TESTOWANIE AUTOKORELACJI SKŁADNIKÓW LOSOWYCH W KMNRL

Łukasz Kwiatkowski

Katedra Ekonometrii i Badań Operacyjnych

Plan wykładu

1) Założenia KMNRL – powtórka
2) Własności estymatora MNK w KMRL i KMNRL – powtórka
3) Konsekwencje niespełnienia założeń o składnikach losowych
4) Testy założeń o składnikach losowych:
    a) Test Jarque-Bera (normalności rozkładu)
    b) Test Ljung-Boxa (autokorelacji)
    c) Test Dubrina-Watsona(autokorelacji)
    d) Test White'a (heteroskedastyczności)
5) Remedia

Założenia Klasycznego Modelu Normalnej Regresji Liniowej (KMNRL)

- Na KMNRL składa się sześć założeń – przypomnijmy je zarówno w zapisie skalarnym (czyli dla każdej pojedynczej obserwacji), jak i macierzowym (dla wszystkich obserwacji jednocześnie)
- Pierwsze trzy z nich dotyczą zmiennych objaśnianej (in.: zależnej) i objaśniających (in.: niezależnych, regresorów, czasami także zwanych predyktorami):

Nr	Zapis skalarny	Zapis macierzowy
1.	$\forall_{t=1,2,\dots,T}\ y_t = x_t\beta + \varepsilon_t$ gdzie $x_t = [1 \ \ x_{t1} \ \ x_{t2} \ \ \dots \ \ x_{tK}],\ k = K+1$	$\underbrace{y}_{T\times 1} = \underbrace{X}_{T\times k}\ \underbrace{\beta}_{k\times 1} + \underbrace{\varepsilon}_{T\times 1}$
2.	$\forall_{\substack{t=1,2,\dots,T \\ i=1,2,\dots,K}}\ x_{ti} -$ znane wartości nielosowe	$X -$ znana macierz nielosowa
3.	$\neg\exists_{(a_1,a_2,\dots,a_K)\in\mathbb{R}^K\setminus\{0\}} \quad \forall_{t=1,2,\dots,T} \sum_{i=1}^{K} a_i x_{ti} = 0$	$rz(X) = k$ $\rightarrow$ macierz X jest pełnego rzędu kolumnowego $\rightarrow$ równoważny warunek: $\det(X'X) \neq 0$
	Po prostu: Wartości żadnej ze zmiennych objaśniających nie dają się zapisać jako liniowa kombinacja wartości pozostałych regresorów (kombinacja o *tych samych* współczynnikach $a_1, a_2, \dots, a_K$ dla poszczególnych obserwacji o nr $t = 1, 2, \dots, T$). Tym samym kolumny macierzy X *nie* są liniowo współzależne, czyli żadna z tych kolumn nie daje się zapisać jako liniowa kombinacja pozostałych	

Założenia Klasycznego Modelu Normalnej Regresji Liniowej (KMNRL)

➢ Kolejne trzy założenia dotyczą już **struktury stochastycznej modelu**, tj. **składnika losowego**:

Nr	Zapis skalarny	Zapis macierzowy
4.	$\forall_{t=1,2,\ldots,T}\ E(\varepsilon_t) = 0$ Wartość oczekiwana każdego ε_t jest wynosi 0	$E(\varepsilon) = 0_{(T\times 1)}$ Wartość oczekiwana całego wektora ε jest równa wektorowi zerowemu
	Po prostu: Zaburzenia losowe $in\ plus$ oraz te $in\ minus$ średnio rzecz biorąc niwelują się	
5.	Dwa "pod-założenia": a) $\forall_{t=1,2,\ldots,T}\ Var(\varepsilon_t) = \sigma^2$ b) $\forall_{\substack{t,s=1,2,\ldots,T \\ t\neq s}}\ Cov(\varepsilon_t, \varepsilon_s) = 0$ $\rightarrow$ wyjaśnienia na kolejnych stronach	$\underbrace{V(\varepsilon)}_{T\times T} = \sigma^2 I_T = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$ gdzie I_T oznacza macierz jednostkową stopnia T $\rightarrow$ na przekątnej macierzy kowariancji, $V(\varepsilon)$, znajdują się wariancje, $Var(\varepsilon_t) = \sigma^2$, zaś poza przekątną kowariancje, $Cov(\varepsilon_t, \varepsilon_s), t \neq s$
6.	$\forall_{t=1,2,\ldots,T}\ \varepsilon_t \sim N$ Każdy składnik losowy ε_t ma rozkład normalny	$\varepsilon \sim N^{(T)}$ Cały (T-wymiarowy) wektor składników losowych ma T-wymiarowy rozkład normalny

Założenia Klasycznego Modelu Normalnej Regresji Liniowej (KMNRL)

➢ **Założenie 5.b**: $\forall_{\substack{t,s=1,2,\ldots,T \\ t\neq s}} Cov(\varepsilon_t, \varepsilon_s) = 0$

→ Kowariancja pomiędzy dowolnymi dwoma (różnymi) składnikami losowymi: $\varepsilon_t, \varepsilon_s$ $(t \neq s)$ wynosi 0.

→ Tym samym współczynnik korelacji także się zeruje:

$$Corr(\varepsilon_t, \varepsilon_s) = \frac{Cov(\varepsilon_t, \varepsilon_s)}{\sqrt{Var(\varepsilon_t)}\sqrt{Var(\varepsilon_s)}} = \frac{Cov(\varepsilon_t, \varepsilon_s)}{\sigma^2} = 0$$

przy czym zachodzi $Var(\varepsilon_s) = Var(\varepsilon_t) = \sigma^2$ (zgodnie z założeniem 5.a)

→ Zatem założenie 5.b oznacza, że **składniki losowe są między sobą nieskorelowane**, przez co wyklucza się ich tzw. ***auto**korelację*.

→ W praktyce, modelując dane w postaci **szeregów czasowych**, reszty z modelu regresji dość często charakteryzują się autokorelacją, co podważa założenie 5.b. Jej występowanie wynika z reguły ze zbyt uproszczonej specyfikacji części systematycznej modelu ($x_t\beta$), niedostatecznie dobrze wychwytującej dynamikę (kształtowanie się w czasie) wartości zmiennej objaśnianej.

→ Na naszych zajęciach **nie** będziemy się zajmować autokorelacją składników losowych w przypadku, gdy przedmiotem modelowania są **dane przekrojowe** (choć autokorelacja jest możliwa także w tego typu danych, np. gdy rozważamy jakieś sąsiadujące ze sobą obiekty, np. powiaty – wówczas możemy przypuszczać, że zaburzenie losowe, które "dotknęło" jeden obiekt / obserwację, wpłynęło w jakiejś mierze także na obiekty sąsiadujące). Mówimy wtedy o tzw. autokorelacji przestrzennej.

Założenia Klasycznego Modelu Normalnej Regresji Liniowej (KMNRL)

➢ **Założenie 5.b** – c.d.

