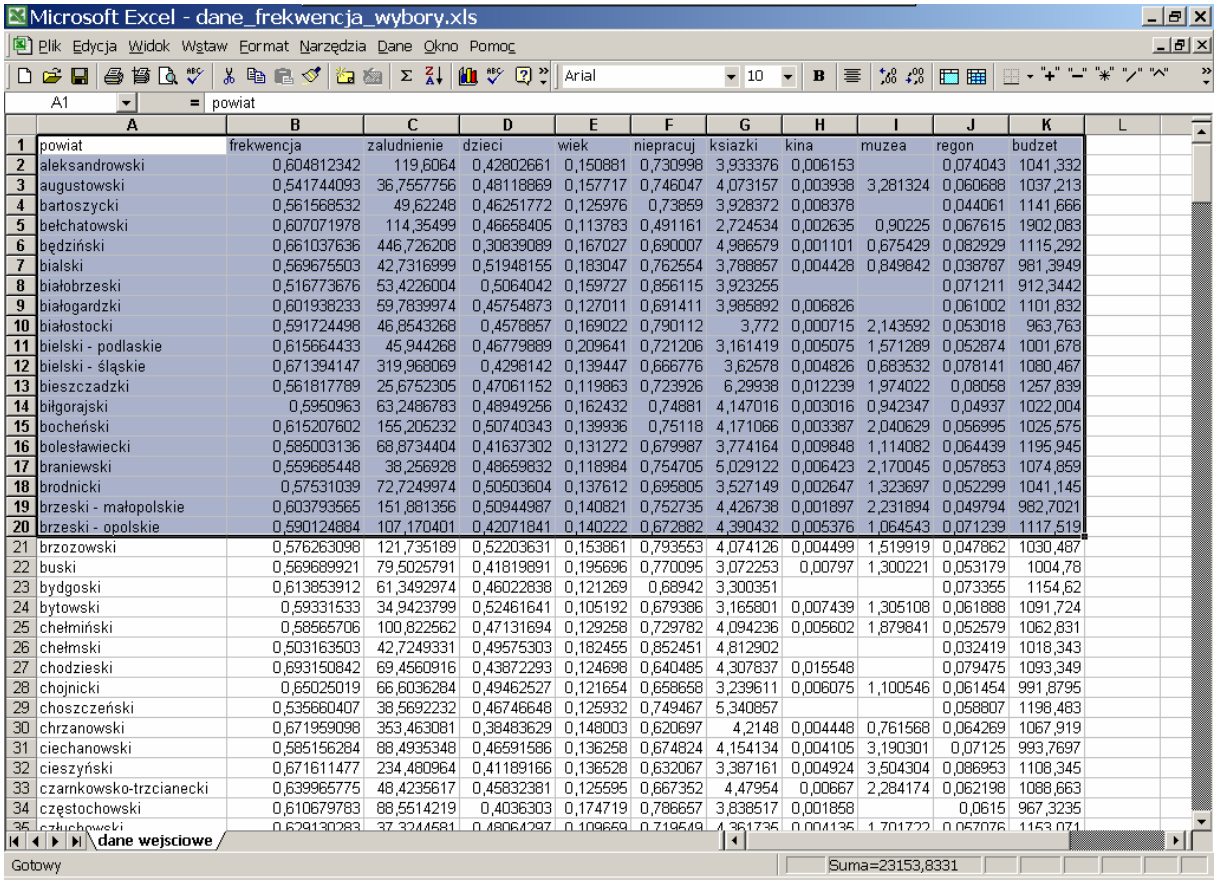
W odniesieniu do danych charakteryzujących wyborców w poszczególnych powiatach uzyskano następujące dane:

- wiek wyborców, jako procentowy udział ludności w wieku poprodukcyjnym,
- zainteresowanie wyborców kulturą, jako przypadającą na jednego mieszkańca powiatu liczbą woluminów w bibliotekach, liczbą miejsc w kinach oraz liczbą muzeów (w tym przypadku na każde 100 000 mieszkańców),
- przedsiębiorczość wyborców, jako liczba jednostek gospodarczych zarejestrowanych w systemie REGON przypadającą na jednego mieszkańca powiatu,
- zamożność wyborców mierzona jest przypadającym na jednego mieszkańca dochodem budżetów gmin wchodzących do danego powiatu.

W odniesieniu do danych makroekonomicznych odnoszących się do powiatu:

- odsetek osób niezatrudnionych, jako stosunek liczby osób niezatrudnionych do liczby osób w wieku produkcyjnym,
- gęstość zaludnienia,
- stosunek liczby osób w wieku przedprodukcyjnym do liczby osób w wieku produkcyjnym („dzieci").

Przykładowe zestawienie danych dla kilkunastu powiatów pokazano na wydrukach.



