TESTOWANIE ZAŁOŻENIA O NORMALNOŚCI ROZKŁADU 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 4.** raczej nie wymaga dodatkowych wyjaśnień – brzmi to całkiem logicznie, że dodatnie i ujemne zaburzenia losowe średnio rzecz biorąc znoszą / niwelują się :)

- **Założenie 5.** obejmuje w istocie, dwa "pod-założenia" – uwydatnione w zapisie skalarnym, natomiast ujęte jako jedno w zapisie macierzowym:

  1) **Założenie 5.a**: $\forall_{t=1,2,\dots,T}\ Var(\varepsilon_t) = \sigma^2$

     $\rightarrow$ Każdy składnik losowy posiada dokładnie *tę samą* (choć nieznaną) wariancję (czyli rozproszenie, rozrzut) – zauważmy, że przy σ^2 nie stoi indeks t.

     $\rightarrow$ Należy podkreślić, że *implicite* zakłada się, że ta wariancja w ogóle *istnieje* i jest *skończona* ($\sigma^2 < \infty$)

     $\rightarrow$ Tę własność jednorodności ("jednakowości") wariancji wszystkich składników losowych nazywamy **homoskedastycznością** (ang. *homoscedasticity, homoskedasticity*).

     $\rightarrow$ Jej zaprzecznie natomiast (czyli niejednorodność, "niejednakowość" wariancji składników losowych) – **heteroskedastycznością** (ang. *heteroscedasticity, heteroskedasticity*). Możemy wtedy zapisać ogólnie: $Var(\varepsilon_t) = \sigma_t^2,\ t = 1, 2, \dots, T$

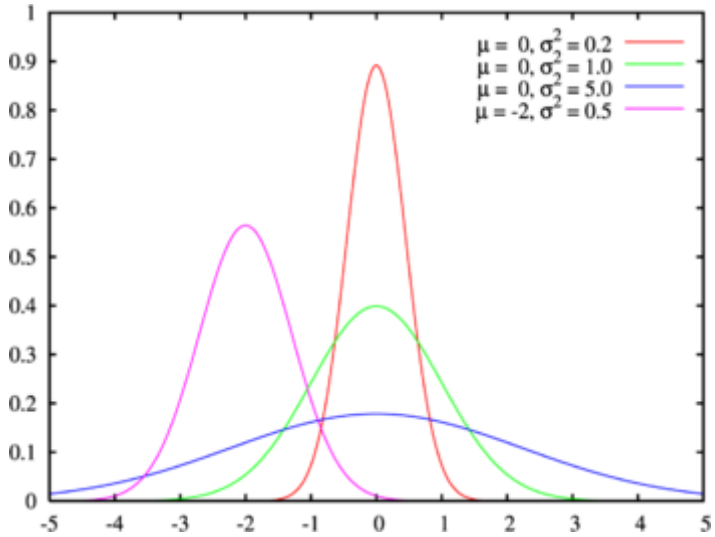
     $\rightarrow$ Warto zauważyć, że z założeń KMNRL wynika, że σ^2 jest wariancją także samej zmiennej objaśnianej: $Var(y_t) = Var(x_t\beta + \varepsilon_t) = Var(\varepsilon_t) = \sigma^2$, gdyż cały komponent $x_t\beta$ jest wielkością stałą, nielosową, co wynika z założenia 2, a przesunięcie zmiennej losowej (tu: ε_t) o pewną wielkość stałą (tu: $x_t\beta$) *nie* zmienia rozproszenia (wariancji) tak powstałej zmiennej, czyli y_t.