→ W dalszym ciągu rozważamy zjawisko **autokorelacji** TYLKO dla **danych czasowych**, w kontekście których mówimy o tzw. **współczynnikach autokorelacji rzędu** $\boldsymbol{r} \in \{\mathbf{1}, \mathbf{2}, \dots\}$:

$$Corr(\varepsilon_t, \varepsilon_{t-r}) = \frac{Cov(\varepsilon_t, \varepsilon_{t-r})}{\sqrt{Var(\varepsilon_t)}\sqrt{Var(\varepsilon_{t-r})}} = \frac{Cov(\varepsilon_t, \varepsilon_{t-r})}{\sigma^2}$$

przy czym zachodzi $Var(\varepsilon_t) = Var(\varepsilon_{t-r}) = \sigma^2$ (zgodnie z założeniem 5.a)

→ O ε_{t-r} mówimy: "$\boldsymbol{r}$**-te opóźnienie składnika losowego**"

→ W praktyce zwykle uwagę koncentrujemy:

- w pierwszej kolejności na współczynniku autokorelacji **rzędu pierwszego**, tj. $Corr(\varepsilon_t, \varepsilon_{t-1})$ – spodziewamy się, że jeśli już autokorelacja miałaby występować, to najprawdopodobniej ta rzędu pierwszego jest najsilniejsza, oczekując, że losowe zaburzenia danego zjawiska w bieżącym okresie (o numerze t) są najsilniej powiązane (skorelowane) z tymi z okresu bezpośrednio go poprzedzającego (o numerze $t-1$)
- następnie na współczynnikach rzędu wyższego od 1, tj. $Corr(\varepsilon_t, \varepsilon_{t-r})$ dla $r = 2, 3, \dots$ – spodziewamy się, że wraz ze wzrostem opóźnienia r autokorelacja staje sie coraz słabsza, wygasa (do zera – w granicznym przypadku przy $r \to \infty$), co pozostaje w zgodzie z intuicją: im "starsze" składniki losowe (w stosunku do bieżącej chwili, t), tym słabszy ich związek/korelacja z bieżącym zakłóceniem losowym, ε_t

Założenia Klasycznego Modelu Normalnej Regresji Liniowej (KMNRL)

➢ **Założenie 5.b** – c.d.

→ Współczynnik autokorelacji jest wielkością **unormowaną** z przedziału $[-1, 1]$

→ Współczynnik autokorelacji rzędu r składników losowych mierzy **kierunek** i **siłę** korelacji bieżącego składnika losowego, ε_t, z jego przeszłą wartością sprzed r okresów, ε_{t-r}

→ **Kierunek autokorelacji**:

- $Corr(\varepsilon_t, \varepsilon_{t-r}) > 0 \Leftrightarrow$ dodatnia autokorelacja rzędu r
- $Corr(\varepsilon_t, \varepsilon_{t-r}) < 0 \Leftrightarrow$ ujemna autokorelacja rzędu r
- $Corr(\varepsilon_t, \varepsilon_{t-r}) = 0 \Leftrightarrow$ brak autokorelacji rzędu r

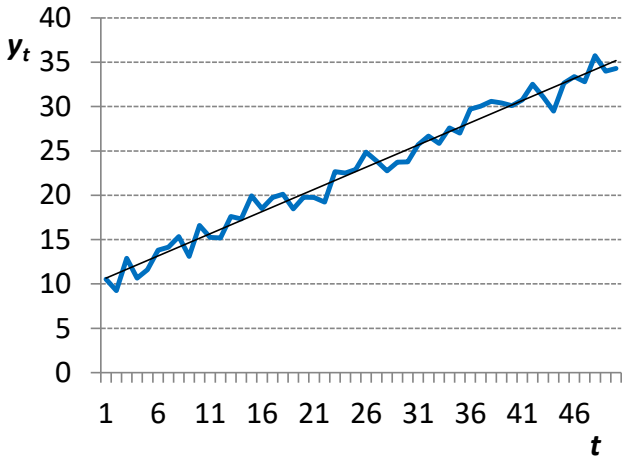
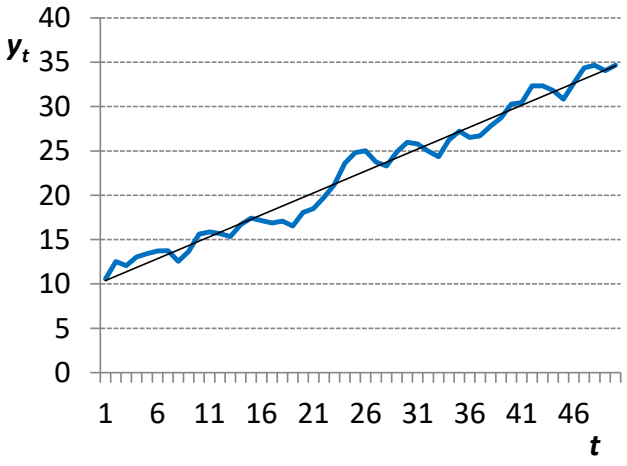
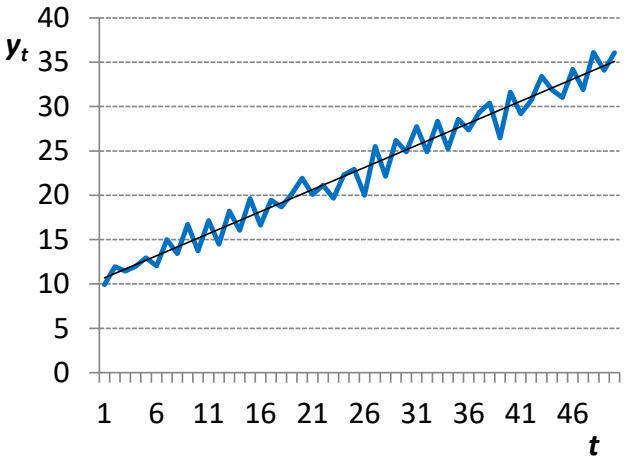
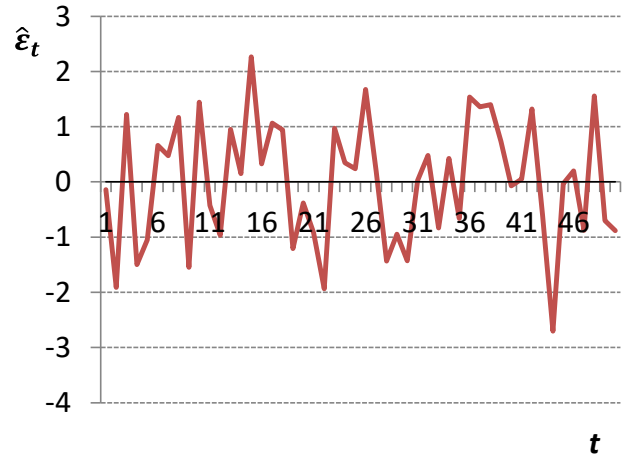
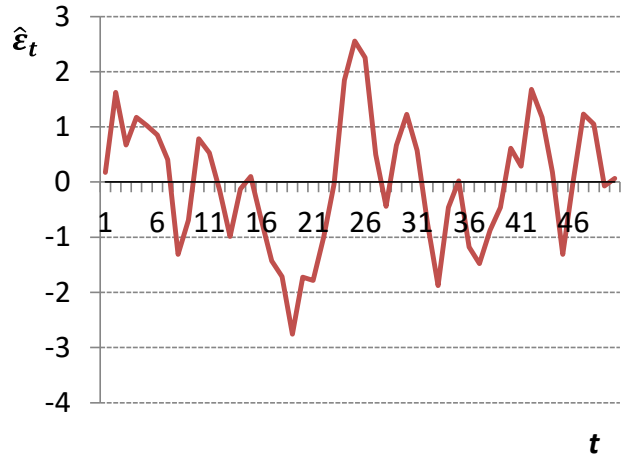
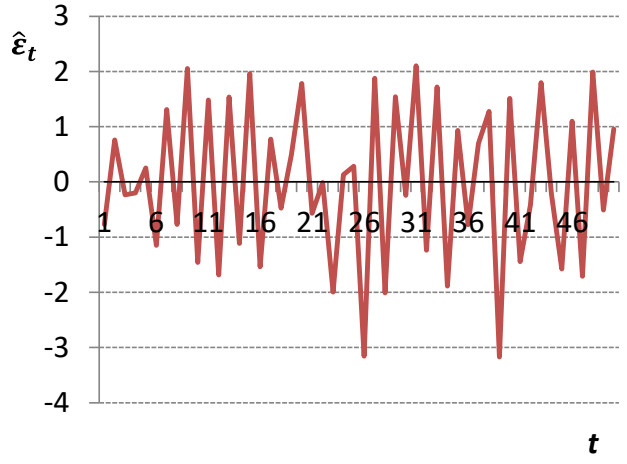
→ **Siła autokorelacji**: Im $|Corr(\varepsilon_t, \varepsilon_{t-r})|$ jest bliższa 1, tym autokorelacja (rzędu r) jest silniejsza

