MIKROEKONOMETRIA
ZAJĘCIA 2

WIKTOR
BUDZIŃSKI

STATYSTYKA TESTOWA

Statystyką testową nazywamy wartość policzoną na pewnej próbie, którą możemy wykorzystać do testowania hipotez statystycznych

- Zazwyczaj jest ona zaprojektowana w ten sposób żeby przy spełnieniu hipotezy zerowej miała ona znany rozkład
  - Jeśli hipoteza zerowa nie jest spełniona to będzie ona odstawała od przewidywanego rozkładu
  - Często wymaga to pewnych dodatkowych założeń o naszych danych

P-value

- P-value danego testu statystycznego mówi nam o **prawdopodobieństwie**, że zaobserwujemy daną wartość statystyki testowej zakładając, że hipoteza zerowa jest spełniona
  - Jeśli to prawdopodobieństwo jest niskie oznacza, że jest mała szansa zaobserwowania takiej statystyki przy hipotezie zerowej
    - Wskazuje to, że hipoteza zerowa może nie być spełniona
  - Dla wielu statystyk tożsame z prawdopodobieństwem popełnienia błędu I rodzaju

CASE STUDY – DETERMINANTY CEN WINA

- Costanigro, M., Mittelhammer, R. C., and McCluskey, J. J. (2009). ***Estimating class-specific parametric models under class uncertainty: local polynomial regression clustering in an hedonic analysis of wine markets***. Journal of Applied Econometrics, 24(7), 1117-1135.
  - Dane dotyczące cen i charakterystyk win ze stanów Kalifornia i Waszyngton w latach 1991-2000

ZADANIE 1

Przeprowadź wstępną analizę danych

1. Przedstaw dane na wykresach
   A. Narysuj histogram zmiennej *Price*
   B. Narysuj scatter plot zależności między zmienną *Price* a *Cases* i *Score*
   C. Narysuj zależność między *Price* a typem wina
2. Policz korelację między *Price*, *Cases* oraz *Score*
   A. Czy są one istotne?
3. Czy średnie ceny wina różnią się między stanami Kalifornia i Waszyngton
   A. Przetestuj to formalnie
4. Czy w stanach Waszyngton i Kalifornia produkują wina różnego typu?

Klasyczny Model Regresji Liniowej (KMRL)

- Postać modelu regresji liniowej: $y_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i$
  - Modelujemy liniową zależność y od zmiennych objaśniających $\mathbf{X}$
  - Parametry modelu $\boldsymbol{\beta}$ są nieznane, ale możemy je oszacować za pomocą posiadanych danych $y, \mathbf{X}$
- Estymator – reguła stosowana w celu znalezienia oszacowania jakiejś wielkości na podstawie danych
- Estymacja parametrów – Metoda Najmniejszych Kwadratów (MNK)
  - Metoda szukania parametrów $\boldsymbol{\beta}$, polegająca na tym, żeby nasza liniowa funkcja dawała jak najmniejsze kwadraty różnic między przewidywanymi a obserwowanymi wartościami
  - Wizualizacja