Microsoft Excel - dane_frekwencja_wybory.xls

Plik Edycja Widok Wstaw Format Narzędzia Dane Okno Pomoc

Arial 10

A1 = powiat

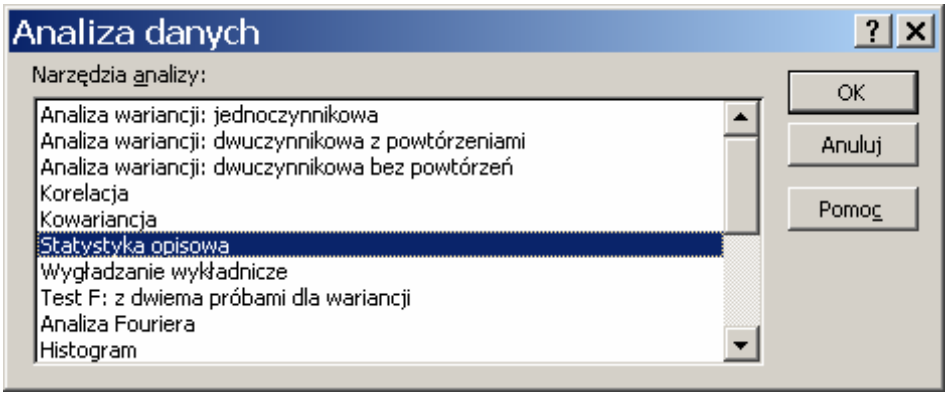
	A	B	C	D	E	F	G	H	I	J	K	L
1	powiat	frekwencja	zaludnienie	dzieci	wiek	niepracuj	ksiazki	kina	muzea	regon	budzet	
2	aleksandrowski	0,604812342	119,6064	0,42802661	0,150881	0,730998	3,933376	0,006153		0,074043	1041,332	
3	augustowski	0,541744093	36,7557756	0,48118869	0,157717	0,746047	4,073157	0,003938	3,281324	0,060688	1037,213	
4	bartoszycki	0,561568532	49,62248	0,46251772	0,125976	0,73859	3,928372	0,008378		0,044061	1141,666	
5	bełchatowski	0,607071978	114,35499	0,46658405	0,113783	0,491161	2,724534	0,002635	0,90225	0,067615	1902,083	
6	będziński	0,661037636	446,726208	0,30839089	0,167027	0,690007	4,986579	0,001101	0,675429	0,082929	1115,292	
7	bialski	0,569675503	42,7316999	0,51948155	0,183047	0,762554	3,788857	0,004428	0,849842	0,038787	981,3949	
8	białobrzeski	0,516773676	53,4226004	0,5064042	0,159727	0,856115	3,923255			0,071211	912,3442	
9	białogardzki	0,601938233	59,7839974	0,45754873	0,127011	0,691411	3,985892	0,006826		0,061002	1101,832	
10	białostocki	0,591724498	46,8543268	0,4578857	0,169022	0,790112	3,772	0,000715	2,143592	0,053018	963,763	
11	bielski - podlaskie	0,615664433	45,944268	0,46779889	0,209641	0,721206	3,161419	0,005075	1,571289	0,052874	1001,678	
12	bielski - śląskie	0,671394147	319,968069	0,4298142	0,139447	0,666776	3,62578	0,004826	0,683532	0,078141	1080,467	
13	bieszczadzki	0,561817789	25,6752305	0,47061152	0,119863	0,723926	6,29938	0,012239	1,974022	0,08058	1257,839	
14	biłgorajski	0,5950963	63,2486783	0,48949256	0,162432	0,74881	4,147016	0,003016	0,942347	0,04937	1022,004	
15	bocheński	0,615207602	155,205232	0,50740343	0,139936	0,75118	4,171066	0,003387	2,040629	0,056995	1025,575	
16	bolesławiecki	0,585003136	68,8734404	0,41637302	0,131272	0,679987	3,774164	0,009848	1,114082	0,064439	1195,945	
17	braniewski	0,559685448	38,256928	0,48659832	0,118984	0,754705	5,029122	0,006423	2,170045	0,057853	1074,859	
18	brodnicki	0,57531039	72,7249974	0,50503604	0,137612	0,695805	3,527149	0,002647	1,323697	0,052299	1041,145	
19	brzeski - małopolskie	0,603793565	151,881356	0,50944987	0,140821	0,752735	4,426738	0,001897	2,231894	0,049794	982,7021	
20	brzeski - opolskie	0,590124884	107,170401	0,42071841	0,140222	0,672882	4,390432	0,005376	1,064543	0,071239	1117,519	
21	brzozowski	0,576263098	121,735189	0,52203631	0,153861	0,793553	4,074126	0,004499	1,519919	0,047862	1030,487	
22	buski	0,569689921	79,5025791	0,41819891	0,195696	0,770095	3,072253	0,00797	1,300221	0,053179	1004,78	
23	bydgoski	0,613853912	61,3492974	0,46022838	0,121269	0,68942	3,300351			0,073355	1154,62	
24	bytowski	0,59331533	34,9423799	0,52461641	0,105192	0,679386	3,165801	0,007439	1,305108	0,061888	1091,724	
25	chełmiński	0,58565706	100,822562	0,47131694	0,129258	0,729782	4,094236	0,005602	1,879841	0,052579	1062,831	
26	chełmski	0,503163503	42,7249331	0,49575303	0,182455	0,852451	4,812902			0,032419	1018,343	
27	chodzieski	0,693150842	69,4560916	0,43872293	0,124698	0,640485	4,307837	0,015548		0,079475	1093,349	
28	chojnicki	0,65025019	66,6036284	0,49462527	0,121654	0,658658	3,239611	0,006075	1,100546	0,061454	991,8795	
29	choszczeński	0,535660407	38,5692232	0,46746648	0,125932	0,749467	5,340857			0,058807	1198,483	
30	chrzanowski	0,671959098	353,463081	0,38483629	0,148003	0,620697	4,2148	0,004448	0,761568	0,064269	1067,919	
31	ciechanowski	0,585156284	88,4935348	0,46591586	0,136258	0,674824	4,154134	0,004105	3,190301	0,07125	993,7697	
32	cieszyński	0,671611477	234,480964	0,41189166	0,136528	0,632067	3,387161	0,004924	3,504304	0,086953	1108,345	
33	czarnkowsko-trzcianecki	0,639965775	48,4235617	0,45832381	0,125595	0,667352	4,47954	0,00667	2,284174	0,062198	1088,663	
34	częstochowski	0,610679783	88,5514219	0,4036303	0,174719	0,786657	3,838517	0,001858		0,0615	967,3235	
35	człuchowski	0,629130283	37,3244581	0,48064297	0,109659	0,719549	4,361735	0,004135	1,701722	0,057076	1153,071	

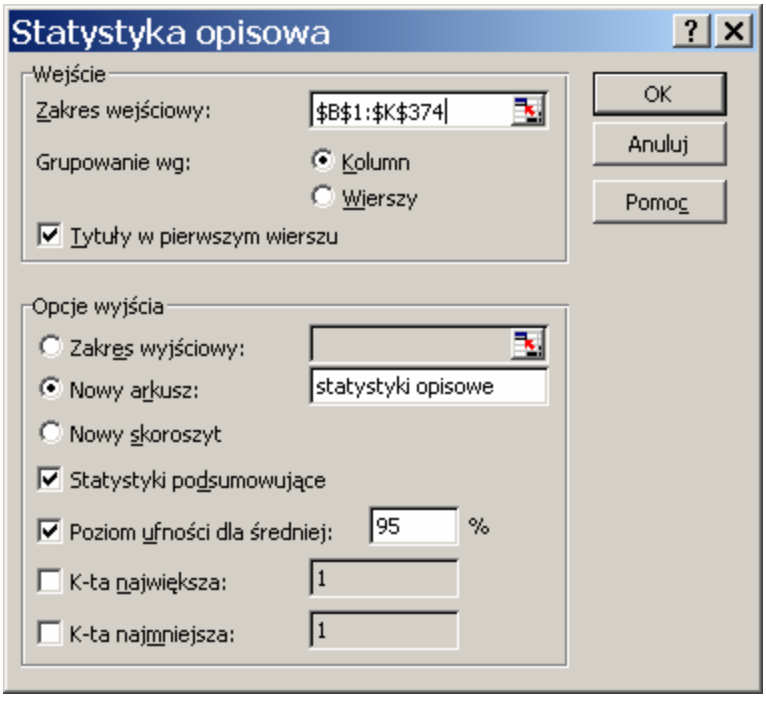
dane wejsciowe

Gotowy

Suma=23153,8331

Zgromadzone dane poddano pierwszej obróbce statystycznej, posługując się narzędziem zawartym w programie *Excel* (w opcji Narzędzia wybieramy Analiza danych):