→ **Skrajne przypadki**:

- $Corr(\varepsilon_t, \varepsilon_{t-r}) = 1 \Leftrightarrow$ idealna autokorelacja dodatnia
- $Corr(\varepsilon_t, \varepsilon_{t-r}) = -1 \Leftrightarrow$ idealna autokorelacja ujemna

Założenia Klasycznego Modelu Normalnej Regresji Liniowej (KMNRL)

➢ **Założenie 5.b** – c.d. – przykłady – dane symulowane: $y_t = 10 + 0{,}5t + \varepsilon_t$ – Z i BEZ autokorelacji ε_t

Brak autokorelacji: $Corr(\varepsilon_t, \varepsilon_{t-1}) = 0$	Autokorelacja dodatnia: $Corr(\varepsilon_t, \varepsilon_{t-1}) = 0{,}8$	Autkorelacja ujemna: $Corr(\varepsilon_t, \varepsilon_{t-1}) = -0{,}8$
Dane z dopasowanym modelem trendu liniowego		
		
Reszty		
		

Założenia Klasycznego Modelu Normalnej Regresji Liniowej (KMNRL)

- Przyglądając się wykresom na poprzedniej stronie możemy zauważyć, że:
  1) Gdy w resztach występuje autokorelacja **dodatnia** (środkowy panel), wówczas ich trajektoria (przebieg szeregu czasowego):
     - jest gładsza (ang. *smooth*) niż w przypadku braku autokorelacji (lewy panel)
     - rzadziej przecina oś poziomą (w zerze)
     - kolejne reszty są do siebie bardziej "podobne", w związku z czym...
     - kolejne reszty relatywnie rzadko/rzadziej (w stosunku do braku autokorelacji) zmieniają znak, wobec czego dają się zauważyć dłuższe serie kolejnych reszt o tym samym znaku
  2) Gdy w resztach występuje autokorelacja **ujemna** (prawy panel), wówczas ich trajektoria (przebieg szeregu czasowego):
     - jest bardziej "szarpana" (ang. *jagged, rugged*) niż w przypadku braku autokorelacji (lewy panel)
     - częściej przecina oś poziomą (w zerze)
     - kolejne reszty są do siebie mniej "podobne"
     - kolejne reszty "nadmiernie często" zmieniają znak

- W przypadku modeli regresji zbudowanych dla danych ekonomicznych, jeśli reszty charakteryzują się autokorelacją, to zwykle jest to autokorelacja **dodatnia** (zaś rzadko kiedy, jeśli w ogóle, ujemna)

Założenia Klasycznego Modelu Normalnej Regresji Liniowej (KMNRL)

- W praktyce nie mamy żadnej innej możliwości wnioskowania o autokorelacji czy o innych własnościach / założeniach co do **składników losowych** (czyli pewnych *nieobserwowalnych* zmiennych losowych), jak tylko poprzez badanie tychże własności w odniesieniu do **reszt**. Te właśnie reszty traktujemy, choć nie do końca formalnie, jako realizacje składników losowych (*warunkowo* względem badanego zbioru danych i tym samym uzyskanych oszacowań parametrów).
- Tak więc:
  1) o *prawdziwych* współczynnikach autokorelacji **składników losowych**, $Corr(\varepsilon_t, \varepsilon_{t-r})$, tylko *zakładamy* na poziomie KMNRL, że są równe 0 (założenie 5.b)
  2) aby jednak zweryfikować, czy to założenie możemy uznać za spełnione w ramach danego oszacowanego modelu, potrzebujemy poddać analizie *oszacowania* tych współczynników, a te uzyskujemy poprzez obliczenie "zwykłych" współczynników korelacji (tj. tych poznanych jeszcze na podstawowym kursie statystyki) dla **reszt**:

$$\underbrace{Corr(\varepsilon_t, \varepsilon_{t-r})}_{\substack{Prawdziwa\ nieznana\ wartość,\\ o\ której\ tylko\ zakładamy, że\ =0}} \approx \underbrace{\widehat{Corr}(\varepsilon_t, \varepsilon_{t-r})}_{Oszacowanie} = \frac{\overbrace{\frac{1}{T-r}\sum_{t=1+r}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_{t-r}}^{\substack{Licznik=oszacowanie\ autokowariancji\\ rzędu\ r}}}{\underbrace{\frac{1}{T-k}\sum_{t=1}^{T} \hat{\varepsilon}_t^2}_{\substack{Mianownik=oszacowanie\ wariancji=\\ =wariancja\ resztowa}}}$$

- Odnośnie wcześniejszych wykresów: dane zostały tam sztucznie wygenerowane, do czego potrzebne było zadanie konkretnych *prawdziwych* wartości współczynnika autokorelacji $Corr(\varepsilon_t, \varepsilon_{t-1})$ i tym samym są tam znane: 0; 0,8; −0,8. Z łatwością możemy też obliczyć ich oszacowania na podstawie uzyskanych reszt, uzyskując wyniki, odpowiednio: 0,03; 0,70; −0,68.

