Testy diagnostyczne

Ekonometria WNE

Sebastian Zalas

University of Warsaw
s.zalas@uw.edu.pl

Heteroskedastycznoość ε

- oznacza brak stałości wariancji. Jej obecność w modelu może prowadzić do błędnego wnioskowania statytycznego
- Można wykryć ją stosując test Breusha-Pagana albo test White'a. Hipoteza zerowa obu testów mówi o homoskedastyczności składnika losowego.
- Jeśli składnik losowy cechuje się heteroskedastycznością należy zastosować odporny estymator wariancji estymatora, np. estymator White'a.

Uwzględnienie nieistotnej zmiennej

- Przyjmijmy poniższą specyfikację modelu:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

  lecz tak naprawdę x_2 nie ma wpływu na y, wtedy $\beta_2 = 0$.

- Co się stanie gdy w regresji zostanie umieszczona zmienna x_2?
  - nieobciążoność pozostaje nienaruszona: $\mathbb{E}[\hat{\beta}_2] = 0$
  - wariancja $\mathbb{V}(\hat{\beta}_1) = \frac{\sigma^2}{(1-R_1^2)\sum_{i=1}^{n}(x_{i,1}-\bar{x}_1)^2}$

- wariancja estymatora $\tilde{\beta}_1$ bez uwzględniania x_2:

$$\mathbb{V}(\tilde{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_{i,1} - \bar{x}_1)^2}$$

Uwzględnienie nieistotnej zmiennej

- Porównajmy wariancje: $\mathbb{V}(\tilde{\beta}_1) < \mathbb{V}(\hat{\beta}_1)$
- Usunięcie nieistotnych zmiennych powinno poprawić precyzję oszacowań
- Uwaga! Sama obserwacja, że zmienna jest (lub nie) istotna statystycznie, nie wystarczy, aby stwierdzić, że zmienna należy (lub nie należy) do modelu.
- Jeśli istnieją argumenty teoretyczne, zmienną należy zachować.

Test RESET

- *Regression Specification Error Test* - test formy funkcyjnej

$$H_0 : \text{liniowa postać modelu jest prawidłowa}$$
$$H_1 : \text{nieliniowa postać modelu jest prawidłowa}$$

- Procedura testowa: szacujemy model i uzyskujemy z niego wartości dopasowane $\hat{y}$.

- Następnie szacujemy pomocniczy model w postaci:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \beta_{k+1}\hat{y}^2 + \beta_{k+2}\hat{y}^3 + \ldots + \beta_{k+p}\hat{y}^p + \nu$$

  Zwykle uwzględniamy $\hat{y}^2$ i $\hat{y}^3$

Test RESET

- Testujemy następujące restrykcje:
  $$\beta_{k+1} = \beta_{k+2} = \ldots = \beta_{k+p} = 0$$
  korzystając z statystyki F która ma rozkład $F(1-\alpha, p, n-k-1)$

- Odrzucenie H_0 wskazuje na to, że liniowa forma modelu nie jest poprawna $\Rightarrow$ problem z założeniem KMRL o liniowości modelu

- Należy spróbować włączyć do modelu czynniki uwzględniające nieliniową zależność

Współliniowość

- Współliniowość deterministyczna - występuje gdy jedna ze zm. objaśniających jest kombinacją liniową innych zmiennych objaśniających

- Np.: x_1 oraz $x_2 = 2 \times x_1 \Rightarrow \hat{\beta}^{MNK}$ jest niezdefiniowany, ponieważ $rz(\boldsymbol{X}^\top \boldsymbol{X})$ nie jest pełny

- Współliniowość stochastyczna pojawia się gdy występuje <u>silna korelacja</u> między dwiema zmiennymi objaśniającymi.