Dodatkowo w arkuszu „statystyki opisowe” możliwe jest obliczenie wartości współczynnika zmienności dla każdej ze zmiennych. Jak widać, wszystkie zmienne opisujące mają zmienność powyżej 10%, a więc wykazują dostateczną zmienność, aby móc je użyć jako zmienne objaśniające, potencjalnie wnoszące coś do wyjaśnienia zjawiska.

frekwencja		*zaludnienie*		*dzieci*		*wiek*		*niepracuj*	
Średnia	0,595824	Średnia	410,6514	Średnia	0,440672	Średnia	0,141856	Średnia	0,682888
Błąd standardov	0,002592	Błąd standa	38,40265	Błąd standa	0,002978	Błąd standa	0,001242	Błąd standa	0,006104
Mediana	0,59428	Mediana	91,03432	Mediana	0,448743	Mediana	0,137846	Mediana	0,709549
Odchylenie stan	0,050066	Odchylenie	741,6784	Odchylenie	0,057524	Odchylenie	0,023992	Odchylenie	0,117881
Wariancja próbk	0,002507	Wariancja	550086,9	Wariancja	0,003309	Wariancja	0,000576	Wariancja	0,013896
Kurtoza	1,220898	Kurtoza	5,533108	Kurtoza	-0,2154	Kurtoza	0,578904	Kurtoza	2,694429
Skośność	-0,450147	Skośność	2,420206	Skośność	-0,404149	Skośność	0,359789	Skośność	-1,377152
Zakres	0,330648	Zakres	4435,671	Zakres	0,298989	Zakres	0,162657	Zakres	0,759517
Minimum	0,378771	Minimum	25,67523	Minimum	0,280739	Minimum	0,058062	Minimum	0,127027
Maksimum	0,70942	Maksimum	4461,346	Maksimum	0,579728	Maksimum	0,220719	Maksimum	0,886544
Suma	222,2422	Suma	153173	Suma	164,3706	Suma	52,91235	Suma	254,7173
Licznik	373	Licznik	373	Licznik	373	Licznik	373	Licznik	373
Poziom ufności(	0,005097	Poziom ufn	75,51359	Poziom ufn	0,005857	Poziom ufn	0,002443	Poziom ufn	0,012002
Zmienność	8,40%	Zmienność	180,61%	Zmienność	13,05%	Zmienność	16,91%	Zmienność	17,26%

ksiazki		*kina*		*muzea*		*regon*		*budzet*	
Średnia	3,77296	Średnia	0,005923	Średnia	2,01698	Średnia	0,068242	Średnia	1198,763
Błąd standa	0,042255	Błąd standa	0,000163	Błąd standa	0,073768	Błąd standa	0,001152	Błąd standa	15,72627
Mediana	3,762972	Mediana	0,005475	Mediana	1,725098	Mediana	0,063003	Mediana	1076,838
Odchylenie	0,816082	Odchylenie	0,002953	Odchylenie	1,22553	Odchylenie	0,022252	Odchylenie	303,7247
Wariancja	0,665989	Wariancja	8,72E-06	Wariancja	1,501923	Wariancja	0,000495	Wariancja	92248,71
Kurtoza	0,806119	Kurtoza	1,375155	Kurtoza	3,274629	Kurtoza	0,923834	Kurtoza	4,098196
Skośność	0,430513	Skośność	0,966765	Skośność	1,605786	Skośność	1,077015	Skośność	1,941262
Zakres	5,394942	Zakres	0,017356	Zakres	7,179566	Zakres	0,122492	Zakres	1964,772
Minimum	1,54071	Minimum	0,000715	Minimum	0,389883	Minimum	0,02754	Minimum	876,8918
Maksimum	6,935652	Maksimum	0,018071	Maksimum	7,56945	Maksimum	0,150032	Maksimum	2841,664
Suma	1407,314	Suma	1,948632	Suma	556,6866	Suma	25,45433	Suma	447138,6
Licznik	373	Licznik	329	Licznik	276	Licznik	373	Licznik	373
Poziom ufn	0,083089	Poziom ufn	0,00032	Poziom ufn	0,145222	Poziom ufn	0,002266	Poziom ufn	30,92357
Zmienność	21,63%	Zmienność	49,85%	Zmienność	60,76%	Zmienność	32,61%	Zmienność	25,34%

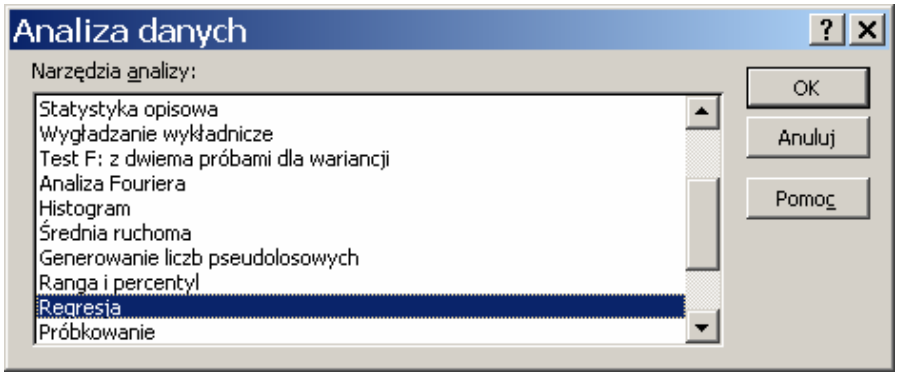
Krok III. Wybór modelu

Przyjmijmy, że budowany jest model liniowy o postaci:

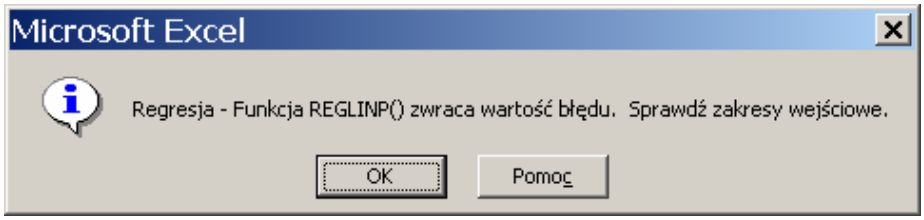
$$y_i = \sum_{j=0}^{k} \alpha_j x_{ij} + \varepsilon_i .$$

Krok IV. Estymacja parametrów strukturalnych

W celu estymacji współczynników regresji w arkuszu „dane wejściowe" uruchamiamy „analizę danych" w opcji „narzędzia" z funkcją regresji:



Uruchomienie tej funkcji powoduje wyświetlenie się komunikatu:



Jest to związane z tym, że niektóre ze zmiennych opisujących mają braki („puste miejsca") – brak jest danych w pojedynczych powiatach[9]. W takiej sytuacji możliwe są dwa rozwiązania:

- wobec faktu, że obserwacji jest bardzo dużo (373 powiaty) eliminujemy ze zbioru danych te powiaty, dla których brak danych,
- uzupełniamy brakujące dane (na przykład wpisujemy tam wartości średnie dla danej zmiennej) – ten sposób szczególnie wtedy jest polecany, kiedy zbiory obserwacji są mało liczne.