Konsekwencje niespełnienia założeń KMNRL o składnikach losowych

- **Założenie 4** (zerowa wartość oczekiwana, $E(\varepsilon_t) = 0$) jest zawsze spełnione, gdy tylko w modelu uwzględniony jest **wyraz wolny**. Istotnie, można wtedy pokazać, że suma reszt MNK w modelu regresji liniowej wynosi 0, a tym samym ich średnia arytmetyczna (jako oszacowanie wartości oczekiwanej) wynosi 0. Zatem nie pozostaje tu nic do testowania. Gdyby jednak w modelu nie uwzględniono wyrazu wolnego, wówczas suma i średnia arytmetyczna reszt mogą być dowolne, co w konwekwencji – choć tego tu nie pokazujemy – prowadzi do mniejszego lub większego obciążenia estymatora MNK (czyli nie jest już prawdą, że średnio rzecz biorąc trafia w to, co trzeba, czyli β). Podsumowując: w modelu regresji liniowej zawsze powinien występować wyraz wolny (z wyjątkiem modelowania sezonowości w parametryzacji bez wyrazu wolnego – choć i wtedy też suma reszt =0).

- **Założenia 5.a i 5.b** (czyli po prostu założenie 5) – tu sprawa jest delikatna :) Co się dzieje, gdy reszty w modelu regresji liniowej charakteryzują się istotną autokorelacją lub heteroskedastycznością? Co do **własności estymatora MNK**, to:

  1) Pozostaje **liniowy** (←własność neutralna, ale najważniejsze, że →) **zgodny i nieobciążony**
  2) **Nie jest już efektywny** – można bowiem wskazać inny estymator (tzw. Uogólnionej MNK), który ma mniejszą wariancję
  3) **Błędy średnie szacunku** obliczane wg dotychczasowej formuły (tj. z macierzy $\hat{V}(\hat{\beta}) = s^2(X'X)^{-1}$) są **niedoszacowane** (czyli obciążone *in minus*, zaniżone)

Konsekwencje niespełnienia założeń KMNRL o składnikach losowych

→ **Własności 1 i 2** sprawiają, że uzyskanym za pomocą zwykłej MNK ocenom parametrów możemy nadal ufać (estymator jest nieobciążony), choć tym razem pozostaje niedosyt w związku z tym, że istnieje też inny estymator (UMNK), który nie dość, że nieobciążony, to jeszcze jest efektywny (a zatem "efektywniejszy" niż MNK). Tym estymatorem się jednak nie będziemy zajmować, ale warto wiedzieć o jego istnieniu :)

→ **Własność 3** jest poważniejsza w konsekwencjach. Okazuje się, że błędy średnie szacunku zwykłej MNK obliczane tak, jak do tego przywykliśmy (tj. z macierzy $\hat{V}(\hat{\beta}) = s^2(X'X)^{-1}$) są niedoszacowane, czyli "przekłamane" (więc nie możemy im do końca ufać) i to w dodatku w atrakcyjny sposób, ponieważ niskie błędy szacunku oznaczają wysoką precyzję wnioskowania o parametrach. Co więcej, analizując postać statystyki testowej *t*-Studenta w teście istotności: $t(\beta_i) = \frac{\hat{\beta}_i}{d(\hat{\beta}_i)}$, zauważamy, że niska wartość $d(\hat{\beta}_i)$ będzie się przekładała na wysoką (co do modułu) wartość statystyki testowej, a tym samym wyższe prawdopodobieństwo odrzucenia H_0 na rzecz H_1, czyli na rzecz wskazania, że dany parametr β_i jest statystycznie istotny... A my przecież lubimy istotność ;) Problem w tym, że nie jesteśmy tu w stanie rozsądzić, czy i na ile "faktycznie" parametr jest istotny, a na ile H_0 o jego nieistotności zostaje odrzucona z powodu zaniżonego błędu średniego szacunku... Ponieważ stastyka *t*-Studenta jest wtedy "przekłamana", **nie możemy ufać wynikom testów czy przedziałom ufności** (które też przecież na niej bazują).

→ Zatem **co można zrobić w sytuacji**, gdy zdiagnozujemy autokorelację lub heteroskedastyczność składnika losowego (czyli założenie 5.a lub 5.b okaże się niespełnione)? Remediów jest kilka – powiemy sobie o nich pod koniec tego materiału. Tymczasem zobaczmy jeszcze, co się dzieje, gdy nie jest spełnione kolejne, 6. założenie KMNRL.

Testy założeń o składnikach losowych – wykaz

- W tym miejscu, zanim przejdziemy do szczegółowego omówienia wybranych procedur testowych, wymieńmy tylko z nazwy te głównie stosowane w praktyce:

- **Testy (homo-/)heteroskedastyczności:**
  1) Goldfelda-Quandta
  2) Breuscha-Pagana
  3) White'a
- **Testy (braku) autokorelacji:**
  1) Durbina-Watsona (i jego modyfikacje)
  2) Ljung-Boxa
- **Testy normalności rozkładu:**
  1) Doornika-Hansena
  2) Kołmogorowa-Smirnowa
  3) Andersona-Darlinga
  4) Shapiro-Wilka
  5) Lilieforsa
  6) Jarque-Bera

→ Warto podkreślić, że wymienione obok: wszystkie testy normalności, a także test Ljung-Boxa (na występowanie autokorelacji) czy test White'a (na heteroskedastyczność) mogą być śmiało wykorzystywane do testowania tych własności także w odniesieniu do "zwykłych" zmiennych (a nie tylko składników losowych), jak np. zmiennej objaśnianej czy zmiennych objaśniających w modelu regresji.

W każdym z tych testów normalności rozkładu układ hipotez wygląda tak samo:

$$\underbrace{H_0: \forall_{t=1,2,\ldots,}\ \varepsilon_t \sim N}_{\substack{składniki\ losowe\ MAJĄ \\ rozkład\ normalny}} \quad \text{vs.} \quad \underbrace{H_1: \neg H_0}_{\substack{składniki\ losowe\ NIE\ mają \\ rozkładu\ normalnego}}$$

Zatem w praktyce, by móc wykorzystać dowolny z nich, nie trzeba znać szczegółów technicznych odnośnie przebiegu samej procedury – wystarczy bowiem samo *p-value* by dowiedzieć się, za którą z hipotez opowiada się dany test. Oczywiście „śpimy bezpiecznie", gdy wszystkie cztery testy jednomyślnie wskazują na tę samą hipotezę.

