Metody Ekonometryczne
Zadania

**Zadanie 1. Wyniki egzaminu**

Plik `Historical.csv` zawiera informacje o indywidualnych wynikach z przedmiotu Ekonometria. W zbiorze danych uwzględniono następujące charakterystyki:

| Nazwa | Opis |
|---|---|
| `Exam` | Liczba punktów z egzaminu końcowego |
| `Final` | Łączna liczba punktów, a więc suma punktów uzyskanych na egzaminie, jak i podczas ćwiczeń |
| `Sex` | Płeć studenta (`F` -kobieta; `M` - mężczyzna) |
| `Year` | Uporządkowana zmienna jakosciowa dla roku egzaminu |
| `Summer` | Zmienna binarna (`1` – egzamin w semestrze letnim) |
| `Instructor` | Zmienna jakościowa określająca identyfikator osoby prowadzącej ćwiczenia |
| `Start` | godzina rozpoczęcia ćwiczeń |

(i) Zaimportuj dane do pakietu `R`. Spróbuj odpowiedzieć czy są systematyczne różnice pomiędzy kobietami a Mężczyznami w wyniku egzaminu końcowego. Porównaj wartości średnie oraz odchylenie standardowe wyniku w tych grupach. Zilustruj różnice korzystając z histogramu (`hist()`) lub szacunków funkcji gęstoci (`density()`).

(ii) Przeprowadź podobną analizę dla poszczególnych dla rozkładu wyników w zależności od osoby prowadzącej zajęcia. Pamiętaj przy tym, że informacja o osobie prowadzącej zajęcia jest dostępna w postaci zmiennej jakościowej.

(iii) Przeprowadź analogiczne ćwiczenie dla innych charakterystyk, tj. `Year`, `Summer` i `Start`. Które z dostępnych charakterystyk w największym stopniu różnicują rozkład wyników egzaminu?

(iv) Oszacuj parametry modelu regresji liniowej, w której zmienną objaśnianą jest wynik egzaminu, a zmiennymi objaśniającymi są zmienne binarne określające osobę prowadzącą ćwiczenia. Wytłumacz uzyskane oszacowania na podstawie wcześniejszych wyników. Dlaczego nie można uwzględnić wszystkich zmiennych binarnych określających osobę prowadzącą ćwiczenia?

(v) Rozszerz specyfikację modelu o płeć i godzinę rozpoczęcia zajęć. Zinterpretuj uzyskane oszacowania. Porównaj uzyskane oszacowania wnioskami płynącymi z analizy warunkowej rozkładów zmiennych.

(vi) Można oczekiwać, że uzyskany wynik z egzaminu końcowego może również zależeć od indywidualnych zdolności czy nakładu pracy studenta podczas semestru. Na podstawie dostępnęgo zbioru danych, spróbuj skonstruować zmienną, która może dobrze mierzyć te charakterystyki. Następnie, oszacuj parametry modelu z poprzedniego punkty, którego specyfikacja została dodatkowo rozszerzona o tę zmienną. Zinterpretuj uzyskane wynik.

**Zadanie 2.** Rozważ następujące równanie dla płac:

$$\ln w_i = \beta_0 + \beta_1 educ_i + \varepsilon_i \tag{1}$$

Gdzie $w_i$ to płaca (za godzinę w USD), a $educ_i$ to liczba lat edukacji.

(i) Na podstawie zbioru danych `CPSSWEducation` w pakiecie `AER` oszacuj parametry modelu (1). Jako $w_i$ wykorzystaj `earnings`, a jako $educ_i$ użyj zmiennej `education`. Zinterpretuuj uzyskane oszacowanie parametru $\beta_1$.

(ii) Czy oszacowanie parametru $\beta_1$ jest statystycznie istotne? Odpowiedź uzasadnij.

(iii) Zinterpretuj ekonomicznie, a następnie przetestuj następującą hipotezę:

$$\mathcal{H}_0: \quad \beta_1 = .1,$$

(iv) Przetestuj następująca hipotezę:

$$\mathcal{H}_0: \quad \beta_1 > .08,$$

(v) Przeprowadź następujący eksperyment.