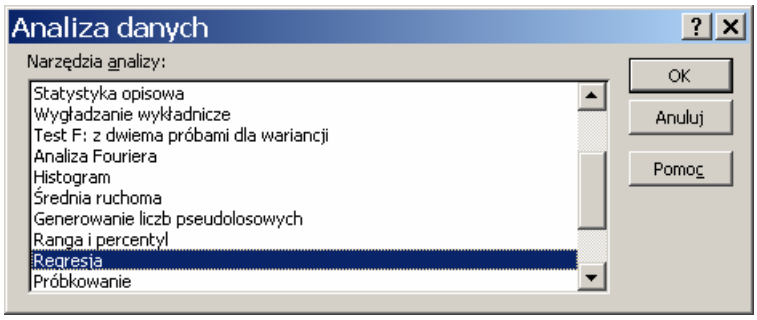
[9] Podobny wniosek można było wysnuć, obserwując parametr „licznik" w arkuszu „statystyki opisowe", gdzie wartości tego parametru różnią się dla poszczególnych parametrów.

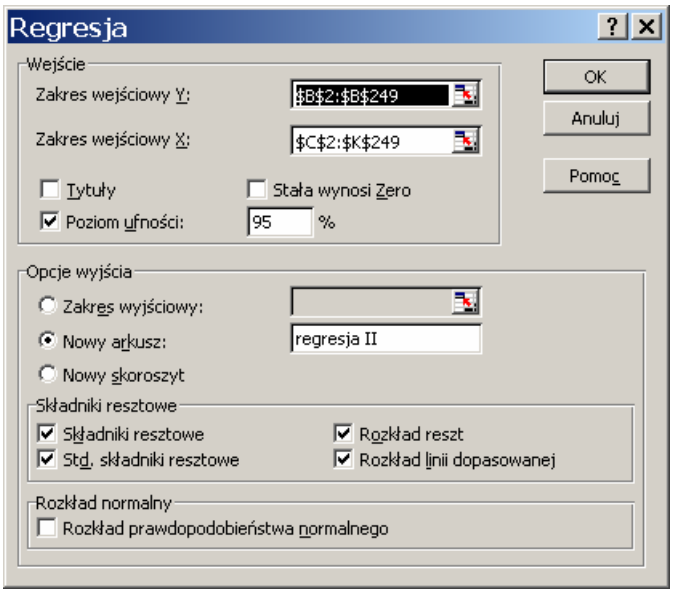
frekwencja		*zaludnienie*		*dzieci*		*wiek*		*niepracuj*	
Średnia	0,606124	Średnia	520,2509	Średnia	0,433953	Średnia	0,141277	Średnia	0,65769
Błąd standa	0,003028	Błąd standa	51,60665	Błąd standa	0,003744	Błąd standa	0,001447	Błąd standa	0,007905
Mediana	0,602895	Mediana	112,2045	Mediana	0,444089	Mediana	0,137494	Mediana	0,686052
Odchylenie	0,047678	Odchylenie	812,7023	Odchylenie	0,058968	Odchylenie	0,02278	Odchylenie	0,124483
Wariancja	0,002273	Wariancja	660485	Wariancja	0,003477	Wariancja	0,000519	Wariancja	0,015496
Kurtoza	0,929246	Kurtoza	2,085415	Kurtoza	-0,386177	Kurtoza	0,57099	Kurtoza	2,111256
Skośność	-0,434476	Skośność	1,788212	Skośność	-0,408313	Skośność	0,521799	Skośność	-1,309795
Zakres	0,306438	Zakres	3582,747	Zakres	0,298989	Zakres	0,13461	Zakres	0,726292
Minimum	0,402981	Minimum	25,67523	Minimum	0,280739	Minimum	0,086109	Minimum	0,127027
Maksimum	0,70942	Maksimum	3608,423	Maksimum	0,579728	Maksimum	0,220719	Maksimum	0,85332
Suma	150,3187	Suma	129022,2	Suma	107,6204	Suma	35,03671	Suma	163,1072
Licznik	248	Licznik	248	Licznik	248	Licznik	248	Licznik	248
Poziom ufn	0,005963	Poziom ufn	101,6452	Poziom ufn	0,007375	Poziom ufn	0,002849	Poziom ufn	0,015569
Zmienność	7,87%	Zmienność	156,21%	Zmienność	13,59%	Zmienność	16,12%	Zmienność	18,93%

ksiazki		*kina*		*muzea*		*regon*		*budzet*	
Średnia	3,676098	Średnia	0,005682	Średnia	1,967688	Średnia	0,072618	Średnia	1234,59
Błąd standa	0,050474	Błąd standa	0,000178	Błąd standa	0,076987	Błąd standa	0,001489	Błąd standa	20,90314
Mediana	3,621736	Mediana	0,005236	Mediana	1,673706	Mediana	0,067501	Mediana	1082,87
Odchylenie	0,794863	Odchylenie	0,002808	Odchylenie	1,212394	Odchylenie	0,023449	Odchylenie	329,183
Wariancja	0,631807	Wariancja	7,88E-06	Wariancja	1,469898	Wariancja	0,00055	Wariancja	108361,5
Kurtoza	0,977485	Kurtoza	0,575461	Kurtoza	4,070299	Kurtoza	0,18473	Kurtoza	2,519159
Skośność	0,40277	Skośność	0,772888	Skośność	1,758683	Skośność	0,878186	Skośność	1,592276
Zakres	5,394942	Zakres	0,015462	Zakres	7,179566	Zakres	0,113	Zakres	1964,772
Minimum	1,54071	Minimum	0,000715	Minimum	0,389883	Minimum	0,037032	Minimum	876,8918
Maksimum	6,935652	Maksimum	0,016177	Maksimum	7,56945	Maksimum	0,150032	Maksimum	2841,664
Suma	911,6724	Suma	1,409139	Suma	487,9867	Suma	18,00924	Suma	306178,4
Licznik	248	Licznik	248	Licznik	248	Licznik	248	Licznik	248
Poziom ufn	0,099414	Poziom ufn	0,000351	Poziom ufn	0,151635	Poziom ufn	0,002933	Poziom ufn	41,17115
Zmienność	21,62%	Zmienność	49,41%	Zmienność	61,62%	Zmienność	32,29%	Zmienność	26,66%

Po wyeliminowaniu niektórych obserwacji powstał nowy zbiór danych. Zawiera on już tylko 248 obserwacji wszystkich zmiennych. Taka operacja usuwania niektórych obserwacji powoduje konieczność powtórnego sprawdzenia parametrów zmienności poszczególnych zmiennych. Wyniki obliczeń przedstawiono na wydruku, s. 117.

