Variáveis Indicadoras

Roteiro

Introdução

Variáveis Binárias

- Modelo estendido para situações em que os parâmetros da regressão são diferentes para algumas das observações de uma amostra.
- Variáveis Binárias (*Dummy Variable*):

  Variáveis explicativas que podem tomar um de dois valores (em geral, 0 ou 1)
- Representam características qualitativas, em eventos que tenham apenas 2 resultados possíveis.

Variável Binária

- Variável Binária (ou Dicotômica):

  Assume os valores:
  - √ 1, se a característica de interesse está presente
  - √ 0, se a característica de interesse não está presente
- As propriedades dos *EMQO* não são afetadas pela presença de variável explicativa binária
  - √ Podem-se construir estimativas intervalares ou testes de significância para seus coeficientes

Variável Binária de Intercepto

Variáveis Binárias de Intercepto

- Permitem a construção de modelos em que alguns (ou todos) os parâmetros da regressão (inclusive o intercepto) variam para algumas observações da amostra

Exemplo – Economia de Imóveis

- Objetivo: Predizer o valor de mercado de uma casa
- Variável-resposta: valor de mercado do imóvel
- Modelo hedônico: o preço é explicado pelo tamanho do imóvel, pela localização, pelo número de quartos, etc.

Um Primeiro Modelo

- O tamanho da casa (S) é a única variável relevante na determinação de seu preço.

$$P_i = b_0 + b_1 S_i + e_i$$

√ P: preço de mercado da casa
√ S: área útil da casa (m^2)
√ $ß_1$: valor de 1 m^2 adicional de área útil;
√ $ß_0$: valor do terreno

Modelo Estendido

- Modelo do preço da casa agregando a localização:

$$P_i = b_0 + b_1 S_i + dD_i + e_i$$

ou seja:
$$E(P_i) = \begin{cases} (b_0 + d) + b_1 S_i & , D_i = 1 \\ b_0 + b_1 S_i & , D_i = 0 \end{cases}$$

√ Se a vizinhança é desejável: $\beta_0 + d$

√ Em outras áreas: β_0

Vizinhança Desejável

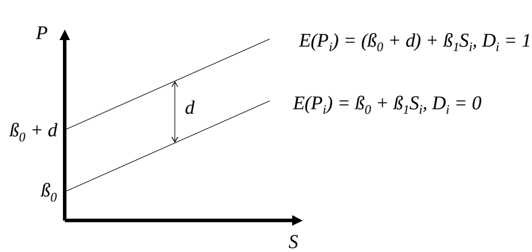
- Seja a variável D_i que representa vizinhança desejável no *i-ésimo* imóvel (universidade, equipamentos urbanos, etc.)

  Assume os valores:

  √ 1, se a propriedade está em uma vizinhança desejável

  √ 0, se a propriedade não está em uma vizinhança desejável

- Supondo $d > 0$:



Interpretação

- d: diferença no preço da casa devido estar localizada em vizinhança desejável

  (*prêmio de localização*)
- Se $d = 0$, não há prêmio de localização para a vizinhança

Variável de Interação

Variáveis de Inclinação

- Se o efeito da localização causar uma variação no coeficiente angular, ou seja, o valor do m^2 é diferente em cada uma das localizações:

$$P_i = b_0 + b_1 S_i + g\,(S_i D_i) + e_i$$

- $S_i D_i$: variável de interação (variável binária de inclinação):

  Capta o efeito de interação da localização e do tamanho da casa

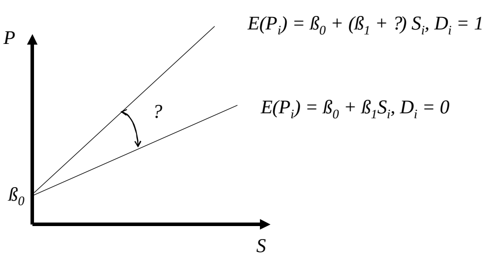
- ou seja:

$$E(P_i) = \begin{cases} b_0 + (b_1 + g)S_i & , D_i = 1 \\ b_0 + b_1 S_i & , D_i = 0 \end{cases}$$

√ Preço do m^2 em local com vizinhança desejável: *ß*$_1$ *+ ?*

√ Preço do m^2 em outras localizações: *ß*$_1$

- Supondo ? > *0*:



Coeficientes de Variáveis Binárias – Inferência

- Se os pressupostos dos modelo estiverem corretos, os *EMQO* têm suas propriedades usuais.
- Além de estimação intervalar pode-se efetuar teste de hipóteses:

√ H_o: ? = 0 vs H_1: ? ? 0

√ H_o: ? = 0 vs H_1: ? > 0

Variável de Intercepto e de Interação

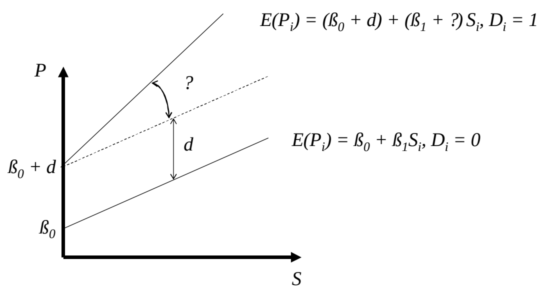
- Se a localização afetar tanto o intercepto quanto o coeficiente angular, então:

$$P_i = b_0 + b_1 S_i + dD_i + g(S_i D_i) + e_i$$

ou seja:

$$E(P_i) = \begin{cases} (b_0 + d) + (b_1 + g) S_i & , D_i = 1 \\ b_0 + b_1 S_i & , D_i = 0 \end{cases}$$

- Supondo *d* e *? > 0*:



Interação entre Fatores Qualitativos

- Situações verificadas:
  - √ Variáveis binárias de intercepto aditivas;
  - √ Efeito das variáveis binárias independentes de qualquer fator qualitativo
- E quando os fatores qualitativos <u>não forem independentes</u>?

Exemplo

- Estimação da equação de regressão de salário, explicado por: experiências, habilidades e outros fatores referentes à produtividade
- Costuma-se incluir as variáveis raça e sexo
  - raça: 1, se branco; 0, caso contrário
  - Sexo: 1, se homem, 0, caso contrário

√ Se a determinação do salário não é discriminatória, então seus coeficientes não serão significativos

---

- A inclusão apenas das variáveis sexo e raça não captará a interação entre estes fatores

√ Ex.: tratamento especial de salário por ser homem **e** branco

- Modelo:

$$Salário_i = b_0 + b_1 experiência_i + d_1 raça_i + d_2 sexo_i + g\,(raça_i \times sexo_i) + e_i$$

$$E(salário_i) = \begin{cases} (b_0 + d_1 + d_2 + g) + b_1 experiência_i & , branco - homem \\ (b_0 + d_1) + b_1 experiência_i & , branco - mulher \\ (b_0 + d_2) + b_1 experiência_i & , não\ branco - homem \\ b_0 + b_1 experiência_i & , não\ branco - mulher \end{cases}$$

d_1 mede o efeito raça; d_2, o efeito de sexo e *?*, o efeito de ser branco e homem

---

Aplicação

Exemplo – Imóveis

- Dados sobre duas vizinhanças (próxima a uma grande universidade e a 3 km de distância)
  - √ Preço das casas (*$*)
  - √ Área: tamanho da área útil (m^2)
  - √ Local: 1, para casas próximas da universidade, 0 caso contrário
  - √ Piscina: 1, se há piscina, 0 caso contrário
  - √ Lareira: 1, se tem lareira, 0 caso contrário
- Dados: *imoveis*

Regressão

- Especifica-se a equação de regressão como:

$$Preço_i = b_0 + b_1 area_i + b_2 idade_i + d_1 local_i + d_2 piscina_i + \\ + d_3 lareira_i + g\,(area_i \times local_i) + e_i$$

  - √ Todos os coeficientes serão positivos, exceto $ß_2$ (depreciação sobre o preço da casa);
  - √ Variável binária de inclinação da interação área x local.

Regression Analysis: preço versus area; idade; ...

```
The regression equation is
preço = 24500 + 76,1 area - 190 idade + 27453 local + 4377 piscina
        + 1649 lareira + 13,0 area*local

Predictor       Coef  SE Coef      T      P
Constant       24500     6192   3,96  0,000
area          76,122    2,452  31,05  0,000
idade        -190,09    51,20  -3,71  0,000
local          27453     8423   3,26  0,001
piscina         4377     1197   3,66  0,000
lareira       1649,2    972,0   1,70  0,090
area*local    12,994    3,320   3,91  0,000

S = 15225,2   R-Sq = 87,1%   R-Sq(adj) = 87,0%

Analysis of Variance

Source           DF           SS           MS        F      P
Regression        6  1,54826E+12  2,58044E+11  1113,18  0,000
Residual Error  993  2,30184E+11    231807076
Total           999  1,77845E+12


Source      DF       Seq SS
area         1  6,28934E+11
idade        1   7218668650
local        1  9,05110E+11
piscina      1   2966900797
lareira      1    482209720
area*local   1   3549887558
```

Resultados do Ajuste

- O modelo se ajusta bem aos dados;
- Com base em testes de significância unilateral, no nível de 5%, rejeitamos a hipótese de que qualquer dos parâmetros seja zero
- Aceitamos a alternativa de que são positivos, exceto o coeficiente idade