(a) Wylosuj ze zbioru danych 2950 obserwacji ze zwracaniem [*wskazówka: wykorzystaj funkcję* `sample()`].

(b) Oszacuj MNK parametry równania (1) na podstawie losowania z wczesniejszego punktu i zachowaj oszacowanie parametru $\beta_1$ w pamięci.

(c) Powtórz powyższe punkty 1000-krotnie (można użyć większej liczby).

Na podstawie uzyskanych w poszczególnych iteracjach (tj. losowaniach) oszacowań naszkicuj wykres gęstości tych oszacowań oraz oblicz średnią i odchylenie standardowe. Porównaj te liczby z wynikami regresji z punktu (i).

(vi) Na podstawie rozkładu uzyskanych oszacowań policz:

(a) W ilu % przypadków uzyskane oszacowania są większe od 0.

(b) W ilu % przypadków uzyskane oszacowania są większe od 0.08.

Czy powyższe wyniki można porównać z punktami (i)-(iii)?

**Zadanie 3. Model grawitacyjny eksportu**

Rozważ następujący model grawitacyjny opsiujący bilateralny eksport:

$$\ln(EX_{ij}) = \beta_0 + \beta_1 \ln(GDP_i) + \beta_2 \ln(GDP_j) + \beta_3 \ln(dist_{ij}) + \varepsilon_{ij} \tag{2}$$

gdzie indeksy $i$ i $j$ oznaczają reporterów/eksporterów oraz partnerów handlowych, $EX_{ij}$ to eksport z $i$-tej do $j$-tej gospodarki, $GDP_i$ to PKB $i$-tej gospodarki, $dist_{ij}$ to odległość geograficzna między $i$-tą a $j$-tą gospodarką oraz $\varepsilon_{ij}$ to składnik losowy.

Zaimportuj następujące dane `Export.csv`. Zbiór ten zawiera następujące charakterystyki:

| Nazwa | Opis |
|---|---|
| `Reporter` | Kod ISO gospodarki eksportującej |
| `Partner` | KOD ISO partnera handlowego |
| `Export` | Eksport brutto towarów w tys. USD; Źródło: UN Comtrade |
| `GDP_Reporter` | PKB gospodarki eksportującej w USD, Źródło: WDI |
| `GDP_Partner` | PKB partnera handlowego w USD, Źródło: WDI |
| `dist` | dystans geograficzny pomiędzy eksporterem a partnerem handlowym, Źródło: CEPII |
| `POP_Reporter` | Populacja gospodarki eksportującej, Źródło: WDI |
| `POP_Partner` | Populacja partnera handlowego, Źródło: WDI |

Odpowiedz na poniższe pytania:

(i) Przenalizuj rozkład Eksportu brutto oraz logarytmu tej zmiennej. W którym przypadku można się spodziewać, że założenie o normalności składnika losowego w modelu regresji liniowej będzie spełnione?

(ii) Oszacuj parametry równania (2). Zinterpretuj uzyskane oszacowania. Czy uzyskane oszacowania są statystycznie istotne?

(iii) Zinterpretuj $R^2$.

(iv) Wykorzystując test RESET przenalizuj poprawność postaci funkcyjnej.

(v) Rozważ następującą hipotezę liniową:

$$\beta_2 = -\beta_3, \tag{3}$$

Zinterpretuj ekonomicznie powyższą hipotezę oraz zweryfikuj ją empirycznie.

(vi) Analogicznie jak w poprzednim punkcie przeanalizuj następującą hipotezę:

$$\beta_1 = \beta_2 = -\beta_3, \tag{4}$$

(vii) Skonstruuj (dwie) zmienne binarne opisujące przynależność eksportera lub partnera do UE, a następnie rozszerz specyfikację modelu (2) o te zmienne. Zinterpretuj uzyskane oszacowania. W jaki sposób się one różnią od Twoich podstawowych oszacowań? Przetestuj czy oszacowania parametrów przy dwóch nowych zmiennych są od siebie statystycznie istotnie różne. Zinterpretuj ekonomicznie wynik tego testu.