→ Poniżej omówimy tylko te wyróżnione na niebiesko. Zaczniemy od testów na autokorelację…

Test Durbina-Watsona

- W teście DW dopuszcza się – i testuje zarazem – występowanie **autokorelacji ε_t tylko rzędu 1**:

$$y_t = x_t\beta + \varepsilon_t \qquad (*)$$

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t \qquad (**)$$

- **Założenia** testu Durbina-Watsona:

  1) Równanie (*) musi zawierać **wyraz wolny;**

  2) Równanie (*) **nie** może zawierać opóźnionej zmiennej objaśnianej (y_{t-1}, y_{t-2}, …)

  3) u_t – **pomocniczy składnik losowy** o rozkładzie normalnym i „swojej własnej" wariancji, $u_t \sim iiN(0, \sigma_u^2)$

  4) Zakłada się, że $|\rho| < 1$ (czyli $\rho \in (-1, 1)$)

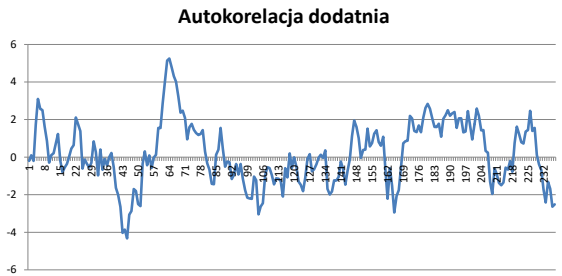
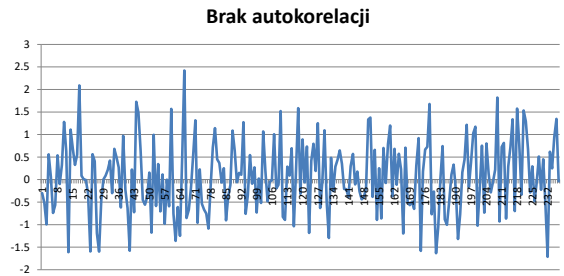
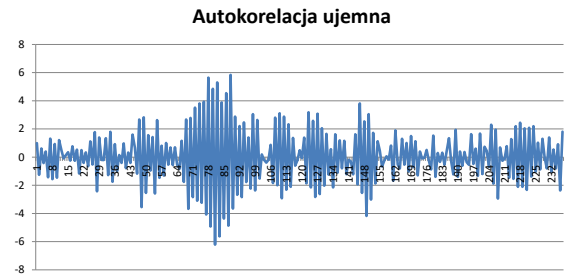
  5) $\rho \equiv \rho_1 = Corr(\varepsilon_t, \varepsilon_{t-1}) = \frac{Cov(\varepsilon_t, \varepsilon_{t-1})}{\sigma^2}$

- **Uwagi**:

  - Zasadniczo $\rho \in [-1, 1]$, ale z pewnych względów technicznych zakłada się tu, że $\rho \in (-1, 1)$
  - Równanie (**) definiuje tzw. **proces autoregresyjny rzędu 1** dla składników losowych
  - Jeżeli $\rho = 0$ (brak autokorelacji: bieżący składnik losowy NIE jest skorelowany z przeszłym), to $\varepsilon_t = u_t$, więc założenie 5.b KMNRL jest spełnione (bo u_t jest spełnia)
  - Jeżeli $\rho \neq 0$ (występuje autokorelacja – w domyśle rzędu 1. – składników losowych) i wówczas założenie 5.b KMNRL NIE jest spełnione $\Rightarrow$ Jakie są tego konsekwencje?

Test Durbina-Watsona

➢ Typy autokorelacji – w zależności od wartości parametru ρ

$\rho > 0$	$\rho = 0$	$\rho < 0$
autokorelacja dodatnia	brak autokorelacji	autokorelacja ujemna
		

→ Zanim się wykona test autokorelacji składników losowych, warto przyglądnąć się właśnie wykresowi szeregu czasowego reszt

Test Durbina-Watsona

➢ Przebieg testu:

1) Po oszacowaniu wyjściwego modelu regreji (*), obliczamy reszty, $\hat{\varepsilon}_t$

2) Obliczamy wartość $d_{emp} = \frac{\sum_{t=2}^{T}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{T}\hat{\varepsilon}_t^2} \in [0,4]$ ($\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1}$ = **pierwszy przyrost reszt**)

- $d_{emp} < 2 \Rightarrow$ podejrzewamy autokorelację dodatnią
- $d_{emp} = 2 \Rightarrow$ brak autokorelacji
- $d_{emp} > 2 \Rightarrow$ podejrzewamy autokorelację ujemną

3) Na podstawie d_{emp} stawiamy odpowiedni układ hipotez:

- $d_{emp} < 2 \Rightarrow$ podejrzewamy autokorelację dodatnią $\Rightarrow$ $\underbrace{H_0: \rho = 0}_{\substack{brak \\ autokorelacji}}$ vs. $H_1: \rho > 0$
- $d_{emp} > 2 \Rightarrow$ podejrzewamy autokorelację ujemną $\Rightarrow$ $\underbrace{H_0: \rho = 0}_{\substack{brak \\ autokorelacji}}$ vs. $H_1: \rho < 0$

4) Obliczamy wartość **statystyki testowej**:

$$DW = \begin{cases} d_{emp} \Leftrightarrow \text{testujemy } \rho > 0 \\ 4 - d_{emp} \Leftrightarrow \text{testujemy } \rho < 0 \end{cases}$$

Test Durbina-Watsona

➢ Przebieg testu – *c.d.*:

5) Z tablic testu DW odczytujemy **dwie wartości krytyczne**: *dolną* (d_L) i *górną* (d_U) – w zależności od T i $k-1$

→ Wartości krytyczne zależą od liczby obserwacji (*T*) oraz liczby regresorów bez wyrazu wolnego ($k-1$)

6) **Reguła decyzyjna**:

- Jeżeli $DW < d_L \Rightarrow$ Na danym poziomie istotności odrzucamy H_0 na rzecz H_1. Stwierdzamy zatem, że składniki losowe charakteryzują się statystycznie istotną autokorelacją dodatnią/ujemną rządu 1.
- Jeżeli $d_L \leq DW \leq d_U \Rightarrow$ Test jest niekonkluzywny (nie rozstrzyga) $\Rightarrow$ G.S. Maddala sugeruje, by „na wszelki wypadek" odrzucić H_0
- Jeżeli $DW > d_U \Rightarrow$ Na danym poziomie istotności NIE mam podstaw do odrzucenia H_0. Stwierdzamy zatem, że składniki losowe NIE charakteryzują się statystycznie istotną autokorelacją.