Po powtórnym przeliczeniu żadna ze zmiennych objaśniających nie utraciła swoich właściwości wyjaśniających – zmienność każdej jest powyżej 10%. Powtórnie zatem przystępujemy do budowy modelu liniowego, choć tym razem na podstawie nowego zredukowanego zbioru obserwacji.





W wyniki estymacji MNK otrzymaliśmy następujące oceny współczynników modelu liniowego:

frekwencja = 0,6624 + 0,000004 zaludnienie + 0,2064 dzieci + 0,0004 wiek
– 0,2498 niepracujący + 0,0073 książki – 1,8967 kina – 0,0023 muzea
+ 1,0561 regon – 0,00006 budżet.

Statystyki regresji	
Wielokrotność R	0,627289731
R kwadrat	0,393492406
Dopasowany R kwadrat	0,370557245
Błąd standardowy	0,037826309
Obserwacje	248

ANALIZA WARIANCJI

	df	*SS*	*MS*	*F*	*Istotność F*
Regresja	9	0,22093524	0,02454836	17,15673174	9,69812E-22
Resztkowy	238	0,34053745	0,00143083		
Razem	247	0,56147269			

	Współczynniki	*Błąd standardowy*	*t Stat*	*Wartość-p*	*Dolne 95%*	*Górne 95%*
Przecięcie	0,662399446	0,05742966	11,5341001	9,65967E-25	0,549263951	0,775534941
zaludnienie	4,44489E-06	5,31552E-06	0,83621094	0,403875003	-6,02659E-06	1,49164E-05
dzieci	0,20639944	0,071452194	2,88863685	0,004225869	0,065639782	0,347159098
wiek	0,0003917	0,125551325	0,00311984	0,997513347	-0,246942365	0,247725765
niepracuj	-0,249782294	0,037375017	-6,68313525	1,64476E-10	-0,323410469	-0,17615412
ksiazki	0,007261591	0,0035502	2,04540355	0,041914863	0,000267755	0,014255427
kina	-1,896668837	0,957900181	-1,98002764	0,048852345	-3,783716593	-0,00962108
muzea	-0,002263574	0,002144399	-1,05557508	0,292232726	-0,006488006	0,001960857
regon	1,056148654	0,188568614	5,60087192	5,8677E-08	0,684671556	1,427625752
budzet	-5,83923E-05	1,58151E-05	-3,69217681	0,000275745	-8,95479E-05	-2,7237E-05

Krok V. Weryfikacja modelu

Zbudowany model ekonometryczny zweryfikujemy na poziomie istotności $\alpha = 0{,}05$.

Dopasowanie modelu do danych empirycznych. Współczynnik dopasowania modelu wynosi $R^2 = 0{,}393492$, a współczynnik zbieżności $\varphi^2 = 60{,}7\%$.

Wniosek. Model wyjaśnia 39,3% zmienności badanej cechy. Świadczy to o słabym dopasowaniu modelu do danych empirycznych.

Istotność układu współczynników regresji. Stawiamy hipotezę o nieistotności układu współczynników regresji (test 1) i weryfikujemy ją za pomocą statystyki o rozkładzie *F* Snedecora o 9 stopniach swobody licznika i 238 stopniach swobody mianownika.

Wartość empiryczna statystyki wynosi $F = 17{,}1567$, a odpowiadający jej krytyczny poziom istotności (istotność *F*) wynosi 9,69812E-22 i jest mniejszy od przyjętego poziomu istotności $\alpha = 0{,}05$. Odrzucamy zatem hipotezę H_0 na korzyść H_1.

Wniosek. Wyniki testu wskazują na zależność frekwencji przynajmniej od jednego z czynników objaśniających uwzględnionych w modelu[10].

[10] Współczynnik determinacji modelu nie jest jednak duży i wynosi 39,3% (wartość dopasowana jest jeszcze mniejsza i wynosi 37%). Oznacza to, że wybrany zestaw zmiennych objaśniających wyjaśnia frekwencję w trakcie wyborów prezydenckich w 2000 roku w sposób liniowy jedynie w niecałe 40% (czyli 60% to „inne" przyczyny takiej, a nie innej frekwencji). Taki wynik powoduje, że utworzony model nie nadaje się do ewentualnych zastosowań i powinniśmy albo go poprawić, albo odrzucić i przystąpić ponownie do etapu analizy problemu. Przypomnijmy, że nie uwzględniamy w naszych rozważaniach istoty wyborów prezydenckich, jaką jest sam kandydat.

Istotność poszczególnych współczynników regresji. Otrzymane wyniki umożliwiają badanie istotności (na poziomie $\alpha = 0{,}05$[11]) poszczególnych zmiennych objaśniających (test 2) na dwa sposoby, co odpowiada stawianiu hipotezy o tym, że poszczególne współczynniki regresji są równe zeru (hipoteza alternatywna: współczynniki równania regresji nie są równe zeru):

- przez porównanie wartości statystyki *t* Studenta z wartością krytyczną,
- przez obserwację przedziału ufności („dolne 95%" – „górne 95%").

Wartość krytyczna statystyki *t* Studenta wynosi ok. 1,96 (próba jest duża), a więc nie mamy podstaw do odrzucenia hipotezy zerowej H_0: $\alpha_j = 0$ dla zmiennych:

- zaludnienie,
- wiek,
- muzea.

Ponieważ wartość krytyczna 1,96 jest wartością przybliżoną, obserwacja przedziałów ufności (czy nie zawierają wartości 0) podpowiada, że nie trzeba usuwać więcej żadnych zmiennych objaśniających.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że parametry strukturalne (α_j) są istotne statystycznie (różne od zera) dla następujących czynników: dzieci, niepracujący, książki, kina, regon, budżet. Dla pozostałych czynników (zaludnienie, wiek, muzea) parametry strukturalne są nieistotne (równe zeru).