Założenia Klasycznego Modelu Normalnej Regresji Liniowej (KMNRL)

- **Założenie 6.**: $\forall_{t=1,2,\ldots,T}\ \varepsilon_t \sim N$ $\rightarrow$ tzn. każdy składnik losowy ma rozkład normalny

  $\rightarrow$ Przy okazji przypomnijmy, że każdy rozkład normalny posiada dwa parametry:

  - wartość oczekiwaną (zwykle oznaczaną symbolem μ; czyt. “mi”)
  - wariancję (zwykle oznaczaną symbolem σ^2)



Źródło: https://pl.wikibooks.org/

$\rightarrow$ Rozkład normalny o parametrach μ i σ^2 oznaczamy w skrócie: $N(\mu, \sigma^2)$

Założenia Klasycznego Modelu Normalnej Regresji Liniowej (KMNRL)

- **Założenia 4-6.** możemy w skrócie zapisać w następujący sposób:

$$\forall_{t=1,2,\ldots,T}\ \varepsilon_t \sim iiN(0,\sigma^2)$$

co oznacza, że składniki losowe są między sobą nieskorelowane (a dokładniej: niezależne; ang. ***i**ndependent*) i każdy z nich podlega temu samemu (ang. ***i**dentical*) rozkładowi normalnemu (ang. ***N**ormal distribution*) o zerowej wartości oczekiwanej i (pewnej stałej, nieznanej) wariancji σ^2.

- Przypomnijmy tu jeszcze, że:
  1) **Założenia 1-5.** (bez 6.) definiują Klasyczny Model Regresji Liniowej (**KMRL**) – składniki losowe mogą mieć tu dowolny rozkład prawdopodobieństwa (byle o wartości oczekiwanej równej 0 i stałej, skończonej wariancji, σ^2)
  2) **Założenia 1-6.** definiują Klasyczny Model **Normalnej** Regresji Liniowej (**KMNRL**)

- Powstaje zasadnicze pytanie: Co zyskujemy dzięki wprowadzeniu założenia 6? Odpowiedzi udzielimy za chwilę, gdy tylko przypomnimy sobie dwa twierdzenia o własnościach estymatora MNK – osobno w KMRL i KMNRL

Własności estymatora MNK w KMRL i KMNRL

➢ **Twierdzenie Gaussa-Markowa – o własnościach estymatora MNK w KMRL (czyli bez założenia 6.)**

Jeżeli spełnione są założenia KMRL, wówczas estymator MNK wektora parametrów strukturalnych jest najlepszy (inaczej: efektywny, najefektywniejszy) w klasie wszystkich liniowych estymatorów nieobciążonych.

→ Przypomijmy, że **estymator najlepszy** (inaczej: efektywny, najefektywniejszy) w danej klasie (grupie) estymatorów to taki, który spośród wszystkich estymatorów w tej klasie ma najmniejszą wariancję (czyli rozproszenie; a tym samym – najwyższą precyzję). Z reguły interesuje nas znalezienie estymtora najefektywniejszego w klasie **estymatorów nieobciążonych**, czyli takich, których wartość oczekiwana jest równa dokładnie temu parametrowi, który chcemy oszacować (czyli "średnio rzecz biorąc dobrze trafia"). Istotnie, co nam po takim estymatorze, którzy ma najmniejszą wariancję, ale średnio rzecz biorąc nie trafia w to, co trzeba (tylko "znosi" go na lewo lub na prawo od wartości parametru)? Zgodnie z powyższym twierdzeniem, estymator MNK – który zarówno w KMRL, jak i KMNRL jest liniową funkcją obserwacji (we wzorze na $\hat{\beta} = (X'X)^{-1}X'y$ wektor obserwacji y "wchodzi" liniowo) – jest najlepszy, lecz "tylko" spośród estymatorów liniowych. "Prawdopodobnie" istnieje estymator, będący jakąś nieliniową funkcją y, który ma jeszcze mniejszą wariancję (ale nie jest on tak łatwy i bezpośredni w zastosowaniu jak MNK).

→ O estymatorze MNK w KMRL mówimy też, że jest BLUE (ang. *Best Linear Unbiased Estimator*)

Własności estymatora MNK w KMRL i KMNRL

- **Twierdzenie (“bezimienne”) o własnościach estymatora MNK w KMNRL**

> Jeżeli spełnione są założenia KMNRL, wówczas estymator MNK wektora parametrów strukturalnych jest najlepszy (inaczej: efektywny, najefektywniejszy) w klasie wszystkich (tj. zarówno liniowych, jak i nieliniowych) estymatorów nieobciążonych.