Test Durbina-Watsona

➢ Uwagi końcowe:

1) W teście DW nie mamy (łatwego) sposobu na wyznaczenie *p-value*

2) Na podstawie wartości d_{emp} można w przybliżeniu oszacować ρ_1: $\hat{\rho}_1 \approx 1 - \frac{d_{emp}}{2}$

3) Formalnie, odrzucenie H_0 implikuje występowanie autokorelacji rzędu 1., ale może być także symptomem występowania autokorelacji wyższego rzędu → test Ljung-Boxa (o którym już niebawem) jest pod tym względem bardziej uniwersalny

4) W sytuacji występowania autokorelacji składników losowych, możemy postąpić dwojako:

- Do zbioru zmiennych objaśniających włączyć **pierwsze opóźnienie zmiennej objaśnianej** (y_{t-1}):

$$y_t = \underbrace{\underbrace{\tilde{x}\tilde{\beta}}_{\substack{dotychczasowe \\ "x_t\beta"}} + \phi y_{t-1}}_{x_t\beta} + \varepsilon_t$$

  Taki model możemy oszacować zwykłą MNK. W tak rozszerzonym modelu składniki losowe mogą (choć nie muszą) się okazać już wolne od autokorelacji (można to *nieformalnie* sprawdzić za pomocą statystyki DW – nieformalnie, bo taki model nie spełnia założeń testu DW; patrzymy tylko czy d_{emp} jest „dostatecznie" bliskie wartości 2, powiedzmy $d_{emp} \in (1{,}85; 2{,}15)$, bez c.d. testu)

- Zastosować tzw. ***Estymowaną Uogólnioną MNK*** **(EUMNK)** – zajmiemy się nią w przyszłości. Niestety, da się ją wykorzystać tylko w przypadku autokorelacji rzędu 1, a ponadto nie prowadzi do rozbudowy modelu (o y_{t-1}), co może przyczynić się w dużym stopniu do poprawy własności predyktywnych modelu

Test Ljung-Boxa (autokorelacji składników losowych)

- Test Ljung-Boxa – ogólniejszy niż test DW (dlaczego?)
- **Nazwa testu**: test Ljung-Boxa (niepoprawnie w niektórych opracowaniach: “Ljunga-Boxa”)
- Wprowadźmy **oznaczenie ρ_l współczynnika autokorelacji rzędu $l \in \{1, 2, \dots\}$** składników losowych:

$$\rho_l = Corr(\varepsilon_t, \varepsilon_{t-l}) = \frac{Cov(\varepsilon_t, \varepsilon_{t-l})}{\sqrt{Var(\varepsilon_t)Var(\varepsilon_{t-l})}}$$

Wartości tych współczynników oczywiście nie znamy – w założniu 6 KMNRL zakładamy tylko, że $\rho_l = 0$ dla każdego $l = 1, 2, \dots$ Rozważany tu test sprawdza, czy faktycznie można uznać, że ciąg kolejnych współczynników ρ_l się zeruje (przy czym to, ile kolejnych współczynników testujemy, musimy ustalić samodzielnie). W tym celu wykorzystuje się *oszacowania* tych współczynników, $\hat{\rho}_l$, które są niczym innym, jak “zwykłymi” współczynnikami korelacji pomiędzy resztami: bieżącymi i tymi opóźnionymi o l okresów:

$$\hat{\rho}_l = \widehat{Corr}(\varepsilon_t, \varepsilon_{t-l}) = \frac{\frac{1}{T-l}\sum_{t=1+l}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_{t-l}}{\frac{1}{T-k}\sum_{t=1}^{T} \hat{\varepsilon}_t^2} \approx \frac{\frac{1}{T-l}\sum_{t=1+l}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_{t-l}}{\sqrt{\frac{1}{T-l}\sum_{t=1+l}^{T} \hat{\varepsilon}_t^2}\sqrt{\frac{1}{T-l}\sum_{t=1}^{T-l} \hat{\varepsilon}_t^2}} =$$

$\rightarrow$ W Excelu:

$$=\text{WSP.KORELACJI}(\hat{\varepsilon}_{(1+l):T}, \hat{\varepsilon}_{1:(T-l)})$$

Ale po kolei – najpierw układ hipotez (dotyczący tych prawdziwych, nieznanych, a w nadziei równych zero współczynników autokorelacji, czyli ρ_l – bez daszka)

Test Ljung-Boxa (autokorelacji składników losowych)

- **Układ hipotez**: $H_0: \rho_1 = \rho_2 = \cdots = \rho_r = 0$ vs. $H_1: \rho_1 \neq 0 \vee \rho_2 \neq 0 \vee \ldots \vee \rho_r \neq 0$

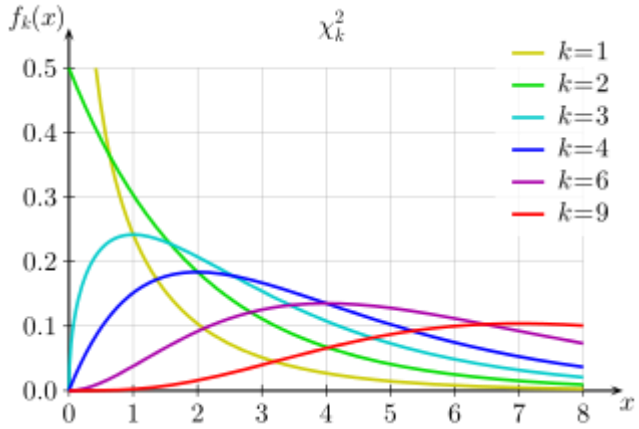
  $\rightarrow$ Zauważmy, że układ hipotez jest tu podobny do tego w teście F

  $\rightarrow$ Wartość r musimy określić samodzielnie. W literaturze można znaleźć zalecenie by przyjąć $r \approx \ln T$. Ponadto, można też przeprowadzić kilkukrotnie ten test, każdorazowo przyjmując inną wartość r, np. $r = 1, 5, 10, 15$.

- **Statystyka testowa**:

$$Q(r) = T(T+2)\sum_{l=1}^{r}\frac{\hat{\rho}_l^2}{T-l} \mid H_0 \dot{\sim} \chi^2(r)$$

$\rightarrow$ Przy prawdziwości H_0 statystyka testu LB ma (asymptotycznie) rozkład χ^2 o r stopniach swobody



Uwaga: na powyższym wykresie liczbę stopni swobody oznaczono symbolem k (zamiast r)

Źródło: https://en.wikipedia.org/wiki/Chi-square_distribution

$\rightarrow$ Wysokie wartości statystyki testowej (w prawym ogonie rozkładu) przemawiają za odrzuceniem H_0

Test Ljung-Boxa (autokorelacji składników losowych)

- **Wartość krytyczna** (dla poziomu istotności α) – oznaczamy ją symbolem χ^2_α, a obliczamy za pomocą funkcji:

  =ROZKŁAD.CHI.ODW(*prawdopodobieństwo* = α; *stopnie_swobody* = r) (stara wersja)

  =ROZKŁ.CHI.ODWR.PS(*prawdopodobieństwo* = α; *stopnie_swobody* = r) (nowa wersja)

- **Zbiór krytyczny** – zawsze prawostronny: $Z_k = (\chi^2_\alpha;\ +\infty)$

- **Konkluzja testu:**
  - Jeżeli $Q(r) \in Z_k$ (lub równoważnie: $Q(r) > \chi^2_\alpha$), wówczas odrzucamy H_0 na rzecz H_1. Stwierdzamy zatem, że składniki losowe NIE są NIESKORELOWANE, tzn. charakteryzują się statystycznie istotną autokorelacją, więc założenie 5.b KMNRL NIE jest spełnione.
  - Jeżeli $Q(r) \notin Z_k$ (lub równoważnie: $Q(r) \leq \chi^2_\alpha$), wówczas nie mamy podstaw do odrzucenia H_0. Możemy zatem uznać, że składniki losowe NIE są zautokorelowane (tzn. NIE występuje statystycznie istotna autokorelacja), więc założenie 5.b KMNRL jest spełnione.