Analiza składników losowych modelu. Przedstawiono fragment obliczeń związanych z resztami modelu.

SKŁADNIKI RESZTOWE - WYJŚCIE

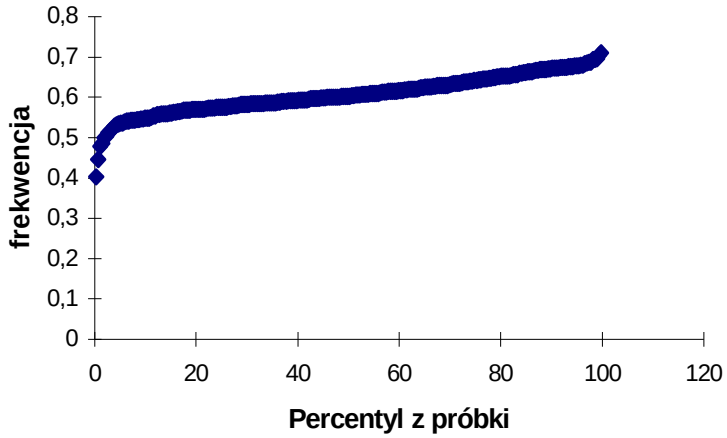
Obserwacja	*Przewidywane frekwencja*	*Składniki resztowe*	*Std. składniki resztowe*
1	0,593804436	-0,052060343	-1,402080856
2	0,609661055	-0,002589077	-0,069728599
3	0,610805229	0,050232408	1,352851202
4	0,580259514	-0,010584012	-0,285046918
5	0,580727407	0,010997091	0,296171912
6	0,586221271	0,029443163	0,792958563
7	0,621107563	0,050286583	1,354310254
8	0,608588357	-0,046770569	-1,259617501
9	0,591461514	0,003634786	0,097891468
10	0,599794263	0,015413339	0,415109602
11	0,583275462	0,001727673	0,046529426
12	0,592300425	-0,032614977	-0,878381375
13	0,605252144	-0,029941754	-0,806386543
14	0,598962066	0,004831499	0,130121155
15	0,61095294	-0,020828056	-0,560937881
16	0,580521564	-0,004258467	-0,114688346
17	0,558532156	0,011157766	0,300499167
18	0,608717725	-0,015402395	-0,414814859
19	0,586211884	-0,000554824	-0,01494244
20	0,614385241	0,057573858	1,550569966

[11] Inne poziomy α możliwe są po zmianie poziomu ufności w oknie „regresja".

NORMALNOŚĆ

Normalności składników losowych nie będziemy w tym przypadku sprawdzać za pomocą testów dokładnych – chcemy pokazać zastosowanie arkusza kalkulacyjnego, który takiego dokładnego zastosowania „wprost” nie daje. Obliczmy zatem wartości statystyk opisowych (korzystając z funkcji zawartej w bloku „narzędzia – analiza danych”):

Średnia	4,94676E-17
Błąd standardowy	0,002357806
Mediana	0,000837488
Odchylenie standardowe	0,037130771
Wariancja próbki	0,001378694
Kurtoza	1,65880605
Skośność	–0,578903645
Zakres	0,253948468
Minimum	–0,165763638
Maksimum	0,08818483
Suma	1,2268E-14
Licznik	248



Rys. 4.1. Percentyle rozkładu normalnego i percentyle z próby

Ponieważ wartości kurtozy i skośności są zawarte w przedziale [–2, 2], a próba jest duża, można więc oczekiwać, że badane wartości reszt mają rozkład normalny. Podobny wniosek otrzymujemy, gdy patrzymy na rysunek, gdzie postać krzywej jest znacząco prosta i równoległa do osi x (rys. 4.1).

Dodatkowo wartość średnia ($4{,}94 \cdot 10^{-17}$, czyli praktycznie zero) dla tak dużej próby nie powoduje odrzucenia hipotezy, gdy w populacji wartość oczekiwana **jest** równa zeru – spełnione jest zatem jedno z założeń Gaussa–Markowa dotyczące reszt.

Wniosek. Możemy zatem przyjąć, że składniki losowe modelu mają rozkład normalny o średniej 0.

Obserwacja	Przewidywane frekwencja	Składniki resztowe e_i	Std. składniki resztowe	$(e_i - e_{i+1})$	$(e_i - e_{i+1})^2$	$(e_i)^2$
1	0,593804436	–0,052060343	–1,402080856			0,002710279
2	0,609661055	–0,002589077	–0,069728599	0,049471266	0,002447406	6,70332E-06
3	0,610805229	0,050232408	1,352851202	0,052821484	0,002790109	0,002523295
4	0,580259514	–0,010584012	–0,285046918	–0,060816419	0,003698637	0,000112021
5	0,580727407	0,010997091	0,296171912	0,021581103	0,000465744	0,000120936
6	0,586221271	0,029443163	0,792958563	0,018446071	0,000340258	0,0008669
7	0,621107563	0,050286583	1,354310254	0,020843421	0,000434448	0,00252874
8	0,608588357	–0,046770569	–1,259617501	–0,097057152	0,009420091	0,002187486
9	0,591461514	0,003634786	0,097891468	0,050405354	0,0025407	1,32117E-05
10	0,599794263	0,015413339	0,415109602	0,011778554	0,000138734	0,000237571
11	0,583275462	0,001727673	0,046529426	–0,013685666	0,000187297	2,98486E-06
12	0,592300425	–0,032614977	–0,878381375	–0,034342651	0,001179418	0,001063737
13	0,605252144	–0,029941754	–0,806386543	0,002673224	7,14612E-06	0,000896509
14	0,598962066	0,004831499	0,130121155	0,034773253	0,001209179	2,33434E-05
15	0,61095294	–0,020828056	–0,560937881	–0,025659555	0,000658413	0,000433808