$\rightarrow$ Zestawiając ze sobą te dwa twierdzenia zauważamy, że przy założeniu normalności rozkładu składników losowych, estymator MNK okazuje się najlepszy już nie tylko wśród estymatorów liniowych (choć w dalszym ciągu jest liniowy, bo jego formuła się przecież nie zmienia), ale także wśród estymatorów nieliniowych.

$\rightarrow$ Możemy teraz przejść do:

- **Konsekwencje (a raczej korzyści) wynikające z założenia 6 KMNRL:**

1) Efektywność estymatora MNK rozszerza się także na klasę estymatorów nieliniowych (choć sam w dalszym ciągu pozostaje liniowy, będąc liniową funkcją obserwacji)
2) Możemy testować hipotezy statystyczne (niewykonalne w KMRL)
3) Możemy budować przedziały ufności (j.w.)

$\rightarrow$ Widzimy zatem, co możemy utracić, gdy **założenie 6** KMNRL (normalność rozkładu składników losowych) nie byłoby spełnione. Przejdźmy tym samym to nieco dokładniejszego omówienia **konsekwencji niespełnienia** założeń KMNRL dotyczących składników losowych.

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) – innym razem

Konsekwencje niespełnienia założeń KMNRL o składnikach losowych

➢ **Założenie 6** (normalność rozkładu składników losowych)

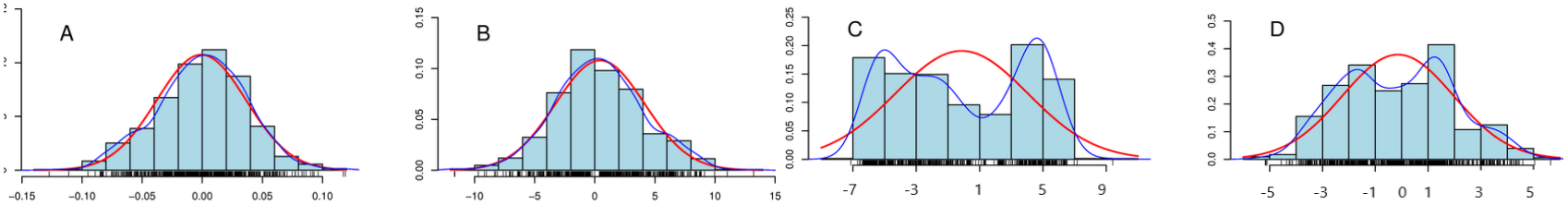
$\rightarrow$ Przypomnijmy, że dwiema z trzech korzyści wynikających z założenia 6 KMNRL jest możliwość przeprowadzania testów statystycznych (*t*, *F*) oraz budowa przedziałów ufności. "Wykonalność" tych technik wnioskowania statystycznego w KMNRL bierze się z tego, że to właśnie przy założeniu, że składniki losowe mają rozkład normalny (o zerowej wartości oczekiwanej i pewnej stałej, nieznanej wariancji) – a nie jakiś inny (choćby też miał zerową wartość oczekiwaną i pewną stałą wariancję) – można pokazać, że wyrażenie (statystyka) postaci $\frac{\widehat{\beta}_i - \beta_i}{d(\widehat{\beta}_i)}$ ma pewien konkretny, dobrze znany (i o dobrze poznanych własnościach) rozkład prawdopodobieństwa, którym jest tu rozkład *t*-Studenta (o $T - k$ stopniach swobody). Jeśli składniki losowe nie miałyby rozkładu normalnego, wówczas owszem, od strony czysto technicznej/obliczeniowej możemy obliczyć sobie wartość tej statystyki w teście (w dalszym ciągu nazywaną tak samo, tj. statystyką *t*-Studenta), ale nie wiemy już, jaki jest jej rozkład, a zatem nie wiemy skąd wziąć wartości krytyczne i jak (tj. w jakim rozkładzie – bo go nie znamy) obliczyć *p-value*. Wobec tego nie możemy sformułować konkluzji testu. Podobnie ma się sprawa z testem *F* – wartość statystyki możemy sobie obliczyć, problem jednak w tym, że nie można już dowieść, że ma ona rozkład *F*, ani nie potrafimy wyprowadzić dokładnej postaci tego prawdziwego rozkładu.

