TESTOWANIE ZAŁOŻENIA HOMOSKEDASTYCZNOŚCI 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)

- **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* ($\sigma^2 < \infty$)

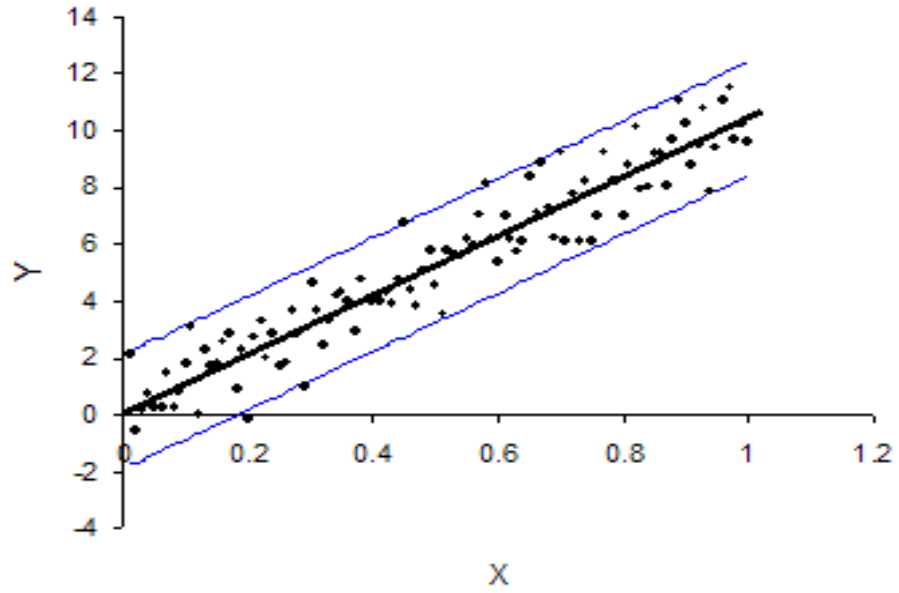
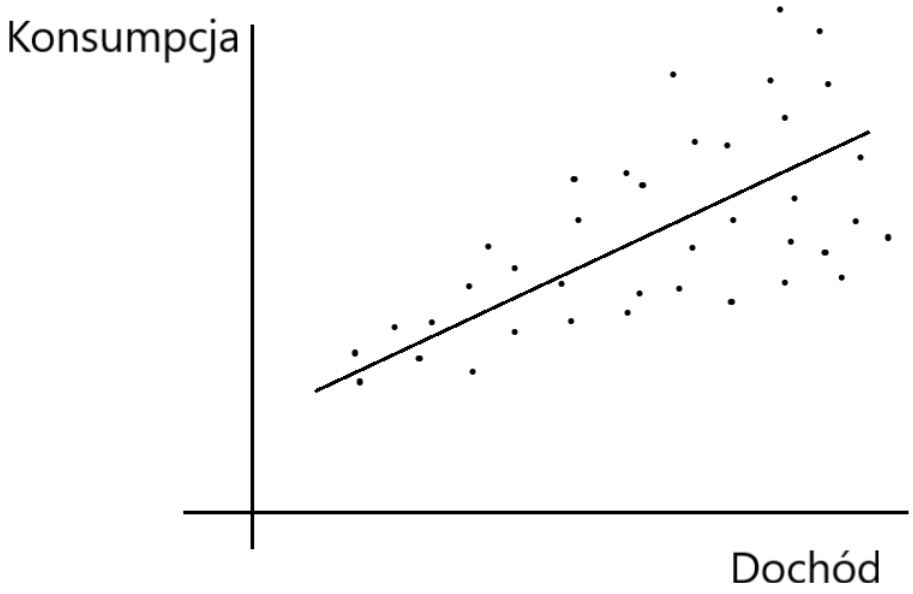
  $\rightarrow$ Tę własność jednorodności ("jednakowości") wariancji wszystkich składników losowych nazywamy **homoskedastycznością** (ang. *homoscedasticity, homoskedasticity*).

  $\rightarrow$ 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 5.a** – c.d. – przykłady:

Reszty homoskedastyczne	**Reszty heteroskedastyczne**
Niebieskie linie reprezentują pasmo: $\hat{y}_t \pm odchylenie\ resztowe\ (s)$ Rozrzut punktów danych wokół dopasowanej prostej jest (w przybliżeniu) jednakowy → Założenie 5.a KMNRL jest spełnione	Wraz ze wzrostem dochodu, wielkość wydatków na konsumpcję także wzrasta, ale wzrasta również jej rozrzut → Założenie 5.a KMNRL nie jest spełnione
	
Źródło: http://www3.wabash.edu/econometrics/EconometricsBook/chap19.htm	Źródło: opracowanie własne (dane umowne)

Konsekwencje niespełnienia założeń KMNRL o składnikach losowych

- **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 White'a (heteroskedastyczności składników losowych)

