Problemas de Especificação

Profa. R. Ballini

Bibliografia:

- Wooldridge, J. M. (2006). Introductory Econometrics – A Modern approach, Chapter 9.

Problemas Adicionais de Especificação e de Dados

Considere o modelo de regressão linear múltipla:

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \ldots + \beta_k X_{kj} + u_j \tag{1}$$

o qual é formulado com base nas seguintes hipóteses:

1) O modelo de regessão é linear nos parâmetros;
2) Os valores de X são fixos;
3) $E(u_j|X_1, \ldots, X_k) = 0$;
4) Ausência de colinearidade perfeita;
5) $Var(u_j|X_1, \ldots, X_k) = \sigma^2$ constante;
6) $cov(u_j, u_i||X_1, \ldots, X_k) = 0, \forall i \neq j$
6) O termo erro u tem distribuição normal.

Problemas Adicionais de Especificação e de Dados

- Se u for correlacionado com x, então x é uma **variável explicativa endógena**.
- Quando uma variável omitida é uma função de uma variável explicativa, há **má especificação da forma funcional**.
- A omissão de uma variável importante pode causar correlação entre o erro e variáveis explicativas, o que pode gerar viés e inconsistência em estimadores MQO.

Problema de Especificação

- Omissão de funções de variáveis independentes não é a única maneira de um modelo sofrer o problema de má especificação da forma funcional.

- Se for necessário utilizar o logaritmo da variável dependente, mas a utilizamos em sua forma original, não obteremos estimadores não-viesados ou consistentes dos efeitos parciais.

- Há testes para detectar esse tipo de problema da forma funcional.

Testes de Erros de Especificação

Considere um modelo com k variáveis independentes:

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \ldots + \beta_k X_{kj} + u_j \qquad (2)$$

Forma simples de verificar se X_k é relevante é testar a significância do b_k estimado com o teste t usual:

$$t = \frac{b_k - \beta_k}{S(b_k)}$$

Para testar se X_3 e X_4 são relevantes ao modelo, devemos usar teste de restrição (F ou LM) e testar a hipótese $\beta_3 = \beta_4 = 0$.

Erro de Especificação – Exemplo

A partir dos dados do arquivo hprice.xlsx, estime o modelo de regressão linear múltipla:

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u$$

em que:

price: preço da casa (em milhares de dólares);

lotsize: tamanho do terreno (em pés^2);

sqrft: área construída (em pés^2);

bdrms: número de quartos;

Adicione os termos quadráticos das variáveis *lotsize* e *sqrft* e analise o modelo.

Erro de Especificação – Exemplo

	Dependent variable:	
	PRICE	
	Linear	Quadrado
	(1)	(2)
LOTSIZE	0.00206771***	0.01144773***
	(0.00064213)	(0.00200699)
SQRFT	0.12277820***	0.01778915
	(0.01323741)	(0.06166963)
BDRMS	13.85252000	20.19265000**
	(9.01014500)	(7.96136000)
I(LOTSIZEˆ2)		−0.00000011***
		(0.00000002)
I(SQRFTˆ2)		0.00001504
		(0.00001284)
Constant	−21.77031000	36.28324000
	(29.47504000)	(74.58646000)
Observations	88	88
R^2	0.67236220	0.75810560
Adjusted R^2	0.66066090	0.74335600
F Statistic	57.46023000*** (df = 3; 84)	51.39820000*** (df = 5; 82)
Note:		*p<0.1; **p<0.05; ***p<0.01

Teste Reset (Regression Specification Test) (Ramsey(1969)[1]

Seja a regressão:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + u \tag{3}$$

Hipótese do teste:

$$H_0 : \quad E(u|x_1, \ldots, x_k) = 0$$

Funções não-lineares de variáveis independentes não devem ser relevantes quando acrescentadas em (3).

---

[1] Ramsey, J. B. *Tests for Specification Errors in Classical Linear Least Squares Regression Analysis*. Journal of the Royal Statistical Society, serie B, vol. 31, pp. 350–371, 1969.

Teste Reset

1. Para implementarmos o teste RESET, teremos que decidir sobre quais funções não-lineares das variáveis explicativas deveremos incluir na equação a ser expandida;

2. Não existe uma resposta direta para esta pergunta.

3. Segundo Ramsey (1969), a inclusão de termos quadráticos e cúbicos se mostrou bastante adequado em várias aplicações.

Teste Reset

Sejam $\hat{Y}_i$ os valores estimados, segundo o modelo proposto em (3). Segundo Ramsey (1969), a partir da equação expandida, dada por:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \delta_1 \hat{Y}^2 + \delta_2 \hat{Y}^3 + u \qquad (4)$$

poderemos testar se em (3) existem problemas de especificação na forma funcional (ausências de não linearidades importantes).