Konsekwencje niespełnienia założeń KMNRL o składnikach losowych

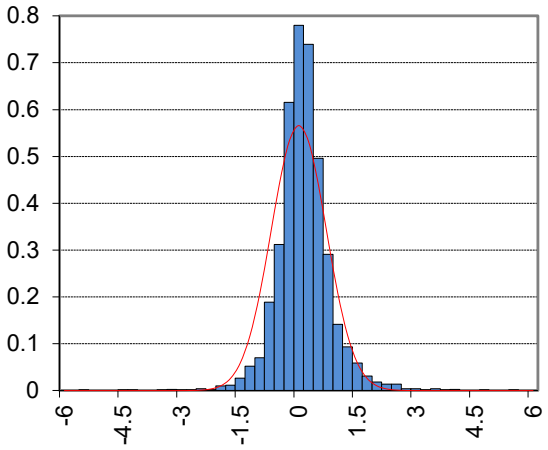
$\rightarrow$ Jakie to ma znaczenie praktyczne? W praktyce zwykle siłą rzeczy i tak otrzymujemy "gdzieś po drodze" wyniki testów istotności (t, F), czy estymacji przedziałowej (ot, choćby prezentuje je zastosowany pakiet statystyczny jako element zestawu wyników domyślnie prezentowanych użytkownikowi po oszacowaniu danego modelu regresji). Gdy przeprowadzimy dodatkowo test normalności reszt i okaże się, że założenie 6 nie daje się obronić, wówczas (patrz kolejne strony):

Konsekwencje niespełnienia założeń KMNRL o składnikach losowych

1) **Po pierwsze**, nie panikuj :) Zbuduj histogram reszt wraz z dopasowaniem rozkładu normalnego i sam oceń "na oko", na ile poważne jest odstępstwo tego pierwszego od krzywej Gaussa – przykłady (niebieska linia to tylko "uciąglony" histogram; czerwona – gęstość scentrowanego w zerze rozkładu normalnego o wariancji równej wariancji resztowej)



→ Na panelach A i B odstępstwo empirycznego rozkładu reszt (czyli ich histogramu) od zakładanego dla składników losowych rozkładu normalnego jest niewielkie (lekka skośność), więc wyniki testów czy przedziałów ufności w modelu regresji z takim rozkładem reszt zapewne nie odbiegają zbytnio od prawidłowych; w przypadkach C i D jest już nieco gorzej (choćby z uwagi na wielomodalność). Ponadto, zwłaszcza gdy przedmiotem modelowania są stopy zwrotów z różnej maści instrumentów finansowych, wówczas rozkład reszt z modelu regresji może przybierać taką oto postać:



← Taki rozkład, który charakteryzuje się **spiczastością** (tj. wyższą niż w rozkładzie normalnym koncentracją wartości w części centralnej) określamy jako **leptokurtyczny** (w przeciwieństwie do rozkładów platykurtycznych, cechujących się „wypłaszczeniem"), a samą własność spiczastości – **leptokurtozą**. Przypomnijmy, że miarą spiczastości rozkładu jest **współczynnik kurtozy** (ozn. Kr; zawsze dodatni). Można pokazać, że:

- dla rozkładu normalnego zawsze $Kr = 3$
- dla rozkładów leptokurtycznych $Kr > 3$
- dla rozkładów platykurtycznych: $0 < Kr < 3$

Bardzo często – zwłaszcza w przypadku danych z rynków finansowych – leptokurtozie towarzyszą tzw. „**grube ogony**" (ang. *fat tails*), które wynikają z wyższej niż ta przewidywana przez rozkład normalny częstości występowania obserwacji „odstających", tj. dalekich od części centralnej rozkładu. Jeśli się dobrze przyjrzeć wykresowi po lewej, to zarówno w prawym, jak i lewym ogonie dotrzeżemy niezerowe belki histogramu, oznaczające, że niektóre reszty osiągnęły aż tak wysokie/niskie wartości.