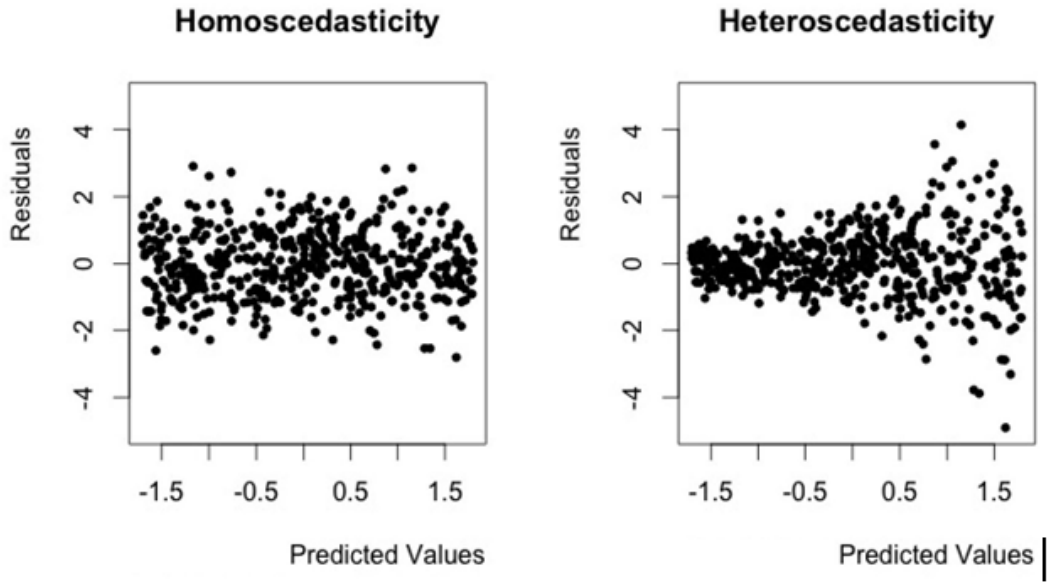
→ Chcemy rozsądzić, czy $Var(\varepsilon_t) = \sigma^2$ (nie zależy od t; homoskedastyczność – zgodnie z założeniami KM(N)RL), czy jednak należy dopuścić, że $Var(\varepsilon_t) = \sigma_t^2$ (heteroskedastyczność)

→ Szacujemy **wyjściowy (główny) model** regresji – zmienną objaśnianą jest y_t

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_K x_{tK} + \varepsilon_t, \qquad k = K + 1$$

→ Obliczamy reszty, $\hat{\varepsilon}_t = y_t - \hat{y}_t$, gdzie $\hat{y}_t = x_t\hat{\beta}$

→ Zanim przejdziemy do testu…



Źródło: Sage research methods

Test White'a (heteroskedastyczności składników losowych)

➢ **Układ hipotez:**

$$H_0: \forall_{t=1,\dots T}\ \sigma_t^2 = \sigma^2 \quad \text{vs.} \quad H_1: \exists_{\substack{t,s \\ t\neq s}}\ \sigma_t^2 \neq \sigma_s^2$$

→ Za pomocą **zwykłej MNK** szacujemy **pomocniczy model** regresji liniowej – dla kwadratów reszt wyjściowego modelu (tego głównego, zbudowanego dla zmiennej objaśnianej y_t):

$$\hat{\varepsilon}_t^2 = \theta_0 + \underbrace{\theta_1 x_{t1} + \theta_2 x_{t2} + \cdots \theta_K x_{t,K}}_{\text{część liniowa}} + \underbrace{\theta_{K+1} x_{t1}^2 + \theta_{K+2} x_{t2}^2 + \cdots + \theta_{K+K} x_{t,K}^2}_{\text{kwadraty zmiennych}}$$

$$+ \underbrace{\theta_{2K+1} x_{t1} x_{t2} + \cdots + \theta_{2K+\frac{K(K-1)}{2}} x_{t,K-1} x_{t,K}}_{\substack{\text{iloczyny mieszane zmiennych;} \\ \text{jest ich } \frac{K(K-1)}{2}}} + \eta_t$$

- w sumie liczba parametrów wynosi $1 + K + K + \frac{K(K-1)}{2} = 1 + \frac{K(K+3)}{2}$
- $\hat{\varepsilon}_t^2$ jest estymatorem (nieznanej!) σ_t^2
- η_t (czyt. *„eta te”*) – pomocniczy składnik losowy

→ Równoważny **układ hipotez**:

H_0: wszystkie parametry (z wyjątkiem θ_0) są równe 0 (czyli $\hat{\varepsilon}_t^2$ jest w przybliżeniu stałe) vs. $H_1: \neg H_0$

Test White'a (heteroskedastyczności składników losowych)

→ Dla uproszczenia zapisu przyjmijmy oznaczenia:

$$w_t = \theta_0 + \theta_1 z_{t1} + \theta_2 z_{t2} + \cdots + \theta_{\frac{K(K+3)}{2}} z_{t,\frac{K(K+3)}{2}} + \eta_t$$

$w_t \equiv \hat{\varepsilon}_t^2$ – zmienna objaśniana; z_{ti} –zmiennej objaśniające

→ Zapis skalarny i macierzowy: $w_t = z_t\theta + \eta_t$ $w = Z\theta + \eta$

$w = \begin{bmatrix} \hat{\varepsilon}_1^2 \\ \hat{\varepsilon}_2^2 \\ \vdots \\ \hat{\varepsilon}_T^2 \end{bmatrix}$ = „y" w regresji pomocniczej, Z = „X" w regresji pomocniczej, θ = wektor parametrów

→ Estymator MNK parametrów modelu pomocniczego: $\hat{\theta} = (Z'Z)^{-1}Z'w$

→ W modelu **pomocniczym** potrzebujemy obliczyć tylko wartość **współczynnika determinacji**:

$$R^2 = 1 - \frac{\sum_{t=1}^{T} \hat{\eta}_t^2}{\sum_{t=1}^{T}(w_t - \bar{w})^2}$$

$$\hat{\eta}_t = w_t - \hat{w}_t; \qquad \hat{w}_t = z_t\hat{\theta}; \qquad \bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$$