Współliniowość

Objawy współliniowości:

- małe zmiany w danych powodują duże wahania oszacowań
- Oszacowania mogą mieć wysokie błędy standardowe i niskie poziom istotności, nawet gdy cały model jest istotny statystycznie a R^2 jest wysoki
- Współczynniki mogą mieć nieintuicyjny znak lub nieracjonalną wielkość

Współliniowość

- Szacujemy model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$
- Wariancja estymatora $\hat{\beta}_1$

$$\mathbb{V}(\hat{\beta}_1) = \frac{\sigma^2}{(1 - R_1^2)\sum_{i=1}^{n}(x_{i,1} - \bar{x}_1)^2}$$

  gdzie R_1^2 to R^2 z regresji $x_1 \sim x_2$

- Im silniej x_1 i x_2 są skorelowane to $\mathbb{V}(\hat{\beta}_1$ rośnie

Współliniowość

- Do wykrywania współliniowości służy wskaźnik inflacji wariancji, VIF:
$$VIF_i = \frac{1}{1 - R_i^2}$$
$VIF_i > 10$ sugeruje istnienie silniej współliniowości.

- Można sprawdzić korelację między zmiennymi objaśniającymi.

- Współliniowość w modelu można zmniejszyć poprzez np. usunięcie zmiennej z wysokim VIF.

Test normalności składnika losowego

- W KMRL składnik losowy ma rozkład normalny. Jeśli tak nie jest, wnioskowanie statystyczne może być błędne.

- Aby przetestować to założenie stosujemy test Jarque-Bery. W istocie sprawdza, czy dane mają skośność (S) i kurtozę (K) pasujące do rozkładu normalnego.

$$S = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n}\sum(x_i - \bar{x}_i)^3}{(\frac{1}{n}\sum(x_i - \bar{x}_i))^{\frac{3}{2}}}$$

$$K = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{n}\sum(x_i - \bar{x}_i)^4}{(\frac{1}{n}\sum(x_i - \bar{x}_i)^2)^2}$$

Test normalności składnika losowego

- Hipoteza zerowa:

$$H_0 : \begin{cases} S = 0 \\ K = 3 \end{cases}$$

  jeśli jest prawdziwa, składnik losowy ma rozkład normalny.

- Statystyka testowa:

$$JB = \frac{n}{6}\left(S + \frac{1}{4}(K-3)^2\right) \sim \chi^2_n$$

  ma rozkład chi kwadrat z n stopniami swobody

- Odrzucenie H_0 może wskazywać na niespełnienie założenia o normalności składnika losowego. Stanowi to problem w małych próbach.

Obserwacje nietypowe

- Jakie obserwację są nietypowe?
  - takie, które odznaczają się dużą różnicą między danymi a wartościami dopasowanymi
  - takie, które silnie różnią się od pozostałych obserwacji
- Warto sprawdzić czy zaburzają oszacowanie modelu:
  - dźwignia, standaryzowane reszty, dystans Cook'a
- Co zrobić?:
  - obserwacje błędne - wyrzucić
  - obserwacje nietypowe - sprawdzić skąd się wzięły; sprawdzić, czy oszacowania są bardzo wrażliwe na nie
  - zastosowanie logarytmów będzie zmniejszało wpływ obserwacji odstających
  - można również przyciąć rozkłady zmiennych po np. 1% z obu stron

Stabilność parametrów - test Chow'a

- Specyfikując model zakładamy, że jego założenia są spełnione dla wszystkich obserwacji.
- Możemy sprawdzić czy parametry są stabilne w całej próbie, czyli czy nie różnią się w podpróbach.
- Szacujemy modele na podpróbach i testujemy czy ich parametry są takie same:

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$$
$$H_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$$

  gdzie $\boldsymbol{\beta}_1$ i $\boldsymbol{\beta}_2$ to wektory parametrów modeli dla dwóch podprób.

Stabilność parametrów - test Chow'a

- Statystyka testowa:

$$F = \frac{\frac{RSS - RSS_1 - RSS_2}{k+1}}{\frac{RSS_1 + RSS_2}{n-2(k+1)}} \sim F(k+1, n-2(k+1))$$

  gdzie:
  - RSS - suma kwadratów reszt modelu szacowanego na całej próbie
  - RSS_1 - suma kwadratów reszt modelu szacowanego na podpróbie #1
  - RSS_2 - suma kwadratów reszt modelu szacowanego na podpróbie #2

- Jeśli H_0 zostanie odrzucona, należy rozważyć np.: oszacowanie modeli na podróbach lub dołączenie do modelu odpowiedniej zm. binarnej.

Pytania? Wątpliwości?

Dziękuję!

e: s.zalas@uw.edu.pl