Ekonometria Dynamiczna

Model liniowy jako narzędzie analizy szeregów czasowych

mgr Paweł Jamer[1]

24 marca 2015

[1] pawel.jamer@gmail.com

Model regresji wielorakiej

Równanie modelu:

$$Y_t = \beta_0 + \sum_{i=1}^{k} \beta_i X_{i,t} + \epsilon_t,$$

dla $t = 1, 2, \ldots, T$.

Założenia:

- $\epsilon_t \sim \mathcal{N}\left(0, \sigma^2\right), t = 1, 2, \ldots, T$;
- ϵ_t oraz ϵ_s są niezależne dla dowolnych s i t takich, że $s \neq t$;
- ϵ_t oraz x_s są niezależne dla dowolnych s i t;
- x_t oraz x_s są nieskorelowane dla dowolnych s i t takich, że $s \neq t$.

Model regresji wielorakiej (macierzowo)

Równanie modelu:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

gdzie

- $\boldsymbol{Y} = \begin{bmatrix} Y_1 & Y_2 & \cdots & Y_T \end{bmatrix}'$ — wektor odpowiedzi;
- $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 & \beta_1 & \dots & \beta_k \end{bmatrix}'$ — wektor parametrów;
- $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1, \epsilon_2, \dots, \epsilon_T \end{bmatrix}'$ — wektor błędów;
- $\boldsymbol{X} = \begin{bmatrix} 1 & X_{1,1} & X_{2,1} & \cdots & X_{k,1} \\ 1 & X_{1,2} & X_{2,2} & \cdots & X_{k,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1,T} & X_{2,T} & \cdots & X_{k,T} \end{bmatrix}$ — macierz eksperymentu.

Metoda Najmniejszych Kwadratów (MNK)

Idea MNK Chcemy tak dobrać parametry $\beta_0, \beta_1, \ldots, \beta_k$ modelu regresji wielorakiej, aby minimalizować sumaryczną odległość punktów $Y_1, Y_2, \ldots, Y_T$ od krzywej regresji.

Funkcja celu MNK

$$Q(\beta) = (\boldsymbol{Y} - \boldsymbol{X}\beta)'(\boldsymbol{Y} - \boldsymbol{X}\beta) \to \min .$$

Estymator MNK

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} .$$

Właściwości MNK

Estymator MNK jest **nieobciążony**, tzn.

$$\mathbb{E}\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}.$$

Macierz kowariancji estymatora MNK przyjmuje postać

$$\mathrm{Cov}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

Twierdzenie Gaussa-Markowa

W modelu regresji wielorakiej estymator MNK jest estymatorem o minimalnej wariancji w klasie estymatorów liniowych nieobciążonych.

Twierdzenie

W modelu regresji wielorakiej estymator MNK jest równy estymatorowi największej wiarogodności.

Residua

Wartości dopasowane

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\beta}.$$

Na potrzeby teorii:

$$\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y},$$

gdzie

- $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$ — macierz daszkowa.

Residua

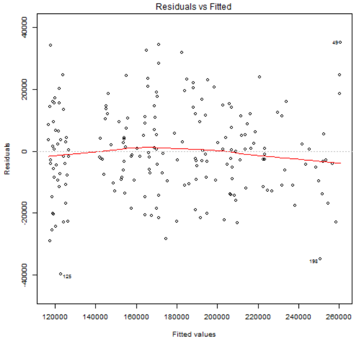
$$\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$$

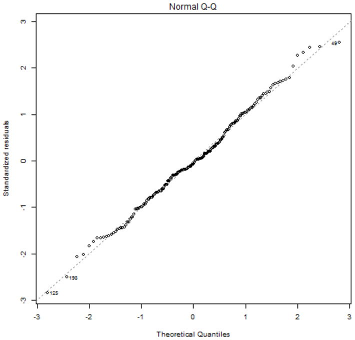
- $\mathbb{E}\left(\boldsymbol{e}\right) = \boldsymbol{0}$,
- $\mathrm{Cov}\left(\boldsymbol{e}\right) = \sigma^2\left(\boldsymbol{I} - \boldsymbol{H}\right)$.

Intuicja Residua $\boldsymbol{e}$ to próbkowy odpowiednik błędów modelu ϵ.

Normalność rozkładu (graficznie)

- Residua w funkcji wartości dopasowanych. Oczekujemy
  - średniej równej zero,
  - jednorodnej wariancji.
- Wykres kwantylowy standaryzowanych residuów. Oczekujemy
  - ułożenia się obserwacji na linii $y = x$.





Normalność rozkładu (test)

Niech $e_1, e_2, \ldots, e_T \sim F$.

Pytanie

Czy dystrybuanta F jest dystrybuantą rozkładu normalnego $F_{\mathcal{N}}$?

Testujemy hipotezę:

$$\begin{cases} H_0: & F = F_{\mathcal{N}} \\ H_1: & F \neq F_{\mathcal{N}} \end{cases}$$

Test Shapiro-Wilka

$$W = \frac{\left[\sum_{t=1}^{[T/2]} a_{T-t+1} \left(e_{T-t+1:T} - e_{t:T}\right)\right]^2}{\sum_{t=1}^{T} e_t^2}$$

Zaleta Test Shapiro-Wilka jest mało wrażliwy na typowe problemy regresji.