➢ **Statystyka testowa testu White'a:**

$$LM_{emp} = T \cdot R^2 | H_0 \sim \chi^2\left(\frac{K(K+3)}{2}\right)$$

$\frac{K(K+3)}{2}$ = liczba parametrów w modelu pomocniczym (**bez** wyrazu wolnego)

→ W sukurs przychodzi nam funkcja =REGLINP(…) (naprawdę „uff…" ;)

Test White'a (heteroskedastyczności składników losowych)

- **Wartość krytyczna – zawsze prawostronna:** χ^2_α

=ROZKŁAD.CHI.ODW(*prawdopodobieństwo* = α; *stopnie_swobody* = $\frac{K(K+3)}{2}$) (stara wersja)

=ROZKŁ.CHI.ODWR.PS(*prawdopodobieństwo* = α; *stopnie_swobody* = $\frac{K(K+3)}{2}$) (nowa wersja)

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

- ***p-value*:**

=ROZKŁAD.CHI(*x* = LM_{emp}; *stopnie_swobody* = $\frac{K(K+3)}{2}$) (stara wersja)

=ROZKŁ.CHI.PS(*x* = LM_{emp}; *stopnie_swobody* = $\frac{K(K+3)}{2}$) (nowa wersja)

- **Uwaga:** Czasami rozważa się „**zredukowaną wersję**" pomocniczego modelu regresji – bez uwzględniania iloczynów mieszanych:

$$\hat{\varepsilon}_t^2 = \theta_0 + \underbrace{\theta_1 x_{t1} + \theta_2 x_{t2} + \cdots \theta_K x_{t,K}}_{część\ liniowa} + \underbrace{\theta_{K+1} x_{t1}^2 + \theta_{K+2} x_{t2}^2 + \cdots + \theta_{K+K} x_{t,K}^2}_{kwadraty\ zmiennych} + \eta_t$$

→ W jakiej sytuacji warto zastosować takie podejście?

→ Wtedy

$$LM_{emp} = T \cdot R^2 | H_0 \sim \chi^2(2K)$$

Remedia – heteroskedastyczność składników losowych

- Typowe **przyczyny** występowania **heteroskedastyczności**:
  I. Na etapie specyfikacji równania modelu pominięto jakieś kluczowe zmienne objaśniające
  II. Nie zlogarytmowano zmiennej objaśnianej (i ewentualnie także zmiennych objaśniających)
  III. Zbiór danych obejmuje (zbyt) niejednorodne, heterogeniczne obserwacje (dotyczy danych przekrojowych)

  Inne…?

- Możliwe **sposoby poradzenia sobie z heteroskedastycznością**:
  I. Odpowiednio do pkt I powyżej – omówiono poniżej
  II. Odpowiednio do pkt II powyżej – omówiono poniżej
  III. Odpowiednio do pkt III powyżej – omówiono poniżej
  IV. EUMNK – omówimy w przyszłości

Remedia – heteroskedastyczność składników losowych

- Ad I. Na etapie specyfikacji równania modelu pominięto jakieś kluczowe zmienne objaśniające

  → Należy się zastanowić nad uwzględnieniem dodatkowych (ekonomicznych) zmiennych

  → Może być wskazane uwzględnienie w modelu **kwadratów i iloczynów mieszanych wyjściowych zmiennych objaśniających** – przykładowo, przy wyjściowo tylko dwóch zmiennych objaśniających otrzymalibyśmy równanie następującej postaci (podobnie jak w regresji pomocniczej testu White'a):

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 x_{t1}^2 + \beta_4 x_{t2}^2 + \beta_5 x_{t1} x_{t2} + \varepsilon_t$$

  Taki model jak najbardziej możemy oszacować za pomocą **zwykłej MNK**, gdyż spełnione jest założenie 1. KM(N)RL (równanie jest w dalszym ciągu *liniowe względem parametrów*).