Konsekwencje niespełnienia założeń KMNRL o składnikach losowych

1) **(c.d.)** Zauważmy, że o sile odstępstwa rozkładu reszt od rozkładu normalnego może nas też informować *p-value* otrzymywane w testach normalności (jak wskażemy to jeszcze poniżej, w każdym takim teście H_0 stanowi, że składniki losowe mają rozkład normalny, zaś H_1 jest jej zaprzeczeniem) – im niższe *p-value*, tym H_0 jest silniej odrzucana.
2) **Po drugie**, otrzymane wcześniej rezultaty testów statystycznych i estymacji przedziałowej musimy teraz traktować z pewną dozą ostrożności i tylko jako (w najlepszym razie) przybliżone (bo zostały uzyskane z wykorzystaniem technik, które mają swoje uzasadnienie/umocowanie w założeniu, które jednak nie jest spełnione). Intuicyjnie, to przybliżenie jest tym lepsze, im mniejsze odstępstwo rozkładu (histogramu) reszt od rozkładu normalnego. Znowu, odstępstwo to możemy zmierzyć za pomocą *p-value*.
3) **Po trzecie** i w końcu, należy się zastanowić nad możliwymi przyczynami tego, że reszty okazują się nie posiadać rozkładu normalnego, co pozwoliłoby także poszukać odpowiedniego remedium. Możliwych tropów jest co najmniej kilka – omówimy je w końcowej części tego opracowania, tej poświęconej właśnie sposobom radzenia sobie w sytuacjach niespełnienia założeń składników losowych w 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 Jarque-Bera (normalności rozkładu)

- **Test normalności** rozkładu składników losowych w danym, oszacowanym modelu regresji także oparty jest na **resztach MNK**, $\hat{\varepsilon}_t = y_t - \hat{y}_t$, zestawionych w wektorze $\hat{\varepsilon} = y - \hat{y}$. To na ich podstawie wnioskujemy o (nie)spełnieniu założeń poczynionych w KMNRL w odniesieniu do składnika losowego.
- **Nazwa testu**: test Jarque-Bera (czyt. "Harke-Bera", pierwszy autor jest hiszpańskiej narodowości)
- **Układ hipotez**: $\underbrace{H_0: \forall_{t=1,2,\ldots,T}\ \varepsilon_t \sim N}_{\text{składniki losowe MAJĄ rozkład normalny}}$ vs. $\underbrace{H_1: \neg H_0}_{\text{nieprawda, że } H_0\text{, zatem składniki losowe NIE mają rozkładu normalnego}}$
- **Statystyka testowa**: $JB_{emp} = \frac{T}{6}\left(Sk(\hat{\varepsilon})\right)^2 + \frac{T}{24}(Kr(\hat{\varepsilon}) - 3)^2 \mid H_0 \dot{\sim} \chi^2(2)$

(Komentarz pod nosem co poniektórych: "Łoo matko..!!" ;)

$\rightarrow Sk(\hat{\varepsilon})$ oznacza "zwykły" (znany jeszcze ze statystyki) **współczynnik skośności** dla reszt, który w arkuszu można obliczyć za pomocą funkcji **=SKOŚNOŚĆ(*wektor_reszt*)**

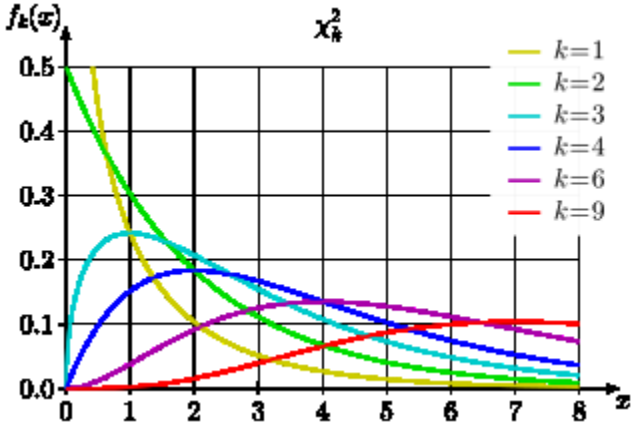
$\rightarrow Kr(\hat{\varepsilon})$ oznacza "zwykły" (znany jeszcze ze statystyki) **współczynnik kurtozy** dla reszt, który w arkuszu można obliczyć za pomocą funkcji **=KURTOZA(*wektor_reszt*)**, ALE – UWAGA!!! – do wyniku tej funkcji należy jeszcze **dodać wartość "3"** (pod funkcją =KURTOZA() Excel rozumie tzw. eksces, który definiuje się jako współczynnik kurtozy pomniejszony o 3)