**Zadanie 4.** Rozważ raz jeszcze model grawitacyjny (2), który został przedyskutowany w zadaniu 3. Wykorzystując dane w pliku `Export.csv` oszacuj parametry modelu (2).

(i) Czy zmienne objaśniające są współliniowe?

(ii) Przetestuj normalność składnika losowego.

(iii) Przenalizuj wykresy kwadratów reszt względem zmiennych objaśniających w modelu (2). Czy wykresy te wskazują na heteroskedastyczność składnika losowego. Jeżeli tak, to przedyskutuj istotę heteroskedastyczności w tym przypadku.

(iv) Przeprowadź test Goldfelda-Quandta na heteroskedastyczność składnika losowego wykorzystując zmienną $dist_{ij}$ jako charakterystykę, która wpływa na wariancję składnika losowego. Czy wynik tego testu potwierdza wcześniejszą analizę z (iii)?

(v) Przeprowadź test White'a z kwadratami i interakcjami zmiennych objaśniających. Przedyskutuj wyniki regresji pomocniczej i skonfrontuj je z analizą z (iii).

(vi) Oszacuj wariancję parametrów wykorzystując estymator odporny na heteroskedastyczność. Porównaj wyniki testów istotności w przypadku wykorzystania tego estymatora oraz w przypadku założenia o homoskedastyczności składnika losowego.

(vii) Oszacuj parametry (2) ważoną MNK. *Teoretyczną* wariancję składnika losowego wyznacz na podstawie regresji pomocniczej, w której zmienną objaśnianą jest logarytm naturalny kwadratów reszt, a zmiennymi objaśniającymi – zmienne objaśniające w równaniu (2) oraz ich kwadraty. Porównaj uzyskane oszacowania z wcześniejszymi.

(viii) Na podstawie oszacowań uzyskanych MNK, MNK z wykorzystaniem odpornych na heteroskedastyczność błędów standardowych oraz ważoną MNK przetestuj hipotezę omówioną w punkcie (v) w zadaniu 3. Przedyskutuj różnice.

**Zadanie 5** (Efektywność estymatora WLS, tj. ważonej metody najmniejszych kwadratów)**.** Rozważ następujący proces generujący dane DGP (*data generating process*) dla danych przekrojowych $y_i$,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{5}$$

gdzie $x_i$ to zmienna objaśniająca, $\beta_0$ to wyraz wolny, $\beta_1$ to parametr strukturalny oraz $\varepsilon_i$ to składnik losowy. Załóż, że składnik losowy jest heteroskedastyczny, tj. $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$, oraz, że wariancja składnika losowa ma następującą postać:

$$\sigma_i^2 = \sigma^2 h(x_i) = \exp(\kappa x_i)\sigma^2. \tag{6}$$

(i) Załóż, że $\beta_0 = 1.5$, $\beta_1 = 1$, $x_i \sim \mathcal{N}(2, 2)$, $\kappa = 1$ oraz $\sigma = 1$. Przeprowadź 1000 symulacji dla danych przekrojowych o liczbie obserwacji równej 250 i wyznacz parametry $\beta_0$ i $\beta_0$. Porównaj rozkład uzyskanych oszacowań punktowych z prawdziwymi wartościami, a więc założeniami o $\beta_1$ i $\beta_1$.

(ii) Dla założeń z poprzedniego punktu rozszerz ćwiczenie symulacyjne tak, aby za każdym razem szacować parametry równania (i) MNK oraz (i) ważoną MNK ze znanymi wagami (a więc korzystając z zależności, tj. $\sigma_i^2 = x_i\sigma^2$), oraz (iii) ważoną MNK z nieznanymi wagami (tj. *szacowanymi* wagami, a więc uzyskanymi na postawie odpowiedniej regresji, w której zmienną objaśnianą są kwadraty reszt). Porównaj rozkłady uzyskanych oszacowań w zależności od wykorzystanej metody.