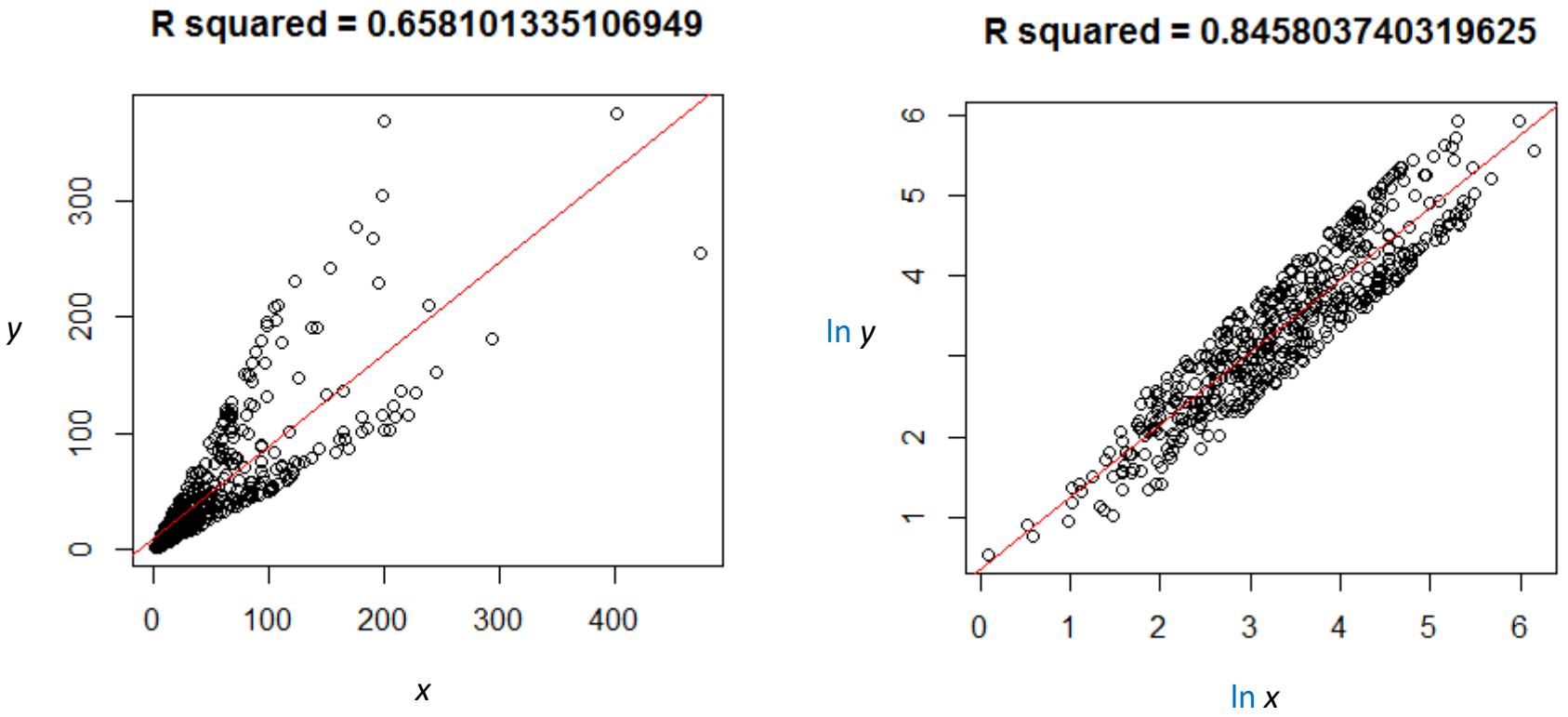
  Taki zabieg umożliwia pewne **przybliżenie potencjalnie *nieliniowych* zależności** pomiędzy zmienną objaśnianą a objaśniającymi, co może w pewien sposób „naprawić" błąd specyfikacji modelu, a w konsekwencji – przyczynić do poprawy własności reszt (nie tylko w odniesieniu do ich wariancji, ale np. także w kwestii normalności rozkładu, o czym też za chwilę). **Ceną**, jaką przychodzi nam jednak zapłacić za takie podejście jest utrata standardowej interpretacji ocen parametrów. Przykładowo, chcąc zinterpretować ocenę parametru β_1, stojącego przy x_{t1}, nie możemy już przyjąć założenia o ustalonych wartościach pozostałych zmiennych objaśniających, gdyż x_{t1} występuje jeszcze – w pewnych nieliniowych funkcjach – przy β_3 i β_5.

  Niemniej można by przejść na ogólniejszy poziom formułowania interpretacji, o którym tu jednak nie będzie mowy… :)

Remedia – heteroskedastyczność składników losowych

- Ad II. Nie zlogarytmowano zmiennej objaśnianej (i ewentualnie także zmiennych objaśniających)

  → No to… logarytmujemy :) Poniższy przykład idealnie obrazuje przypadek, gdzie logarytmizacja „y-ka" oraz „x-a" faktycznie pomaga.

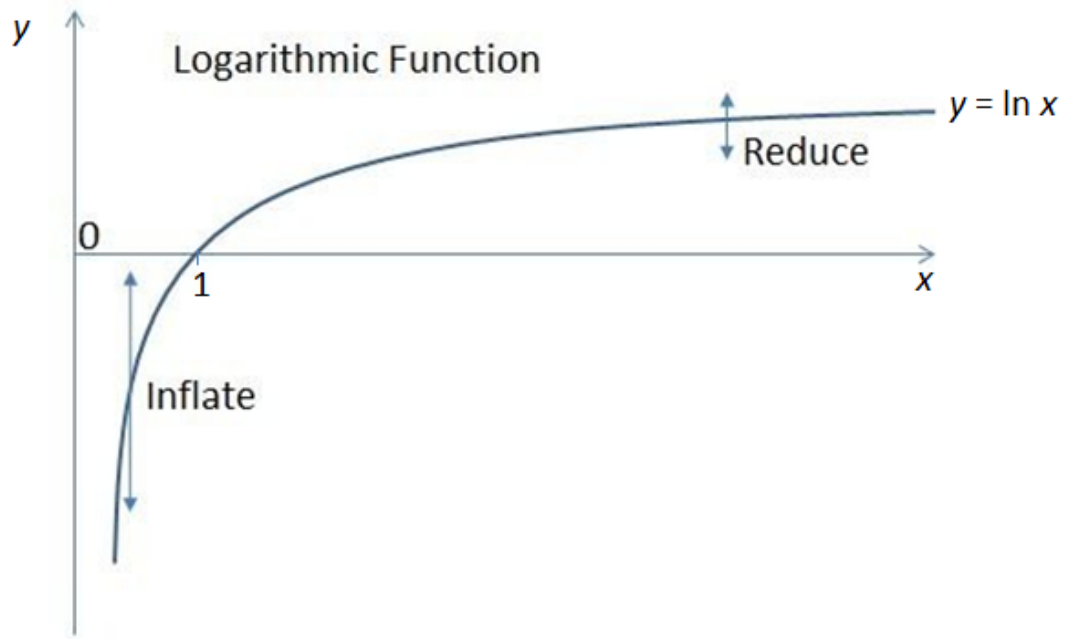


Źródło: https://stats.stackexchange.com (zmieniono etykiety osi)

Remedia – heteroskedastyczność składników losowych

→ ALE dlaczego transformacja logarytmiczna pomaga?