$\rightarrow$ Symbol $\chi^2(2)$ oznacza tzw. **rozkład χ^2** (mała grecka litera "chi"; czyt. "**chi-kwadrat**", ale nie "czi-kwadrat" ;) o dwóch stopniach swobody (to już kolejny rozkład, którego parametr określamy tym mianem). Rozkład ten jest podobny do rozkładu F – także określony tylko na półosi dodatniej.

$\rightarrow$ Symbol $\dot{\sim}$ oznacza, że statystyka testowa ma dany rozkład tylko *asymptotycznie*, tj. przy $T \rightarrow \infty$

Test Jarque-Bera (normalności rozkładu)

- Funkcja gęstości rozkładu $\chi^2(2)$:



Uwaga: na powyższym wykresie symbol k oznacza liczbę stopni swobody

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

→ Wysokie wartości statystyki testowej (w prawym ogonie rozkładu) przemawiają za odrzuceniem H_0

***Pytanie**: Niech χ^2_α oznacza stosowną wartość krytyczną. Jakiej zatem postaci będzie zbiór krytyczny?*

Test Jarque-Bera (normalności rozkładu)

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

  =ROZKŁAD.CHI.ODW(*prawdopodobieństwo* = α; *stopnie_swobody* = 2) (stara wersja)

  =ROZKŁ.CHI.ODWR.PS(*prawdopodobieństwo* = α; *stopnie_swobody* = 2) (nowa wersja)

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

- **Konkluzja testu:**
  - Jeżeli $JB_{emp} \in Z_k$ (lub równoważnie: $JB_{emp} > \chi^2_\alpha$), wówczas odrzucamy H_0 na rzecz H_1. Stwierdzamy zatem, że składniki losowe w oszacowanym modelu NIE podlegają rozkładowi normalnemu, więc założenie 6 KMNRL NIE jest spełnione. Tym samym do uzyskanych przy tym założeniu wyników testów istotności czy estymacji przedziałowej należy podchodzić z ostrożnością, traktując je tylko jako przybliżone.
  - Jeżeli $JB_{emp} \notin Z_k$ (lub równoważnie: $JB_{emp} \leq \chi^2_\alpha$), wówczas nie mamy podstaw do odrzucenia H_0. Możemy zatem uznać, że składniki losowe w oszacowanym modelu podlegają rozkładowi normalnemu, więc założenie 6 KMNRL jest spełnione i wyniki testów istotności czy estymacji przedziałowej są wiarygodne (UWAGA: o ile NIE występuje autokorelacja czy heteroskedastyczność reszt).

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

  =ROZKŁAD.CHI(*x* = JB_{emp}; *stopnie_swobody* = 2) (stara wersja)

  =ROZKŁ.CHI.PS(*x* = JB_{emp}; *stopnie_swobody* = 2) (nowa wersja)

Remedia – odstępstwo rozkładu składników losowych od rozkładu normalnego

- Aby poszukać rozwiązania, należy się wpierw zastanowić, z czego może wynikać odstępstwo rozkładu reszt od rozkładu normalnego. Możliwości jest kilka:

I. Występująca heteroskedastyczność reszt

→ Patrz remedia dotyczące heteroskedastyczności

II. Na etapie specyfikacji równania pominięto jakieś kluczowe zmienne objaśniające

→ Także wtedy, gdy skutkiem tego nie jest heteroskedastyczność

III. Niezlogarytmowanie zmiennej objaśnianej (i ewentualnie także zmiennych objaśniających)

→ Także wtedy, gdy skutkiem tego nie jest heteroskedastyczność

IV. W modelowanym zbiorze danych występują tzw. obserwacje nietypowe/odstające (ang. *outliers*)

→ W przypadku danych przekrojowych można spróbować usunąć takie obserwacje, ale…

→ W przypadku szeregów czasowych już nie; są jednak inne sposoby – jeden z nich omówiony został na kolejnym slajdzie

