ACH3657

Métodos Quantitativos para Avaliação de Políticas Públicas

Aula teórica 08

Regressão Múltipla

Alexandre Ribeiro Leichsenring

alexandre.leichsenring@usp.br

Organização

Grau de Ajuste

Assim como na regressão simples, podemos definir:

$$\text{SQT} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad (1)$$

$$\text{SQE} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \quad (2)$$

$$\text{SQR} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (3)$$

Usando o mesmo argumento no caso da regressão simples, podemos mostrar que:

$$\text{SQT} = \text{SQE} + \text{SQR}$$

Coeficiente de determinação (R^2)

Exatamente como no caso da regressão simples,

$$R^2 = \frac{\text{SQE}}{\text{SQT}} = 1 - \frac{\text{SQR}}{\text{SQT}}$$

Observações sobre R^2

- $R^2 = [\text{Cor}(y_i,\ \hat{y}_i)]^2$
- R^2 nunca diminui quando adicionamos outra variável independente na regressão
- O fato de R^2 nunca diminuir faz dele um instrumento fraco para decidir se uma variável deve ou não ser adicionada ao modelo
- O fator que deve determinar se uma variável pertence ao modelo é se a variável tem na *população* um efeito parcial sobre y diferente de zero
- Mais à frente, trataremos de como testar essa hipótese (inferência estatística)
- Veremos também que, quando usado apropriadamente, R^2 permite-nos testar um grupo de variáveis com a finalidade de ver se ele é importante para explicar y

Exemplo: Determinantes de nmgrad

Da regressão de *nmgrad*, a equação resultante é:

$$\hat{nmgrad} = 1,29 + 0,453\ nmem + 0,0094\ tac$$

$$n = 141,\ R^2 = 0,176$$

- *nmem* e *tac* explicam juntos cerca de 17,6% da variação em *nmgrad*
- Isso pode não parecer uma porcentagem alta mas há muitos outros fatores – incluindo formação da família, personalidade, qualidade da educação do ensino médio, afinidade com o curso escolhido – que contribuem para o desempenho dos estudantes
- Se *nmem* e *tac* explicassem toda a variação em *nmgrad*, então o desempenho no curso superior seria predeterminado pelo desempenho no ensino médio!

Hipótese RLM.1 (Linear nos parâmetros)

No modelo populacional, a variável dependente y está relacionada à variável independente x e ao erro (ou perturbação) u como

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

- $\beta_0, \beta_1, \ldots, \beta_k$ são os parâmetros desconhecidos do modelo
- u é o erro aleatório não-observável ou um termo de perturbação aleatória

Hipótese RLM.2 (Amostragem Aleatória)

Temos uma amostra aleatória de n observações

$$\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : 1 = 1, 2, ..., n\}$$

proveniente do modelo populacional descrito na Hipótese RLM.1

Hipótese RLM.3 (Média condicional zero)

O erro u tem um valor esperado igual a zero, dados quaisquer valores das variáveis independentes:

$$\mathbf{E}(u|x_1, x_2, \ldots .x_k) = 0$$

Mal especificação implica, geralmente, em violação:

- Omitir um fator importante que está correlacionado com qualquer uma das variáveis $x_1, x_2, \ldots, x_k$ faz com que a hipótese RLM.3 não se sustente
- Se esquecemos de incluir o termo quadrático $rend^2$ na estimação da função consumo:

$$cons = \beta_0 + \beta_1\ rend + \beta_2\ rend^2 + u$$

- Se usamos o nível da variável e, de fato, é o log da variável que aparece no modelo populacional, ou vice-versa
- Há outros modos pelos quais u pode estar correlacionado com uma variável explicativa (como o problema do erro de medida)

➤ Quando a hipótese RLM.3 se mantém, dizemos que temos variáveis explicativas **exógenas**.

➤ Se x_j for correlacionado com u por alguma razão, então se diz que x_j, é uma variável explicativa **endógena**

Hipótese RLM.4 (Colinearidade não perfeita)

Na amostra (e, portanto, na população), nenhuma das variáveis independentes é constante, e não há relações lineares *exatas* entre as variáveis independentes.

- A hipótese de colinearidade não perfeita somente diz respeito às variáveis independentes
- Estudantes de econometria iniciantes tendem a confundir as hipóteses RLM.4 e RML.3

  ➤ RLM.4 não diz nada sobre a relação entre u e as variáveis explicativas

- A hipótese RLM.4 é mais complicada que sua contrapartida na regressão simples (envolve as relações entre todas as variáveis independentes)
- Se uma variável independente é uma combinação linear exata de outras variáveis independentes, dizemos que o modelo sofre de *colinearidade perfeita*

  ➤ Nesse caso: não pode ser estimado por MQO

Observações

- A hipótese RLM.4 permite que as variáveis independentes sejam correlacionadas

  ➤ Elas apenas não podem ser correlacionadas perfeitamente.

- Se não permitíssemos correlação entre variáveis independentes, regressão múltipla não seria muito útil para a análise econométrica
- A maneira mais simples como duas variáveis independentes podem ser perfeitamente correlacionadas é quando uma variável é um múltiplo da outra

  ➤ Pode acontecer quando um pesquisador, inadvertidamente, coloca a mesma variável medida em unidades diferentes dentro da equação de regressão.

- Outra maneira de as variáveis independentes serem perfeitamente colineares ocorre quando uma variável independente pode ser expressa como uma função linear exata de duas ou mais das outras variáveis independentes:

$$votoA = \beta_0 + \beta_1 gastoA + \beta_2 gastoB + \beta_3 gastototal + u$$

- A solução é simples: retire qualquer uma das três variáveis do modelo.