(Oczywiście, logarytmizacja jest możliwa tylko, gdy dana zmienna przyjmuje wartości dodatnie)



Źródło: https://blog.minitab.com/en/applying-statistics-in-quality-projects/how-could-you-benefit-from-a-box-cox-transformation (z drobnymi zmianami)

Remedia – heteroskedastyczność składników losowych

→ Dodatkowe argumenty przemawiające za stosowaniem **transformacji logarytmicznej**:

I. Zwróćmy uwagę, że w KMNRL $\varepsilon_t \in \mathbb{R}$ (ponieważ ε_t podlega rozkładowi normalnemu, którego nośnikiem[1] jest właśnie cały zbiór liczb rzeczywistych) oraz $y_t = x_t\beta + \varepsilon_t$. Wobec tego – co do zasady – zbiór wartości zmiennej objaśnianej to „całe" $\mathbb{R}$, a nie np. tylko $\mathbb{R}_+ = (0, +\infty)$. Zatem jeśli pewna zmienna ekonomiczna, dajmy na to A_t, przyjmuje wartości tylko dodatnie ($A_t \in \mathbb{R}_+$, w domyśle dla każdego $t = 1, 2, \ldots, T$), to $\ln A_t \in \mathbb{R}$. Wydaje się, że choćby z punktu widzenia **zgodności nośników** zmiennych: ogólnie $y_t \in \mathbb{R}$ w KMNRL, oraz rozważanej tu dla przykładu $\ln A_t \in \mathbb{R}$, to właśnie ta ostatnia, $\ln A_t$, powinna być zmienną objaśnianą, a nie bezpośrednio $A_t \in \mathbb{R}_+$.

II. Postać logarytmiczna zmiennej objaśnianej może także wynikać z pewnych modeli wywodzących się z **teorii ekonomii**. Przykładem jest tu choćby znany Państwu model funkcji produkcji Cobba i Douglasa, który jest przykładem tzw. *modelu potęgowego*. Alternatywnie rozważa się także tzw. *model wykładniczy*. Omówmy sobie teraz obydwie te specyfikacje (patrz kolejne slajdy), w ramach krótkiego, aczkolwiek ważnego wtrętu :)

[1] Nośnikiem danej zmiennej losowej (jej rozkładu prawdopodobieństwa) nazywamy taki zbiór jej wartości, w których funkcja gęstości / masy prawdopodobieństwa przyjmuje wartości *stricte* dodatnie (a nie równe zero). Przykładowo, nośnikiem rozkładu normalnego czy *t*-Studenta jest cała oś $\mathbb{R}$, podczas gdy dla rozkładów F czy χ^2 jest nim tylko pół oś dodatnich liczb rzeczywistych, $\mathbb{R}_+$. W jeszcze innym przypadku, nośnikiem może być po prostu pewien odcinek, jak na przykład w rozkładzie jednostajnym określonym na odcinku powiedzmy $[a, b]$ – nośnikiem jest właśnie ten odcinek; poza nim funkcja gęstości takiego rozkładu (przyjmująca stałą wartość w $[a, b]$) przyjmuje wartość zero.

Remedia – heteroskedastyczność składników losowych: model potęgowy

- Niech $A_t \in \mathbb{R}_+$ oznacza pewną zmienną ekonomiczną, której zmienność chcemy wyjaśnić za pomocą zmiennych objaśniających $Z_{ti} \in \mathbb{R}_+$.
- **Model potęgowy:**

$$A_t = \alpha_0 Z_{t1}^{\beta_1} Z_{t2}^{\beta_2} \dots Z_{tK}^{\beta_K} e^{\varepsilon_t}$$

→ Powyższe równanie nie spełnia założenia 1 KMNRL, gdyż: i) składnik losowy nie jest addytywny, ii) część deterministyczna modelu nie stanowi liniowej kombinacji parametrów i zmiennych objaśniających. Wobec tego logarytmujemy wyjściowe równanie i reparametryzujemy $\ln \alpha_0$, przyjmując $\beta_0 = \ln \alpha_0$ (wszak dowolna funkcja parametru też jest parametrem, czyli nieznaną stałą):

$$\underbrace{\ln A_t}_{y_t} = \beta_0 + \beta_1 \underbrace{\ln Z_{t1}}_{x_{t1}} + \beta_2 \underbrace{\ln Z_{t2}}_{x_{t2}} + \dots + \beta_K \underbrace{\ln Z_{tK}}_{x_{tK}} + \varepsilon_t$$

Tym samym model spełnia już założenie 1 KMNRL (daje się zapisać w postaci $y_t = x_t \beta + \varepsilon_t$).

→ Terminy „zmienna objaśniana" i „zmienna objaśniająca" odnoszą się do modelu *ekonometrycznego*, a nie wyjściowej (potęgowej) specyfikacji, zatem są tu nimi, odpowiednio, $y_t \equiv \ln A_t$ oraz $x_{ti} \equiv \ln Z_{ti}$.