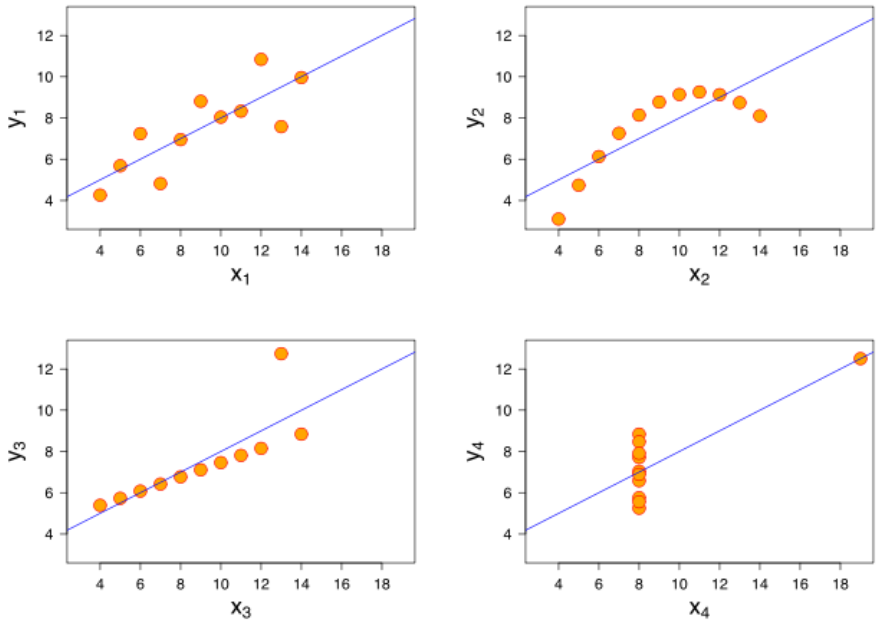
METODA NAJMNIEJSZYCH KWADRATÓW

Jeśli spełnione są określone założenia (wymogi) to estymator MNK

1. Nieobciążony (wartość oczekiwana szacowanych parametrów będzie równa prawdziwej wartości)
2. Zgodny (im większa próba, tym szacowane parametry będą bliższe prawdziwej wartości)
3. Efektywny (estymator jest efektywny gdy ma najmniejszą wariancję ze wszystkich estymatorów)

ZAŁOŻENIA KMRL

1. Liniowa postać funkcji jest faktycznie prawidłowa



- Jeśli tak nie jest
  - Można estymować model nieliniowy (innymi metodami estymacji niż MNK)
  - Można spróbować dokonać jakiejś transformacji zmiennych objaśnianej lub objaśniających (np. logarytm, kwadrat), która da zależność bliższą liniowej

ZAŁOŻENIA KMRL

2. Zmienne objaśniające są liniowo niezależne
   - Żadnej ze zmiennych objaśniających nie da się wyrazić za pomocą liniowej kombinacji pozostałych
   - Zwykle zakłada się także, że ich średnie i wariancje są skończone
   - Jeśli jakieś zmienne objaśniające są współliniowe – nie da się dla nich określić jednoznacznych parametrów

ZAŁOŻENIA KMRL

3. Błędy są sferyczne $\operatorname{var}\left(\varepsilon \middle| \mathbf{X}\right) = \sigma^2 I_n$
   - Brak autokorelacji $\underset{i \neq j}{\forall} E\left(\varepsilon_i \varepsilon_j \middle| \mathbf{X}\right) = 0$ czyli $\operatorname{cov}\left(\varepsilon_i, \varepsilon_j\right) = 0$
     - Błędy pomiędzy obserwacjami są nieskorelowane
     - Może być niespełnione np. dla szeregów czasowych, danych panelowych itp.
     - Jeśli tak nie jest – estymator MNK nadal jest nieobciążony i zgodny, ale jest nie już efektywny
       - Można zastosować np. Uogólnioną MNK
   - Homoskedastyczność
     - Składnik losowy ma tę samą wariancję dla wszystkich obserwacji $E\left(\varepsilon_i^2 \middle| \mathbf{X}\right) = \sigma^2$
     - Jeśli tak nie jest – estymator MNK nadal jest nieobciążony i zgodny, ale jest nie już efektywny
       - Można zastosować np. Ważoną MNK lub odporne macierze kowariancji
   - Czasem zakłada się dodatkowo normalność składnika losowego
     - Nie jest niezbędne, ale jest potrzebne do dowodów własności testów dla skończonych prób
     - Jeśli błędy są normalne to estymator MNK jest ekwiwalentny do estymatora Metody Największej Wiarygodności (MNW)

ZAŁOŻENIA KMRL

4. Zmienne objaśniające są egzogeniczne
   - Są znane z pewnością, nie trzeba traktować ich jak zmienne losowe (np. z powodu błędów pomiaru)
   - Nie są skorelowane z błędem losowym $E\left(\varepsilon \middle| \mathbf{X}\right) = 0$
   - Czyli (m.in.) błędy losowe mają średnią 0
   - Jeśli tak nie jest (są endogeniczne, czyli korelacja y i $\mathbf{X}$ jest możliwa naturalnie, bez zależności przyczynowo-skutkowej) estymator MNK przestaje być zgodny
     - Można zastosować metodę zmiennych instrumentalnych (znaleźć zmienne, które wpływają na $\mathbf{X}$, ale nie mają wpływu na y)
     - Można zastosować metodę analizy wielomianowej (ang. *multivariate analysis*), w której dowolna zmienna może być traktowana jako objaśniana

ZADANIE 2

Wykorzystując dane *me_wine*.dta przygotuj model regresji hedonicznej dla cen wina

1. Przeprowadź regresję objaśniającą ceny wina skupiając się na wpływie liczbie wyprodukowanych skrzynek wina oraz ocenie jego smaku
2. Przeprowadź podstawowe testy diagnostyczne

Analiza hedoniczna – analiza przy pomocy regresji, w której wartość dobra (przybliżona ceną) jest rozbita na wartość poszczególnych charakterystyk tego dobra

BŁĘDY STANDARDOWE

Błąd standardowy jest miarą niepewności oszacowania parametrów modelu

- Im większa jego wartość tym mniejsza precyzja oszacowań