Teste Reset

Sob H_0, o modelo (3) foi especificado corretamente.

Teste RESET nada mais é do que a implementação de um teste de restrição nos coeficientes, que usa a estatística F para avaliar

$$H_0 : \delta_1 = \delta_2 = 0$$

no modelo expandido (4).

A significância desta estatística sugere algum problema com não-linearidades no modelo original.

A distribuição da estatística F, sob H_0 e admitindo a validade das suposições de 1 a 5, é aproximadamente $F_{[2;n-(k+2)]}$, para amostras grandes.

Teste de Davidson-MacKinnon (1981) (Modelos *Non Nested*)

Modelos contendo conjuntos distintos de regressores

Exemplo: analisar se uma variável independente deve estar em nível ou na forma logarítmica. Ou seja,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u \tag{5}$$

Contra o modelo:

$$Y = \beta_0 + \beta_1 log(X_1) + \beta_2 log(X_2) + u \tag{6}$$

Davidson-MacKinnon (1981): se o modelo (5) se mantiver com $E(u|X_1, X_2) = 0$, então os valores estimados do modelo (6) são não significantes em (5).

Logo, estima-se (6) por MQO, para obter os valores estimados $\check{y}$.

Teste é baseado na estatística t sobre $\check{y}$

Teste de Davidson-MacKinnon (1981) (Modelos *Non Nested*)

Ou seja,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \theta \check{y} + erro \tag{7}$$

Como os valores de $\check{y}$ são funções não lineares de X_1 e X_2, $\check{y}$ deve ser não significativo se (5) é o modelo adequado.

Logo, uma estatística t significante é uma rejeição de (5).

Aplicação

A partir dos dados do arquivo hprice1.wf1, estime o modelo em que todas as variáveis estão em forma de nível;

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u$$

Após estime o modelo em que as variáveis *price, lotsize* e *sqrft* estão em logaritmos, ou seja,

$$price = \beta_0 + \beta_1 log(lotsize) + \beta_2 log(sqrft) + \beta_3 bdrms + u$$

A partir do teste de Davidson e MacKinnon, verifique qual o modelo adequado.

Critérios para Seleção de Modelos

1. R-Quadrado

$$R^2 = \frac{SQReg}{SQTotal} = 1 - \frac{SQRes}{SQTotal}$$

- Comparar R^2 de dois ou mais modelos: variável dependente deve ser a mesma e o número de variáveis independentes deve ser o mesmo

2. R-Quadrado Ajustado

$$\bar{R}^2 = 1 - \frac{SQRes/(n-k)}{SQTotal/(n-1)} = 1 - (1 - R^2)\frac{n-1}{n-k}$$

- $\bar{R}^2 \leq R^2$
- $\bar{R}^2$ pune o acréscimo de novos regressores
- $\bar{R}^2$ é mais adequado que o R^2 para comparação de dois ou mais modelos com número de variáveis independentes distintos, mas a variável dependente deve ser a mesma.

Exemplo - Adequação do modelo

	Dependent variable:	
	PRICE	
	Linear	Log
	(1)	(2)
LOTSIZE	0.00206771***	
	(0.00064213)	
SQRFT	0.12277820***	
	(0.01323741)	
log(LOTSIZE)		61.45707000***
		(12.30436000)
log(SQRFT)		224.97340000***
		(29.84882000)
BDRMS	13.85252000	19.35056000**
	(9.01014500)	(8.84913500)
Constant	−21.77031000	−2,026.41700000***
	(29.47504000)	(209.33610000)
Observations	88	88
R^2	0.67236220	0.67779720
Adjusted R^2	0.66066090	0.66629000
F Statistic (df = 3; 84)	57.46023000***	58.90179000***
Note:		*p<0.1; **p<0.05; ***p<0.01

3. Critério de Informação de Akaike

$$AIC = \frac{2k}{n} + log\left(\frac{SQRes}{n}\right)$$

- Comparar dois ou mais modelos: menor valor de AIC

Critério para Seleção de Modelos

4. Critério de Informação de Schwarz

$$SIC = \frac{k}{n} log(n) + log\left(\frac{SQRes}{n}\right)$$

- Comparar dois ou mais modelos: menor valor de SIC