Współczynnik determinacji

Definiujemy:

- $\text{SST} = \sum_{t=1}^{T} \left(Y_t - \overline{Y}\right)^2$,
- $\text{SSR} = \sum_{t=1}^{T} \left(\hat{Y}_t - \overline{Y}\right)^2$,
- $\text{SSE} = \sum_{t=1}^{T} \left(Y_t - \hat{Y}_t\right)^2$.

Zachodzi:

$$\text{SST} = \text{SSR} + \text{SSE}.$$

Obserwacja Jeśli wykres rozproszenia silnie skupia się wokół prostej regresji, to

$$\text{SSE} \ll \text{SST}.$$

Współczynnik determinacji

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

Istotność równania regresji

Pytanie

Czy równanie regresji ma sens jako całość?

Testujemy hipotezę:

$$\begin{cases} H_0: & \hat{\beta}_1 = \hat{\beta}_2 = \ldots = \hat{\beta}_k = 0 \\ H_1: & (\exists i)\, \hat{\beta}_i \neq 0 \end{cases}$$

Test F

$$F = \frac{SSR/k}{SSE/(T-k-1)} \sim \mathbb{F}^{[k,T-k-1]}$$

Istotność parametrów

Pytanie

Czy zmienna X_i ma istotny wpływ na zmienną Y?

Testujemy hipotezę:

$$\begin{cases} H_0: & \hat{\beta}_i = 0 \\ H_1: & \hat{\beta}_i \neq 0 \end{cases}$$

Test t

$$t_i = \frac{\hat{\beta}_i}{\mathrm{SE}_{\hat{\beta}_i}} \sim t^{[T-k-1]}$$

Współliniowość

Problem Jeśli wśród zmiennych objaśniających występują zmienne współliniowe, to macierz $\boldsymbol{X}'\boldsymbol{X}$ nie jest odwracalna.

Czynnik inflacji wariancji

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2},$$

gdzie

- R_i^2 — współczynnik determinacji w modelu, w którym zmienną objaśnianą jest X_i, a zmiennymi objaśniającymi zmienne $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_k$.

Heurystyka Kiedy $\mathrm{VIF}_j > 10$, to zmienna j wykazuje dużą współliniowość ze zmiennymi pozostałymi.

Skorygowany współczynnik determinacji

Problem Współczynnik R^2 rośnie wraz z dokładaniem kolejnych zmiennych do modelu.

Skorygowany współczynnik determinacji

$$\overline{R}^2 = 1 - \frac{\mathrm{SSE}/(T-k-1)}{\mathrm{SST}/(T-1)}$$

Kryteria informacyjne

Problem Wzrost liczby obserwacji T wywiera zbyt silny wpływ na skorygowany wpsółczynnik determinacji $\overline{R}^2$.

Kryterium informacyjne Akaike (AIC)

$$\text{AIC} = T \ln \frac{\boldsymbol{e}'\boldsymbol{e}}{T} + 2\,(k+1)$$

Bayesowskie kryterium informacyjne (BIC)

$$\text{BIC} = T \ln \frac{\boldsymbol{e}'\boldsymbol{e}}{T} + (k+1)\ln T$$

Selekcja modelu

Pytanie

Jak ze zbioru zmiennych objaśniających $X_1, X_2, \ldots, X_k$ wybrać w sposób automatyczny te, które w najlepszy sposób opisują zmienną odpowiedzi Y?

- **forward**: zaczynamy od modelu pustego i w kolejnych krokach wybieramy spośród zmiennych jeszcze nie wybranych tą, która ma najbardziej pozytywny wpływ na kryterium.
- **backward**: zaczynamy od modelu pełnego i w kolejnych krokach odrzucamy spośród zmiennych znajdujących się w modelu tą, której brak wywrze najbardziej istotny wpływ na kryterium informacyjne.
- **both**: naprzemienne stosowanie metod poprzednich.

Predykcja

Niech:

- $\hat{\beta}$ — wektor estymatorów MNK regresji wielorakiej.
- $\boldsymbol{X}_{new}$ — macierz zawierająca nowe wartości zmiennych objaśniających, dla których nie są znane wartości zmiennej objaśnianej.

Problem Chcemy wyznaczyć prawdopodobne wartości zmiennych objaśnianych $\hat{Y}_{new}$

Predykcja

$$\hat{Y}_{new} = \boldsymbol{X}_{new}\hat{\beta}$$

Przedział ufności dla predykcji:

$$\hat{Y}_{new} - t_{\alpha/2}^{[T-k-1]}\hat{\mathrm{SE}}_{new} \leqslant Y_{new} \leqslant \hat{Y}_{new} + t_{\alpha/2}^{[T-k-1]}\hat{\mathrm{SE}}_{new}$$

Pytania?

Dziękuję za uwagę!