→ Ponieważ $\beta_i = \frac{\partial \ln A_t}{\partial \ln Z_{ti}} \equiv El_{A_t|Z_{ti}}$ (dla $i = 1, \dots, K$), jego ocenę, $\hat{\beta}_i$, interpretujemy podług szablonu:

Jeżeli wartość zmiennej Z_{ti} (← bez logarytmu) *byłaby wyższa o 1%* (*przy ustalonych wartościach pozostałych zmiennych* $Z_{t1}, \dots, Z_{t,i-1}, Z_{t,i+1}, \dots, Z_{tK}$)*, wówczas wartość zmiennej* A_t *byłaby wyższa* ($\hat{\beta}_i > 0$)/niższa ($\hat{\beta}_i < 0$) *przeciętnie o* $|\hat{\beta}_i|$%. (← zapisano w module, bo w tym miejscu znak już pomijamy; wartości (modułu) oceny parametru NIE przemnażamy już przez 100%).

Remedia – heteroskedastyczność składników losowych: model wykładniczy

- Niech $A_t \in \mathbb{R}_+$ oznacza pewną zmienną ekonomiczną, której zmienność chcemy wyjaśnić za pomocą zmiennych objaśniających $Z_{ti} \in \mathbb{R}$ (Z_{ti} nie musi przyjmować wartości tylko w $\mathbb{R}_+$).
- **Model wykładniczy:**

$$A_t = \alpha_0 \alpha_1^{Z_{t1}} \alpha_2^{Z_{t2}} \dots \alpha_K^{Z_{tK}} e^{\varepsilon_t}$$

→ Powyższe równanie nie spełnia założenia 1 KMNRL – z tych samych dwóch powodów, co model potęgowy. Wobec tego logarytmujemy wyjściowe równanie i reparametryzujemy $\ln \alpha_i$ ($i = 0, 1, \dots, K$), przyjmując $\beta_i = \ln \alpha_i$:

$$\underbrace{\ln A_t}_{y_t} = \beta_0 + \beta_1 \underbrace{Z_{t1}}_{x_{t1}} + \beta_2 \underbrace{Z_{t2}}_{x_{t2}} + \dots + \beta_K \underbrace{Z_{tK}}_{x_{tK}} + \varepsilon_t$$

Tym samym model spełnia już założenie 1 KMNRL (daje się zapisać w postaci $y_t = x_t \beta + \varepsilon_t$).

→ Terminy „zmienna objaśniana” i „zmienna objaśniająca” odnoszą się do modelu *ekonometrycznego*, a nie wyjściowej (potęgowej) specyfikacji, zatem są tu nimi, odpowiednio, $y_t \equiv \ln A_t$ oraz $x_{ti} \equiv Z_{ti}$.

Remedia – heteroskedastyczność składników losowych: model wykładniczy

→ Jak interpretować ocenę parametru β_i, stojącego przy danej zmiennej Z_{ti} $(i = 1, 2, \dots, K)$?

Zauważmy, że jeżeli wartość zmiennej Z_{ti} wzrosłaby o 1 swoją jednostkę (przy ustalonych wartościach pozostałych zmiennych objaśniających), to wartość zmiennej objaśnianej, $\ln A_t$, zmieni się przeciętnie o (właśnie) $\hat{\beta}_i$ (wzrośnie, gdy $\hat{\beta}_i > 0$, lub spadnie, gdy $\hat{\beta}_i < 0$). Jak zmiana $\ln A_t$ o $\hat{\beta}_i$ przekłada się na zmianę bezpośrednio zmiennej ekonomicznej A_t? Rozpiszmy to sobie:

$$\ln A_t + \hat{\beta}_i = \ln A_t + \ln e^{\hat{\beta}_i} = \ln\left(A_t e^{\hat{\beta}_i}\right)$$

Zatem zmiana $\ln A_t$ o $\hat{\beta}_i$ jest równoznaczna z $e^{\hat{\beta}_i}$-krotną zmianą „samego" A_t. Ponieważ teraz mówimy o relatywnej (multiplikatywnej) zmianie wartości zmiennej A_t, z reguły lepiej będzie ją wyrazić w procentach. Przykładowo, gdyby czynnik $e^{\hat{\beta}_i}$ był równy 2, wówczas $A_t \cdot 2$ oznacza dwukrotny wzrost A_t, czyli wzrost o 100%. Podobnie, gdyby $e^{\hat{\beta}_i} = 0{,}7$, wówczas $A_t \cdot 0{,}7$ oznacza spadek wartości A_t o 30%. W ten sposób możemy zapisać ogólnie, że zmiana $\ln A_t$ o $\hat{\beta}_i$, powodując $e^{\hat{\beta}_i}$-krotną zmianą „samego" A_t, powoduje wzrost/spadek wartości A_t przeciętnie o $\left(e^{\hat{\beta}_i} - 1\right)100\%$.

Zauważmy, że zawsze $e^{\hat{\beta}_i} > 0$. Z drugiej strony, gdy $\hat{\beta}_i > 0$, wówczas $\left(e^{\hat{\beta}_i} - 1\right)100\% > 0$ (ponieważ $e^{\hat{\beta}_i} > 1$), natomiast przy $\hat{\beta}_i < 0$, wówczas $\left(e^{\hat{\beta}_i} - 1\right)100\% < 0$ (ponieważ $e^{\hat{\beta}_i} \in (0, 1)$). Oznacza to, że podobnie jak dotychczas – i zgodnie z intuicją – znak oceny parametru informuje nas bezpośrednio o kierunku zmiany (wzroście/spadku) wartości zmiennej A_t.