AUTOKORELACJA

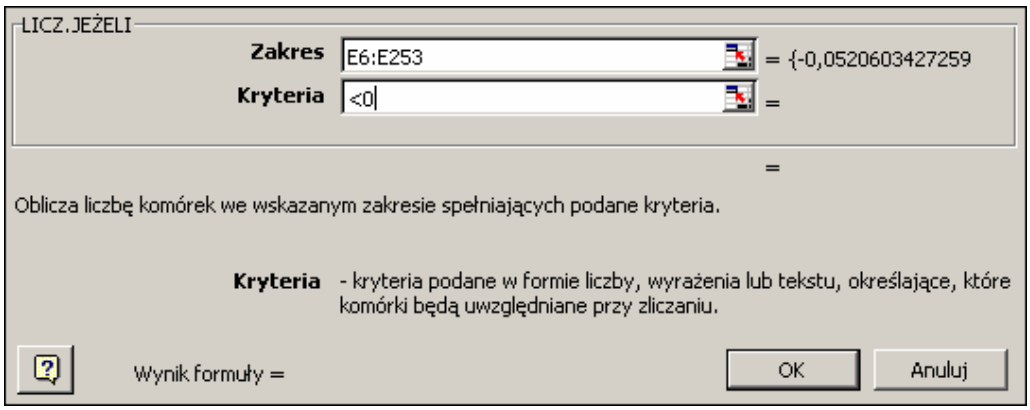
W celu sprawdzenia występowania autokorelacji obliczymy wartość statystyki Durbina–Watsona (test 7). Dane dotyczące reszt przekopiowaliśmy do arkusza o nazwie „autokorelacja”, gdzie dokonamy obliczeń wartości statystyki *d*. Fragment obliczeń przedstawiamy na wydruku, s. 121.

Wartość statystyki Durbina–Watsona wynosi 1,9675. Dla $n = 248$ obserwacji oraz $k = 6$ zmiennych objaśniających wartość krytyczna dla $\alpha = 0{,}05$ tego rozkładu wynosi w przybliżeniu $d_L = 1{,}57$, $d_U = 1{,}78$, a zatem nie ma podstaw do odrzucenia hipotezy zerowej o braku autokorelacji.

Wniosek. Nie ma podstaw do odrzucenia hipotezy o braku autokorelacji składników losowych rzędu pierwszego.

SYMETRIA

W arkuszu „autokorelacja” obliczamy też liczbę reszt, która jest mniejsza od zera. Korzystamy w tym celu z funkcji statystycznej LICZ.JEŻELI, która zwraca nam w tym przypadku wartość 123.



Ponieważ wszystkich obserwacji jest 248, stąd wartość statystyki *t* Studenta wynosi 0,1267 (test 12). Odpowiednia wartość krytyczna dla $\alpha = 0{,}05$ i $n > 30$ wynosi 1,96.

Wniosek. Nie ma podstaw do odrzucenia hipotezy, że rozkład składników losowych jest symetryczny.

LOSOWOŚĆ

Zbadanie losowości reszt przeprowadzimy na podstawie testu liczby serii (test 13). W tym celu w arkuszu „autokorelacja” najpierw dokonujemy uporządkowania od najmniejszej do największej reszt według wielkości, jaką jest przewidywana frekwencja, a następnie obliczymy liczbę serii reszt o tym samym znaku. Serie ze znakiem minus zostały zacieniowane. Przy operacji zacieniowania skorzystano z funkcji logicznej „JEŻELI”. Znajdujemy następnie liczbę wartości ujemnych i dodatnich reszt.

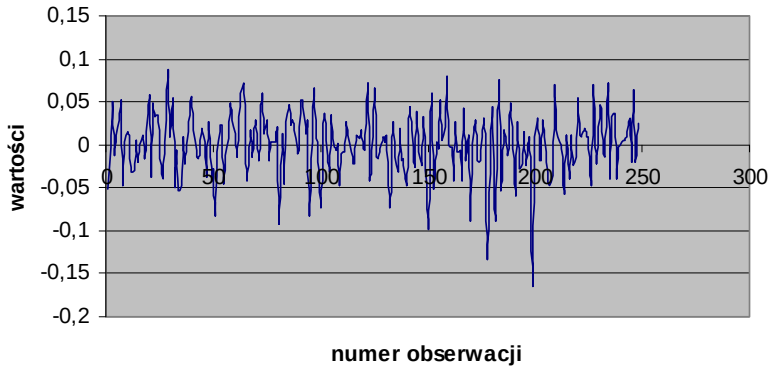
Obserwacja	Operacja porządkowania Przewidywane frekwencja	Składniki resztowe e_i	Składniki ujemne	Składniki dodatnie
25	0,546362482	-0,00455529	1	0
165	0,550895552	-0,00701929	1	0
74	0,555088106	0,0128275	0	-1
57	0,556695581	0,01684073	0	-1
...				
132	0,558155288	-0,07306934	1	0
17	0,558532156	0,01115777	0	-1
113	0,682138173	0,01198043	0	-1
118	0,696882514	0,0081574	0	-1
95	0,713335004	-0,08237419	1	0
100	0,715514406	-0,07363095	1	0
			123	-125

Tak zliczona liczba serii wynosi $K = 130$. Na poziomie $\alpha = 0{,}05$ znajdujemy, że wartość krytyczna K_α dla 123 wartości ujemnych oraz 125 wartości dodatnich wynosi około 51.

Wniosek. Nie ma podstaw do odrzucenia hipotezy o losowości reszt modelu.

HOMOSKEDASTYCZNOŚĆ

Do zbadania homoskedastyczności składników losowych (stałości wariancji) korzystamy zazwyczaj z testu Goldfelda–Quandta (test 15).



Rys. 4.2. Reszty modelu ekonometrycznego

Jeżeli dla obserwacji od 1 do 150 poszczególne wartości reszt zmieniają się dosyć równomiernie, to dla obserwacji 151–248 tak już nie jest. Podzielmy zatem próbkę na dwie części 1–150 i 151–248 i skorzystajmy z tego, że próba jest rzeczywiście bardzo duża i użyjmy wprost testu F Snedecora zamiast testu Goldfelda–Quandta:

$$F = \frac{\max\left(S_1^2, S_2^2\right)}{\min\left(S_1^2, S_2^2\right)}$$

gdzie: S_1^2, S_2^2 – wariancje wybranych części próby:

wariancja 1 (150 elementów):	0,001182714
wariancja 2 (98 elementów):	0,001681073
min wariancji	0,001182714
max wariancji	0,001681073
wartość statystyki *F*	1,421369015
wartość krytyczna $\alpha = 0{,}05$	1,364025337
wartość krytyczna $\alpha = 0{,}02$	1,474695921

Jak widać hipoteza o stałości wariancji jest możliwa do zaakceptowania jedynie na poziomie $\alpha = 0{,}02$.

Wniosek. Nie ma podstaw do odrzucenia hipotezy o równości wariancji składników losowych.