(iii) Przeprowadź powyższe obliczenia dla różnych wartości $\kappa$, np. $\kappa \in \{0.5, 2\}$. Czy Twoje wnioski z poprzedniego punktu pozostają te same ?

(iv) Przeprowadź obliczenia z punktu (ii) dla różnych wielkości zbioru danych, np. $N \in \{30, 100, 10000\}$. Czy Twoje wnioski z punktu (ii) pozostają te same ?

**Zadanie 6** (Autokorelacja)**.** Używając danych *USMacroG* z pakietu *AER* oszacuj prosty model popytu na pieniądz postaci:

$$\log \frac{M}{P} = \beta_0 + \beta_1 i + \beta_2 \log Y, \tag{7}$$

gdzie $\frac{M}{P}$ jest miarą realnego pieniądza w gospodarce, $i$ jest stopą procentową, a $Y$ realnym produktem

(i) Czy $i$ to realna czy nominalna stopa procentowa? Dlaczego w specyfikacji nie została ona zlogarytmowana? Uzasadnij.

(ii) Zinterpretuj uzyskane szacunki

(iii) Narysuj wykres reszt $\epsilon_t$ w czasie oraz scatterplot reszt $\epsilon_t$ jako funkcję opóznionych reszt $\epsilon_{t-1}$. Co z nich wynika? Jaka jest korelacja między tymi zmiennymi losowymi?

(iv) Przetestuj reszty testami: Durbina-Watsona, Ljunga-Boxa oraz Breuscha-Godfreya. Jaki rząd dla testów powinno się dobrać? Czy yesty te potwierdzają wyniki analizy z poprzedniego punktu.

(v) Zastosuj test Ljunga-Boxa w pętli dla opóznienia od 1 do 16, zaspisując jedynie p-values z każdej iteracji. Jakie są wnioski?

(vi) Przeprowadź wnioskowanie o istotności parametrów w korygując macierz wariancji metodą Neweya-Westa. Czy wnioski się zmieniają?

(vii) Dokonaj estymacji równania (7) przy wykorzystaniu metody Cochrane'a-Orcutt'a (ustaw parametr convergence = 6). Porównaj otrzymane wyniki z początkowymi.

(viii) Jakie może być źródło autokorelacji reszt? Przejrzyj wykresy zmiennych, które są używane w modelu. Jak można przekształcić zmienne lub model, aby ograniczyć problem autokorelacji?

**Zadanie 7** (Efektywność estymatora FGLS Cochrane'a -Orcutta)**.** Rozważ następujący proces generujący dane DGP (*data generating process*) dla szeregu czasowego $y_t$,

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \tag{8}$$

gdzie $x_t$ to zmienna objaśniająca, $\beta_0$ to wyraz wolny, $\beta_1$ to parametr strukturalny oraz $\varepsilon_t$ to składnik losowy. Załóż również, że składnik losowy charakteryzuje się autokorelacją pierwszego rzędu, tj.

$$\varepsilon_t = \rho\varepsilon_{t-1} + \eta_t, \tag{9}$$

gdzie $\rho$ to parametr autoregresyjny a $\eta_t$ to idiosynkratyczne zaburzenie losowe. Wiadomo, że $\eta_t \sim \mathcal{N}\left(0, \sigma_\eta^2\right)$.

(i) Załóż, że $\beta_0 = 1.5$, $\beta_1 = 1$, $x_t \sim \mathcal{N}(2, 2)$, $\rho = 0.8$ oraz $\sigma_\eta = 1$. Przeprowadź 1000 symulacji dla szeregu czasowego o długości 100 obserwacji i wyznacz parametry $\beta_0$ i $\beta_1$. Porównaj rozkład uzyskanych oszacowań punktowych z prawdziwymi wartościami, a więc założeniami o $\beta_0$ i $\beta_1$.

(ii) Dla założeń z poprzedniego punktu rozszerz ćwiczenie symulacyjne tak, aby za każdym razem szacować parametry równania MNK oraz iterowaną UMNK (metodą Cochrane'a -Orcutta dostępna w pakiecie `orcutt`). Porównaj rozkłady uzyskanych oszacowań w zależności od wykorzystanej metody.