Remedia – heteroskedastyczność składników losowych: model wykładniczy

→ Podsumowując, ocenę parametru β_i w modelu wykładniczym interpretujemy według szablonu:

Jeżeli wartość zmiennej Z_{ti} (← bez logarytmu) *byłaby wyższa o 1 <u>jednostkę</u>* (*przy ustalonych wartościach pozostałych zmiennych* $Z_{t1}, \dots, Z_{t,i-1}, Z_{t,i+1}, \dots, Z_{tK}$*), wówczas wartość zmiennej* A_t *byłaby wyższa* ($\hat{\beta}_i > 0$)/niższa ($\hat{\beta}_i < 0$) *<u>przeciętnie</u> o* $\left|\left(e^{\hat{\beta}_i} - 1\right)100\right|$ %. (← zapisano w module, bo w tym miejscu znak już pomijamy).

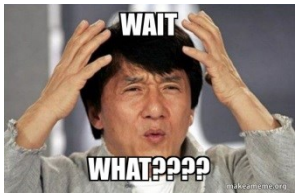
Remedia – heteroskedastyczność składników losowych: model wykładniczy

- Kiedy do modelowania pewnej zmiennej ekonomicznej możemy chcieć zastosować ***model wykładniczy*** (zmienna objaśniana – w postaci logarytmicznej; zmienne objaśniające – bez takiej transformacji, tj. w swojej wyjściowej postaci) – trzy główne powody:

1) W modelowaniu szeregów czasowych – gdy chcemy uwzględnić **trend deterministyczny** w modelu (przykładowo, w równaniu $\ln A_t = \beta_0 + \beta_1 t + \cdots + \varepsilon_t$ zmienna czasowa t jest BEZ logarytmu):
   - Gdy A_t przejawia wzrost wykładniczy, wówczas $\ln A_t$ charakteryzuje się trendem liniowym

   - Ze względu na interpretację oceny parametru przy trendzie – w powyższym równaniu, gdzie zmienna t jest niezlogarytmowana, moglibyśmy sformułować interpretację w stylu *„Z okresu na okres (przy ustalonych wartościach pozostałych zmiennych objaśniających), wartość zmiennej A_t wzrasta/spada (w zależności od znaku $\hat{\beta}_1$) przeciętnie o* $\left|\left(e^{\hat{\beta}_i} - 1\right)100\right|$ %.

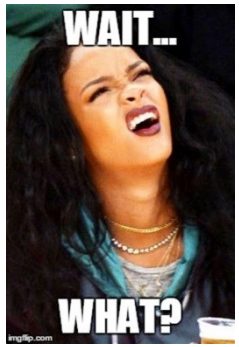
     → Zauważmy, że gdyby zmienną t uwzględnić w modelu w postaci zlogarytmowanej, tj. $\ln A_t = \beta_0 + \beta_1 \ln t + \cdots + \varepsilon_t$, wówczas parametr β_1 byłby elastycznością A_t względem czasu, a zatem początek interpretacji brzmiałby tak: *„Jeżeli wartość zmiennej czasowej wzrosłaby o 1% (przy ustalonych itd…) itd…”*, co jest bardzo nieintuicyjne, bo co niby miałby oznaczać *wzrost numeru okresu o 1%...?*



Remedia – heteroskedastyczność składników losowych: model wykładniczy

- Kiedy do modelowania pewnej zmiennej ekonomicznej możemy chcieć zastosować ***model wykładniczy*** (zmienna objaśniana – w postaci logarytmicznej; zmienne objaśniające – bez takiej transformacji, tj. w swojej wyjściowej postaci) – trzy główne powody (c.d.):

2) Jeżeli w modelu chcemy uwzględnić jakiś **regresor** będący zmienną ekonomiczną **wyrażoną w procentach**, np. stopę bezrobocia, stopę inflacji, stopę procentową – także z powodu interpretacji oceny stojącego przy niej parametru. Jeżeli NIE zlogarytmujemy takiej zmiennej, wówczas w intepretacji będziemy mogli mówić o wzroście jej wartości o 1 *punkt procentowy* (np. wzrost stopy bezrobocia z 10% do 11%). W przeciwnym razie, gdybyśmy zlogarytmowali taką zmienną, wtedy musielibyśmy przejść na interpretację elastycznościową, która jest raczej nieintuicyjna w przypadku zmiennych wyjściowo wyrażonych w procentach, gdyż wymagałaby rozważania wzrostu takiej zmiennej dosłownie o 1% (zamiast punktu procentowego), co oznacza konieczność wyobrażenia sobie „*procentu z procentu*”.



Model wykładniczy

- Kiedy do modelowania pewnej zmiennej ekonomicznej możemy chcieć zastosować ***model wykładniczy*** (zmienna objaśniana – w postaci logarytmicznej; zmienne objaśniające – bez takiej transformacji, tj. w swojej wyjściowej postaci) – trzy główne powody (c.d.):

3) Jeżeli w modelu chcemy uwzględnić jakiś regresor będący **zmienną licznikową**, tj. przyjmującą (z reguły niewielkie) wartości ze zbioru liczb naturalnych, np. liczba dzieci w rodzinie. Powodem, dla którego takiej zmiennej również nie poddajemy transformacji logarytmicznej jest znowu kwestia interpretacji – gdyby w sytuacji zmiennej objaśnianej w postaci zlogarytmowanej, zmienną objaśniającą licznikową również poddano logarytmizacji, wówczas parametr stojący przy tej ostatniej byłby elastycznością. Wtedy może okazać się „nienaturalne" mówienie o 1%-wym wzroście owej zmiennej licznikowej (bo np. cóż miałby oznaczać wzrost liczby dzieci w rodzinie o 1%??? :)