- ***p-value*** – obliczane za pomocą funkcji:

  =ROZKŁAD.CHI(*x* = $Q(r)$; *stopnie_swobody* = r) (stara wersja)

  =ROZKŁ.CHI.PS(*x* = $Q(r)$; *stopnie_swobody* = r) (nowa wersja)

Remedia – czyli co robić, gdy zał. 5.a, 5.b czy 6 KMNRL nie jest spełnione

- W tej sekcji wskażemy najpopularniejsze sposoby poradzenia sobie w sytuacjach występowania (istotnej) autokorelacji czy też heteroskedastyczności składników losowych, jak i niespełnienia założenia o normalności ich rozkładu. Zacznijmy od…

- **Autokorelacja składników losowych** – dwa sposoby:

  I. Włączenie do modelu dodatkowych zmiennych objaśniających w postaci **opóźnionych wartości zmiennej objaśnianej** (y_{t-1}, y_{t-2}, …)

  ALTERNATYWNIE (czyli w wyjściowym modelu, bez stosowania rozwiązania I):

  II. Zastosowanie estymatora tzw. **Estymowanej Uogólnionej Metody Najmniejszych Kwadratów**, (EUMNK) – zamiast zwykłej MNK – w celu wyznaczenia nowych ocen punktowych parametrów oraz poprawnych (nieobciążonych, jak w zwykłej MNK) błędów średnich szacunku

Remedia – autokorelacja składników losowych

- **Ad I.** Pierwszym sposobem jest włączenie do modelu regresji **opóźnionych wartości zmiennej objaśnianej** (w dodatkowych charakterze zmiennych objaśniających), co możemy zapisać:

$$y_t = \underbrace{\tilde{x}_t\tilde{\beta}}_{\substack{\text{dotychczasowe}\\ \text{regresory wraz}\\ \text{z ich parametrami}}} + \underbrace{\phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p}}_{\substack{\text{dodatkowe zmienne = opóźnione wartości y-ka,}\\ \text{wraz z "nowymi" parametrami}}} + \varepsilon_t$$

gdzie $p \geq 1$ oznacza maksymalne opóźnienie uwzględnione w modelu. Całe równanie daje się zapisać w znanej nam już postaci: $y_t = x_t\beta + \varepsilon_t$, gdzie $x_t = [\tilde{x}_t \ y_{t-1} \ y_{t-2} \ldots \ y_{t-p}]$, zaś $\beta = [\tilde{\beta}' \ \phi_1 \ \phi_2 \ \ldots \ \phi_p]'$. Z tego wynika, że tak rozbudowany model jest w dalszym ciągu modelem regresji *liniowej*. Z drugiej strony, **nie** jest już *stricte* KMNRL, gdyż założenie 2 (o nielosowości wszystkich "x-sów") **nie** jest spełnione: te dodatkowe regresory, $y_{t-1}, y_{t-2}, \ldots, y_{t-p}$ – jako opóźnione wartości y_t, który jest przecież losowy (zmienną losową) – też są zmiennymi losowymi. Tym samym otrzymujemy przykład tzw. **Modelu Normalnej Regresji Liniowej z Losowymi Zmiennymi Objaśniającymi** (MoNoReLLZO ← szczyt twórczości akronimistycznej ;). Okazuje się, że taki model (przy pewnych dodatkowych warunkach technicznych) można szacować za pomocą zwykłej MNK – estymator ten jest wówczas nieobciążony (choć nie jest już liniowy, gdyż w macierzy X znajdują się teraz wartości zmiennej objaśnianej, zatem wyrażenie $\hat{\beta} = (X'X)^{-1}X'y$ NIE jest już liniową funkcją obserwacji na "y-ku"). Ponadto, jeśli tylko spełnione jest założenie, że składniki losowe w tym modelu są nieskorelowane między sobą (brak autokorelacji), wówczas możemy także obliczać (i im ufać :) błędy średnie szacunku w "standardowy" sposób, tj. z macierzy $\hat{V}(\hat{\beta}) = s^2(X'X)^{-1}$, a także testować hipotezy i budować przedziały ufności – tak, jak dotychczas.

Remedia – autokorelacja składników losowych

UWAGA: W szczególnym przypadku, model postaci

$$y_t = \cancel{\underbrace{\tilde{x}_t\tilde{\beta}}_{\substack{\text{dotychczasowe regresory wraz} \\ \text{z ich parametrami}}}} + \underbrace{\phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p}}_{\substack{\text{dodatkowe zmienne = opóźnione wartości y-ka,} \\ \text{wraz z "nowymi" parametrami}}} + \varepsilon_t$$

czyli po prostu

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

określamy mianem **modelu autoregresyjnego rzędu *p*** (ang. *autoregressive model*, *autoregression*), **AR(*p*)**.

→ Modele AR(*p*) są niezwykle ważne! → przedmiot *Analiza szeregów czasowych*

Zatem na model postaci

$$y_t = \underbrace{\tilde{x}_t\tilde{\beta}}_{\substack{\text{dotychczasowe} \\ \text{regresory wraz} \\ \text{z ich parametrami}}} + \underbrace{\phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p}}_{\substack{\text{dodatkowe zmienne = opóźnione wartości y-ka,} \\ \text{wraz z "nowymi" parametrami}}} + \varepsilon_t$$

można spojrzeć jako pewne uogólnienie modelu AR(*p*), polegające na włączeniu do zbioru zmiennych objaśniających (oprócz opóźnień y_t) pewnych dodatkowych, tym razem „zewnętrznych” regresorów, zawartych w wektorze $\tilde{x}_t$.

Remedia – autokorelacja składników losowych

$\rightarrow$ Warto tu jeszcze wyjaśnić, DLACZEGO włączenie opóźnień zmiennej objaśnianej do modelu miałoby "wyczyścić" składnik losowy z autokorelacji? Udzielimy sobie tylko intuicyjnej odpowiedzi, bez wchodzenia w formalizm. Otóż bardzo wiele zjawisk (nie tylko społeczno-gospodarczych) charakteryzuje się tym, że ich poziom w danym okresie (t) jest "jakoś podobny" do tego z wcześniejszych okresów (najpewniej $t-1$, ale też i $t-2$, $t-3$ itd.). Przykładów można by tu podać bez liku: inflacja, stopa bezrobocia, notowania na rynkach kapitałowych, stopy procentowe, zużycie energii elektrycznej itp. Oznacza to, że wiele szeregów czasowych odznacza się po prostu autokorelacją (dodatnią). Aby ująć w modelu regresji to skorelowanie danej zmiennej y_t w okresie bieżącym z jej przeszłością ($y_{t-1}, y_{t-2}, \ldots$) wprowadzamy do niego właśnie opóźnione wartości tejże zmiennej, co w konsekwencji pozwoli lepiej modelować jej dynamikę (kształtowanie się w czasie). Co się dzieje, gdy tego nie zrobimy? Można (zdecydowanie nieformalnie) powiedzieć, że wszystko to, co nie zostało ujęte w części systematycznej modelu regresji (tj. w komponencie $x_t\beta$) siłą rzeczy "ląduje" w składniku losowym (ε_t). Tak więc, jeśli autokorelacja danego zjawiska (y_t) nie została ujęta i "wymodelowana" poprzez właściwą specyfikację modelu w jego komponencie systematycznym (poprzez uwzględnienie w nim opóźnień zmiennej objaśnianej), wówczas reszty z takiego modelu będą się charakteryzować autokorelacją. Swoją drogą zaznaczmy także, że podobne "przeniesienie" będzie miało miejsce, gdybyśmy modelując jakiś szereg czasowy z ewidentną sezonowością pominęli zupełnie zmienne sezonowe w specyfikacji równania regresji – wówczas ta sezonowość będzie widoczna w trajektorii reszt.