KOINCYDENCJA

Na koniec[12] zbadamy warunek koincydencji dla istotnych zmiennych objaśniających. Mamy następujące pary zmiennych:

Pary zmiennych	sign(α_i)	sign(r_{ji})	Czy zachodzi koincydencja?
Dzieci	+	–	Nie
Niepracujący	–	–	Tak
Książki	+	–	Nie
Kina	–	+	Nie
Regon	+	+	Tak
Budżet	–	+	Nie

	frekwencja	*zaludnienie*	*dzieci*	*wiek*	*niepracuj*	*ksiazki*	*kina*	*muzea*	*regon*	*budzet*
frekwencja	1									
zaludnienie	0,320604	1								
dzieci	-0,270792	-0,644055	1							
wiek	-0,211996	-0,109141	-0,13329	1						
niepracuj	-0,523161	-0,726206	0,644561	0,249563	1					
ksiazki	-0,113874	-0,445261	0,286816	0,06337	0,340703	1				
kina	0,015484	0,14989	-0,214927	-0,06098	-0,239733	0,149651	1			
muzea	-0,033424	-0,133431	0,165257	-0,08732	0,058624	0,217963	0,216138	1		
regon	0,491019	0,583707	-0,678582	-0,248286	-0,71946	-0,261834	0,254759	0,03191	1	
budzet	0,357875	0,755455	-0,611117	-0,253814	-0,81482	-0,280365	0,306717	-0,028493	0,760461	1

Wniosek. Brak koincydencji w przypadku zmiennych objaśniających: dzieci, książki, kina oraz budżet.

Krok III′. Ponowny wybór modelu

Usuwamy z modelu zmienne, które okazały się nieistotne (zaludnienie, wiek, muzea) oraz zmienne, dla których nie zachodził warunek koincydencji (dzieci, książki, kina oraz budżet), otrzymujemy następujący model:

$$\text{frekwencja} = 0{,}6597 - 0{,}1349\ \text{niepracujący} + 0{,}4832\ \text{regon}$$

[12] Należało badanie koincydencji przeprowadzić na początku weryfikacji modelu, jednak zbyt „szybko” zmniejszylibyśmy liczbę zmiennych opisujących, co nie pozwoliłoby na pokazanie wszystkich trudności, z jakimi moglibyśmy się spotkać. Tak to już bywa z rzeczywistymi przykładami.

Odpowiednie obliczenia wykonane zostały za pomocą: analiza danych – regresja. Poniżej przedstawiamy otrzymane wyniki.

PODSUMOWANIE - WYJŚCIE

Statystyki regresji	
Wielokrotn	0,548576
R kwadrat	0,300935
Dopasowa	0,295229
Błąd stand	0,040026
Obserwacj	248

ANALIZA WARIANCJI

	df	*SS*	*MS*	*F*	*Istotność F*
Regresja	2	0,168967	0,084484	52,73418	8,98E-20
Resztkowy	245	0,392506	0,001602		
Razem	247	0,561473			

	Współczynni	*d standardo*	*t Stat*	*Wartość-p*	*Dolne 95%*	*Górne 95%*
Przecięcie	0,659756	0,028763	22,93761	6,15E-63	0,603102	0,716411
niepracuj	-0,134894	0,029457	-4,579357	7,43E-06	-0,192915	-0,076873
regon	0,483155	0,156377	3,089676	0,002235	0,17514	0,79117

Model ten spełnia wymagania formalne, jednak jego przydatność jest jeszcze mniejsza, gdyż współczynnik determinacji wynosi ok. 30%.

Krok VI. Wnioskowanie na podstawie modelu

Podsumowanie: Należy znaleźć model (prawdopodobnie nieliniowy), którego współczynnik determinacji będzie większy od 50%. Model ten powinien uwzględniać podstawową zmienną objaśniającą, jaką jest sam kandydat na prezydenta, oraz kilka innych zmiennych, takich jak kultura polityczna, tradycje, działalność środków masowego przekazu i in. Nie leży to jednak w zakresie celów, jakie przyświecają tej książce, dlatego pozostaniemy przy stwierdzeniu, że model

$$\text{frekwencja} = 0{,}6597 - 0{,}1349\ \text{niepracujący} + 0{,}4832\ \text{regon}$$

nie opisuje poprawnie badanego problemu.

Literatura

[1] DITTMANN P., *Metody prognozowania sprzedaży w przedsiębiorstwie*, Wydawnictwo Akademii Ekonomicznej im. Oskara Langego, Wrocław 2000.

[2] DOMAŃSKI C., *Testy statystyczne*, PWE, Warszawa 1990.

[3] DOUGHERTY C., *Introduction to Econometrics*, Oxford University Press, London 2002.

[4] *Ekonometria*, praca zbiorowa pod red. A. Welfe, PWE, Warszawa 1998.

[5] *Ekonometria. Zbiór zadań*, praca zbiorowa pod red. A. Welfe, PWE, Warszawa 2003.

[6] GALANC T., *Metody wspomagania procesu zarządzania. Decyzyjne modele liniowe i prognozowanie ekonometryczne*, Oficyna Wydawnicza Politechniki Wrocławskiej 1998.

[7] GŁADYSZ B., KOŁWZAN W., MERCIK J., *Wielookresowy model ekonometryczny zarządzania aktywami banku*, Zastosowania Badań Operacyjnych, Łódź 1996.

[8] GŁADYSZ B., KOŁWZAN W., MERCIK J., *Eksperymenty z modelami ekonometrycznymi w prognozowaniu dla celów zarządzania aktywami i pasywami banku* [w:] Metody i zastosowania badań operacyjnych, praca zbiorowa pod redakcją T. Trzaskalika, WUAE, Katowice 1998.

[9] GOLDBERGER A. S., *Teoria ekonometrii*, PWE, Warszawa 1975.

[10] GREŃ J., *Statystyka matematyczna. Modele i zadania*. PWN, Warszawa 1982.

[11] HELLWIG Z., *Elementy rachunku prawdopodobieństwa i statystyki matematycznej*, PWN, Warszawa 1972.

[12] *Metody ekonometryczne. Przykłady i zadania*, praca zbiorowa pod red. S. Bartosiewicz, PWE, 1980.

[13] NOWAK E., *Zarys metod ekonometrii. Zbiór zadań*. PWN, Warszawa 1994.

[14] PAWŁOWSKI Z., *Ekonometria*, PWN, Warszawa 1975.

[15] STUDENMUND A. H., *Using Econometrics*, A Practical Guide, Addison Wesley Longman Inc, 2001.

[16] SZMIGIEL C., MERCIK J., *Ekonometria*, Wydawnictwo Wyższej Szkoły Zarządzania i Finansów we Wrocławiu, Wrocław 2000.

[17] *Wprowadzenie do ekonometrii w przykładach i zadaniach*, praca zbiorowa pod red. K. Kukuły, PWN, Warszawa 1999.