Remedia – heteroskedastyczność składników losowych: model potęgowo-wykładniczy

- **Model potęgowo-wykładniczy:**
  - Zmienna objaśniana – w logarytmie
  - Niektóre zmienne objaśniające – w logarytmie; pozostałe – w swojej wyjściowej postaci

$$A_t = \alpha_0 Z_{t1}^{\beta_1} Z_{t2}^{\beta_2} \dots Z_{tK_1}^{\beta_{K_1}} \alpha_1^{W_{t1}} \alpha_2^{W_{t2}} \dots \alpha_{K_2}^{W_{tK_2}} e^{\varepsilon_t} \quad /\ln \quad (\ln \alpha_i = \beta_{K_1+i}, \text{ dla } i = 1, 2, \dots, K_2; \ln \alpha_0 = \beta_0)$$

$$\underbrace{\ln A_t}_{y_t} = \beta_0 + \beta_1 \underbrace{\ln Z_{t1}}_{x_{t1}} + \cdots \beta_{K_1} \underbrace{\ln Z_{tK_1}}_{x_{tK_1}} + \beta_{K_1+1} \underbrace{W_{t1}}_{x_{t,K_1+1}} + \cdots \beta_{K_1+K_2} \underbrace{W_{tK_2}}_{x_{t,K_2+1}} + \varepsilon_t$$

Model potęgowo-wykładniczy – estymacja

Przypomnijmy

$$\underbrace{\ln A_t}_{y_t} = \beta_0 + \beta_1 \underbrace{\ln Z_{t1}}_{x_{t1}} + \cdots \beta_{K_1} \underbrace{\ln Z_{tK_1}}_{x_{tK_1}} + \beta_{K_1+1} \underbrace{W_{t1}}_{x_{t,K_1+1}} + \cdots \beta_{K_1+K_2} \underbrace{W_{tK_2}}_{x_{t,K_1+K_2}} + \varepsilon_t$$

Powyższy model spełnia już 1. założenie KMNLR (posiada addytywny składnik losowy, zaś część systematyczna jest liniowa względem parametrów). Zatem:

1) Da się zapisać w ogólnej postaci:

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_K x_{tK} + \varepsilon_t$$

gdzie:

- $K = K_1 + K_2$ (liczba parametrów BEZ wyrazu wolnego, β_0); $k = K + 1$ (… i Z wyrazem wolnym)
- $y_t \equiv \ln A_t$
- $x_{ti} \equiv \ln Z_{ti}$ dla $i = 1, \dots, K_1$
- $x_{ti} \equiv W_{ti}$ dla $i = K_1 + 1, \dots, K_1 + K_2$

2) jego parametry możemy oszacować za pomocą MNK: $\hat{\beta} = (X'X)^{-1}X'y$, przy czym:

$$y = \begin{bmatrix} \ln A_1 \\ \ln A_2 \\ \vdots \\ \ln A_T \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & \ln Z_{11} & \dots & W_{1,K_2} \\ 1 & \ln Z_{21} & \dots & W_{2,K_2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \ln Z_{T1} & \dots & W_{T,K_2} \end{bmatrix}_{(T \times k)}$$

Remedia – heteroskedastyczność składników losowych: model potęgowo-wykładniczy

→ A teraz wracamy już do omawiania możliwych przyczyn występowania i wynikających z nich sposobów zaradzenia heteroskedastyczności – **przypomnijmy** (ze slajdu nr 44):

- Typowe **przyczyny** występowania **heteroskedastyczności**:
  I. Na etapie specyfikacji równania modelu pominięto jakieś kluczowe zmienne objaśniające
  II. Niezlogarytmowano zmiennej objaśnianej (i ewentualnie także zmiennych objaśniających)
  III. Zbiór danych obejmuje (zbyt) niejednorodne, heterogeniczne obserwacje (dotyczy danych przekrojowych)

  Inne…?

  → Do omówienia pozostał nam jeszcze punkt III (i jeszcze EUMNK, ale to w osobnym odcinku)

Remedia – heteroskedastyczność składników losowych

- Ad III. Zbiór danych obejmuje (zbyt) niejednorodne, heterogeniczne obserwacje (dotyczy danych przekrojowych)

  Przykładowo, chcemy modelować **przychody przedsiębiorstw** w Polsce – w podziale przedsiębiorstw ze względu na wielkość wyróżnia się 4 grupy:

  1) mikroprzedsiębiorstwo – zatrudnienie średnioroczne pracowników to mniej niż 10 osób i przychody netto lub suma bilansowa są mniejsze lub równe 2 mln EUR
  2) małe przedsiębiorstwo – zatrudnienie średnioroczne pracowników to mniej niż 50 osób i przychody netto lub suma bilansowa są mniejsze lub równe 10 mln EUR
  3) średnie przedsiębiorstwo – zatrudnienie średnioroczne pracowników to mniej niż 250 osób i przychody netto są mniejsze lub równe 50 mln EUR lub suma bilansowa jest mniejsza lub równa 43 mln EUR
  4) duże przedsiębiorstwo – wszystkie pozostałe, które nie wpisują się w powyższe kryteria.

  → Różna wielkość zarówno samych przychodów przedsiębiorstw w poszczególnych grupach, jak i – co właśnie problematyczne – wariancji tych przychodów

  → Modelowanie próby przedsiębiorstw z różnych grup „w jednym worku” *vs.* modelowanie każdej grupy z osobna

  → Ponadto, przedsiębiorstwa danej wielkości, ale z różnych sektorów/branż także mogą się charakteryzować niejednorodną wariancją…

  → Jakie jeszcze inne przykłady niejednorodnych zbiorów danych możecie podać?