Teorema

Sob as hipóteses RLM.1 a RLM.4:

$$\mathbf{E}(\hat{\beta}_j) = \beta_j, j = 1, \ldots, k$$

para qualquer valor do parâmetro populacional β_j.

➤ Em outras palavras, os estimadores de MQO são estimadores não-viesados dos parâmetros da população

Observações

- Quando dizemos que MQO é não-viesado sob as hipóteses RLM. 1 a RLM.4, estamos dizendo que o *procedimento* pelo qual as estimativas de MQO foram obtidas é não-viesado
- Esperamos que tenhamos obtido uma amostra que nos dê uma estimativa próxima do valor da população, mas isso não pode ser garantido.

Inclusão de Variáveis Irrelevantes em um Modelo de Regressão

- Vamos avaliar a inclusão de uma variável irrelevante no modelo: incluída no modelo mas sem efeito parcial sobre y na população (seu coeficiente populacional é zero.)
- Suponha que especificamos o modelo:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

  e que ele satisfaz as hipóteses RLM.1 a RLM.4. mas que x_3 não tem efeito sobre y uma vez que x_1 e x_2 estão controlados (ou seja, $\beta_3 = 0$)
- x_3 pode ou não ter correlação com x_1 e x_2
- Em termos de esperanças condicionais:

$$\mathbf{E}(y|x_1, x_2, x_3) = \mathbf{E}(y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- Como não sabemos que $\beta_3 = 0$, somos inclinados a estimar a equação com x_3:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

Qual o efeito de incluir x_3 quando seu efeito populacional é zero?

- Em termos da inexistência de viés de $\hat{\beta}_1$ e $\hat{\beta}_2$ não há nenhum efeito!
- A inexistência de viés implica $\mathbf{E}(\hat{\beta}_j) = \beta_j$ para qualquer valor de β_j, incluindo $\beta_j = 0$!

Conclusão

- Incluir uma ou mais variáveis irrelevantes no modelo de regressão múltipla (superespecificar o modelo) não afeta a inexistência de viés dos estimadores de MQO
- Porém, incluir variáveis irrelevantes não é inócuo:

  $\Rightarrow$ pode ter efeitos indesejáveis sobre as variâncias dos estimadores de MQO (veremos mais adiante)

- Suponha omitimos uma variável que, realmente, pertence ao modelo verdadeiro (ou populacional)
- É o chamado problema de excluir uma variável relevante ou de *subespecificar o modelo*
- Suponha que o modelo populacional tem duas variáveis explicativas relevantes:

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

- Suponha que nosso interesse esteja β_1
- Suponha que somente executamos somente uma regressão simples de y sobre x_1:

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$$

- Neste caso:

$$\mathbf{E}(\tilde{\beta}_1) = \beta_1 + \beta_2 \tilde{\delta}_1$$

onde $\tilde{\delta}_1$ é o coeficiente de inclinação da regressão de x_2 sobre x_1:

$$\tilde{x}_2 = \tilde{\delta}_0 + \tilde{\delta}_1 x_1$$

- $\beta_2\tilde{\delta}_1$ é o viés em $\tilde{\beta}_1$

Resumo do viés

	$\mathbf{Corr}(x_1,x_2) > 0$	$\mathbf{Corr}(x_1,x_2) < 0$
$\beta_2 > 0$	viés positivo	viés negativo
$\beta_2 < 0$	viés negativo	viés positivo

Exemplo (Equação do salário horário)

Suponha que o modelo $y = \beta_0 + \beta_1\ educ + \beta_2\ aptid + u$ satisfaça RLM.1 a RLM.4.

O conjunto de dados no arquivo WAGE1.RAW não contém dados sobre aptidão, de modo que estimamos β_1, a partir da regressão simples:

$$\widetilde{\log(salario)} = 0,584 + 0,083\ educ$$

$$n = 526,\ R^2 = 0,186$$

- Esse é somente o resultado de uma única amostra, de modo que não podemos dizer que 0,083 é maior que β_1
- O retorno verdadeiro da educação poderia ser menor ou maior que 8,3% (nunca saberemos com certeza).
- Entretanto, sabemos que a média dos estimadores de todas as amostras aleatórias deve apresentar viés para cima.

Exemplo

Suponha que, no nível fundamental do ensino, a nota média dos estudantes de um exame padronizado seja determinado por

$$notmed = \beta_0 + \beta_1\ gasto + \beta_2 taxpob + u$$

- $gasto$ é o gasto público por estudante
- $taxpob$ é a taxa de pobreza das crianças da escola

- Usando dados do distrito da escola, temos somente observações da percentagem de estudantes com uma nota de aprovação e gastos públicos por estudante
- Não temos informações sobre taxas de pobreza
- Assim, estimamos β_1, a partir da regressão simples de $notmed$ sobre $gasto$

O que podemos inferir sobre o provável viés de $\tilde{\beta}_1$?

Podemos obter o viés provável em $\tilde{\beta}_1$:

1. β_2 é provavelmente negativo: há evidência de que crianças que vivem na pobreza têm, em média, notas mais baixas em testes padronizados
2. O gasto público médio por estudante é, provavelmente, negativamente correlacionado com a taxa de pobreza: quanto maior a taxa de pobreza menor o gasto público médio por estudante:

$$\Rightarrow Corr(x_1, x_2) < 0$$

$\Rightarrow \tilde{\beta}_1$ terá viés positivo

Implicações:

- Pode ser que o efeito verdadeiro do gasto público fosse zero ($\beta_1 = 0$)
- Entretanto, a estimativa de β_1, da regressão simples será, geralmente, maior que zero
- Isso poderia nos levar a concluir que os gastos públicos são importantes quando eles não são