Regressão Estimada

- Equação estimada do modelo:

```
Preço = 24.500 + 76,1 area - 190 idade + 27.453 local + 4.377 piscina
  + 1.649 lareira + 13,0 area*local
```

- Casas próximas à Universidade (*local=1*)

```
Preço = (24.500 + 27.453) + (76,1 + 13,0) area - 190 idade +
        4.377 piscina + 1.649 lareira
Preço = 51.953 + 89,1 area - 190 idade + 4.377 piscina + 1.649 lareira
```

- Casas distantes da Universidade (*local=0*)

```
Preço = 24.500 + 76,1 area - 190 idade + 4.377 piscina + 1.649 lareira
```

Conclusões

- Prêmio de localização estimado, para lotes próximos à universidade: *$27.453*
- Preço por m^2:
  - √ Próximo à universidade: *$89,11*
  - √ Distantes da universidade: *$76,12*
- Depreciação: *$190,09* por ano
- Aumento do preço devido à piscina: *$4.377,16*
- Aumento do preço devido à lareira: *$1.649,17*

Variáveis Qualitativas com várias Categorias

Variáveis Qualitativas Não-Binárias

- Muitos fatores têm mais de duas categorias:
  - √ Regiões de um país: sul, sudeste, centro-oeste, nordeste e norte
  - √ Nível de instrução: menos que médio, médio, superior, pós-graduação

Cuidados na Construção

- Exemplo: Salário explicado pela experiência e nível de instrução
- Variáveis binárias para nível de instrução:
  - √ E_0: 1, menos que ensino médio; 0, caso contrário
  - √ E_1: 1, nível médio; 0, caso contrário
  - √ E_2: 1, nível universitário; 0, caso contrário
  - √ E_3: 1, pós-graduado; 0, caso contrário

- Modelo especificado:

$$Salário_i = b_0 + b_1 experiência_i + d_1 E_{i1} + d_2 E_{i2} + d_3 E_{i3} + e_i$$

$$E(salário_i) = \begin{cases} (b_0 + d_3) + b_1 experiência_i & , \text{pós} - \text{graduado} \\ (b_0 + d_2) + b_1 experiência_i & , \text{nível universitário} \\ (b_0 + d_1) + b_1 experiência_i & , \text{nível médio} \\ b_0 + b_1 experiência_i & , \text{menos que médio} \end{cases}$$

- A inclusão de todas as variáveis criaria colineariadade exata, já que:

  $E_0 + E_1 + E_2 + E_3 = 1$

- Solução: omitir uma variável binária (**grupo de referência)**
- $ß_0$: representa salário-base para trabalhador sem qualquer experiência e sem diploma de ensino médio.
- Não importa qual variável seja omitida, embora haja escolha mais conveniente
- A não-omissão levará à impossibilidade de ajuste

Resumo

Variáveis *Dummy*

- Variáveis Binárias Qualitativas, usadas para indicar a presença ou ausência de determinado fenômeno

  Assumem apenas o valor 0 ou 1

  Exemplo

  Existência ou não de piscinas numa regressão do preço de casas
  - √ $X_i = 1$, se a casa tem piscina
  - √ $X_i = 0$, se a casa não tem

Tipos de Variáveis *Dummy*

- Aditiva: altera o intercepto
- Multiplicativa: altera o coeficiente angular
- Mista: altera o intercepto e o coeficiente

- Podem ser usadas também para avaliar qualitativamente situações com mais de 2 alternativas possíveis

  Exemplo

  A qualidade da condição do piso da casa boa, média ou ruim
  - √ Usam-se $p - 1$ variáveis, sendo p o número de possibilidades

$$X_i = \begin{cases} 1, \text{ se o piso está em boas condições} \\ 0, \text{ se não} \end{cases}$$

$$X_{i+1} = \begin{cases} 1, \text{ se o piso está em condições médias} \\ 0, \text{ se não} \end{cases}$$

- Deixa-se de fora a possibilidade de as condições serem ruins. Esta ocorre quando $X_i = 0$ e $X_{i+1} = 0$

  Ou seja, o piso está em condições ruins quando não está em boas condições ($X_i = 0$) nem em condições médias ($X_{i+1} = 0$)

- O método dos mínimos quadrados não tem respostas se informam-se *p* variáveis (no exemplo, 3) ao invés de $(p - 1)$ variáveis
  - √ É inadequado informar-se apenas uma variável, com os valores *1* (boa), *2* (média) e *3* (ruim).
  - √ Neste caso, se entenderia que a condição *3* (ruim) é *3* vezes tão ruim quanto a condição boa (*1*)

Referências

Bibliografia Recomendada

- Hill, R. C., Griffiths, W. E. e Judge, G. (Saraiva)
  Econometria
- Gujarati, D. N. (Pearson)
  Econometria Básica
- Maddala, G. S. (LTC)
  Introdução à econometria