→ W praktyce zdarza się czasami, że podejmowane według powyższych wytycznych odpowiednie próby „uratowania" założenia o normalności składników losowych spełzają jednak na niczym… Wówczas pozostaje ograniczyć się tylko do oceny "siły" odstępstwa histogramu reszt od rozkładu normalnego (wykres + *p-value* z testów normalności), oraz do jawnego sformułowania tej klauzuli o "konieczności zachowania pewnej dozy ostrożności" przy interpretacji wyników estymacji przedziałowej i testów statystycznych, wynikającej z wówczas tylko przybliżonego ich charakteru.

V. Zateż też… może po prostu rozkład normalny nie jest właściwy w modelowaniu danego zjawiska

Remedia – odstępstwo rozkładu składników losowych od rozkładu normalnego

Ad IV. Metoda radzenia sobie z obserwacją nietypową w modelowanym szeregu czasowym

Wyobraźmy sobie, że modelujemy szereg czasowy $\{y_t; t = 1, 2, \dots, T\}$, z wartościami zmiennej objaśnianej ułożonymi oczywiście w wektorze y, wykorzystując model regresji:

$$y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_K x_{tK} + \varepsilon_t$$

Dalej, przyjmijmy, że któregoś konkretnego numeru $t^* \in \{1, 2, \dots, T\}$ mamy do czynienia z wartością odstającą, y_{t^*}. Jeżeli model w swojej wyjściowej postaci nie „poradzi" sobie z wyjaśnieniem/modelowaniem tej obserwacji, wówczas odpowiadająca jej reszta MNK, $\hat{\varepsilon}_{t^*}$, też będzie wartością odstającą (na tle pozostałych reszt). W celu „wychwycenia"/„zamodelowania" tej wartości y_{t^*} należy wprowadzić do modelu dodatkową zmienną objaśniającą w postaci tzw. **zmiennej zero-jedynkowej** (inaczej: *typu 0-1, binarnej, dychotomicznej*), zwanej także **zmienną sztuczną** (ang. *dummy variable*):

$$D_t = \begin{cases} 1 \Leftrightarrow t = t^* \\ 0 \Leftrightarrow t \neq t^* \end{cases}$$

Ilustracja: Niech

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 D_t + \varepsilon_t, \qquad t = 1, 2, \dots, T$$

Przyjmijmy, że y_2 stanowi wartość ostającą, tj. $t^* = 2$. Wówczas macierz X wyglądałaby tak:

	X		
t	"1"	x_{t1}	$x_{t2} = D_t$
1	1	…	0
2 = t*	1	…	**1**
3	1	…	0
4	1	…	0
5	1	…	0
6	1	…	0
…	…	…	…

UWAGA: Dobrze mieć jakieś wyjaśnienie faktograficzne lub przynajmniej hipotetyczne („story") danej obserwacji odstającej. Wtedy wystąpienie tej przyczyny utożsamiamy właśnie z przyjęciem przez zmienną D_t wartości równej 1.

Remedia – odstępstwo rozkładu składników losowych od rozkładu normalnego

→ O czym informuje nas ocena parametru przy zmiennej D_t? Kontynuując powyższą ilustrację, zauważmy:

- Dla $t = t^*$ zachodzi $D_t = D_{t^*} = 1$, zatem: $y_{t^*} = \beta_0 + \beta_1 x_{t^*1} + \beta_2 + \varepsilon_{t^*}$
- Gdyby nie była wystąpiła przyczyna, dla której y_{t^*} stała się wartością odstającą, tj. w sytuacji, gdyby $D_{t^*} = 0$, wówczas otrzymalibyśmy po prostu: $y_{t^*} = \beta_0 + \beta_1 x_{t^*1} + \varepsilon_{t^*}$

→ $\hat{\beta}_2$ informuje nas o…?