$\rightarrow$ Z powyższych rozważań wynika, że "niekorzystne" (z punktu widzenia spełniania założeń modelu) własności reszt oszacowanego modelu mogą być po prostu symptomem błędu specyfikacji modelu, a tym samym cenną wskazówką odnośnie tego, w którym kierunku go rozbudować.

Remedia – autokorelacja składników losowych

$\rightarrow$ Wracając do kwestii "zaradzenia" problemowi autokorelacji składników losowych, powstaje pytanie o to, **ile powinno wynosić p**, czyli innymi słowy, **ile opóźnień zmiennej objaśnianej należy uwzględnić w modelu regresji**, gdy okaże się, że reszty faktycznie są zautokorelowane. Są różne podejścia – tu podamy sobie jedno z tych najczęściej wykorzystywanych w praktyce, a zarazem bardzo intuicyjne: dokładaj kolejne opóźnienia aż do momentu, gdy reszty przestaną przejawiać autokorelację. Zatem na początek szacujemy "zwykły" model (tj. bez opóźnień). Jeśli reszty okażą się zautokorelowane (w świetle wyników stosownego testu statystycznego, który zaprezentujemy później), wówczas uwzględnij w modelu tylko pierwsze opóźnienie, tj. y_{t-1}. Po oszacowaniu (zwykłą MNK) tak rozbudowanej specyfikacji, ponownie przeprowadź test autokorelacji reszt – jeśli ponownie wskaże on na występowanie statystycznie istotnej autokorelacji, wówczas uwzględnij w modelu kolejne opóźnienie, tj. y_{t-2}. Po oszacowaniu takiego równania, ponownie zbadaj autokorelację reszt itd... W praktyce częstokroć okazuje się, że uwzględnienie maksymalnie dwóch opóźnień jest już wystarczające.

Testem, którym każdorazowo w powyższym postępowaniu można wykorzystać do badania autokorelacji reszt (w modelach z kolejnymi opóźnieniami „y-ka") jest test Ljung-Boxa. Jeśli zaś chodzi o test Durbina-Watsona, to formalnie w sytuacji uwzględnienia opóźnionej zmiennej objaśnianej w modelu NIE można go stosować. Niemniej częstokroć i tak się go poniekąd wykorzystuje – obliczając wartość statystyki d_{emp}, a następnie patrząc tylko, czy d_{emp} jest „dostatecznie" bliskie wartości 2, powiedzmy $d_{emp} \in (1{,}85; 2{,}15)$. Wtedy nie przeprowadza się pełnej procedury testowej (nie wyznacza się wartości krytycznych itd.), a tylko właśnie przygląda się samej wartości d_{emp} (dlatego jest to podejście *nieformalne*).

Remedia – autokorelacja składników losowych

- **Ad II.** Drugi sposób to zastosowanie estymatora tzw. ***Estymowanej Uogólnionej MNK*** **(EUMNK)** – w szczegółach zajmiemy się nią w przyszłości, a tu uczynimy tylko kilka uwag.

  → Gdy występuje autokorelacja składników losowych, wówczas estymator zwykłej MNK, $\hat{\beta}$, jest nadal estymatorem liniowym, zgodnym i nieobciążonym (ta ostatnia własność jest dla nas kluczowa), ale traci swoją efektywność. Oznacza to, że oceny uzyskane za pomocą zwykłej MNK są w dalszym ciągu poprawne i „sensowne", ale wówczas lepszy (czyli o mniejszym rozproszeniu) może – choć nie musi – okazać się estymator EUMNK, który:
  - **nie** jest estymatorem **liniowym** (bo się okaże nieliniową funkcją obserwacji na zmiennej objaśnianej); niemniej jednak w dalszym ciągu, podobnie jak MNK, da się go łatwo zastosować
  - jest **zgodny** i **nieobciążony** (jak zwykła MNK)
  - jest ***asymptotycznie*** **efektywny**, czyli jest efektywny (o najmniejszym rozproszeniu spośród wszystkich estymatorów nieobciążonych, w tym MNK), ale tylko *asymptotycznie*, tj. przy założeniu, że $T \to \infty$. Zatem w skończonej próbie (z którą, *de facto*, mamy do czynienia zawsze :) tak naprawdę nie wiemy, czy zastosowanie EUMNK da nam precyzyjniejsze oceny parametrów niż te ze zwykłej MNK. Możemy mieć jednak nadzieję, że im bardziej liczebną próbą dysponujemy, tym w istocie tak właśnie jest.
  - Błędy średnie szacunku estymatora EUMNK nie są już niedoszacowane (tak jak te w zwykłej MNK), ale z drugiej strony mają charakter tylko (znowu) *asymptotyczny*, czyli są prawidłowe, o ile $T \to \infty$. W praktyce zatem (gdzie zawsze przecież $T < \infty$) błędy średnie szacunku z EUMNK są tylko przybliżone, choć słusznie jest oczekiwać, że to przybliżenie jest tym dokładniejsze, im wyższą liczbą obserwacji dysponujemy.

Remedia – autokorelacja składników losowych

→ Niestety, zastosowanie EUMNK ma tu pewne ograniczenia:

1. Po pierwsze, metodę tę możemy wykorzystać tylko w sytuacji autokorelacji rzędu 1, a nie wyższego.

2. Po drugie, choć metoda ta da nam „poprawione” oceny parametrów i poprawne (w odróżnieniu od tych ze zwykłej MNK, przynajmniej w przybliżeniu) błędy średnie szacunku parametrów, tak jednak nie ubogaci struktury modelu – poprzez włączenie dodatkowych zmiennych objaśniających w postaci opóźnień zmiennej objaśnianej, co ma miejsce w przypadku rozwiązania I. Jest to akurat o tyle ważne, że włączenie tego typu dodatkowych regresorów do modelu zwykle w dużym stopniu przyczynia się do poprawny nie tylko jego dopasowania „w próbie”, ale – co nawet ważniejsze – własności predyktywnych. Szczegółami zajmiemy się w przyszłości.