Variáveis Indicadoras

Variáveis preditoras qualitativas (variáveis *dummy*)

- Exemplos de variáveis explanatórias qualitativas: compras (sim; não), sexo (masculino e feminino), tipo de firma (valores, ações, capital e comercial), regiões (nordeste, centro e sul), estação do ano (verão, outono, inverno e primavera).
- Um economista deseja relacionar a velocidade com que um novo seguro é adotado *(Y)* com o tamanho da firma (X_1) e o tipo de firma. A variável resposta é medida em número de meses passados entre o tempo que a primeira firma adotou a inovação e o tempo que uma dada firma adotou. A variável X_1 é dada em milhões de dólares. A segunda variável preditora é qualitativa e é dada em duas classes. Para que a variável qualitativa possa ser usada no modelo, deve-se usar indicadores quantitativos(variáveis indicadoras) para as classes da mesma.

Exemplo

- Para o exemplo da inovação de um seguro, onde a variável qualitativa tem duas classes, podemos definir duas variáveis indicadoras, X_2 e X_3 do seguinte modo:

$$X_2 = \begin{cases} 1 & \textit{firma de capital} \\ 0 & \textit{outros casos} \end{cases}$$

$$X_3 = \begin{cases} 1 & \textit{firma de comércio} \\ 0 & \textit{outros casos} \end{cases}$$

- Para o exemplo, pensaríamos em usar um modelo de primeira ordem, dado por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \quad (1)$$

Exemplo

- Esta abordagem intuitiva de designar variáveis indicadoras para cada classe da variável qualitativa, infelizmente, nos traz grandes dificuldades computacionais. Por exemplo, suponha que temos *n=4* observações, as primeiras duas sendo para firmas de capital (X_2*=1* e X_3*=0),* e as duas últimas sendo para firmas de comércio (X_2*=0* e X_3*=1).* A matriz de delineamento **X**, fica:

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & 1 & 0 \\ 1 & X_{21} & 1 & 0 \\ 1 & X_{31} & 0 & 1 \\ 1 & X_{41} & 0 & 1 \end{bmatrix}$$

- Note que a primeira coluna é igual a soma da terceira com a quarta. Portanto, as colunas são linearmente dependentes. Isto tem um efeito sério sobre a matriz $\mathbf{X^TX}$**:**

Exemplo

- Uma maneira simples de resolver este problema é retirar uma das variáveis indicadoras. No exemplo, podemos retirar a variável X_3.
- Uma variável qualitativa com *c* classes será representada por *c-1* variáveis indicadoras, cada uma delas recebendo os valores 0 e 1.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \quad (2)$$

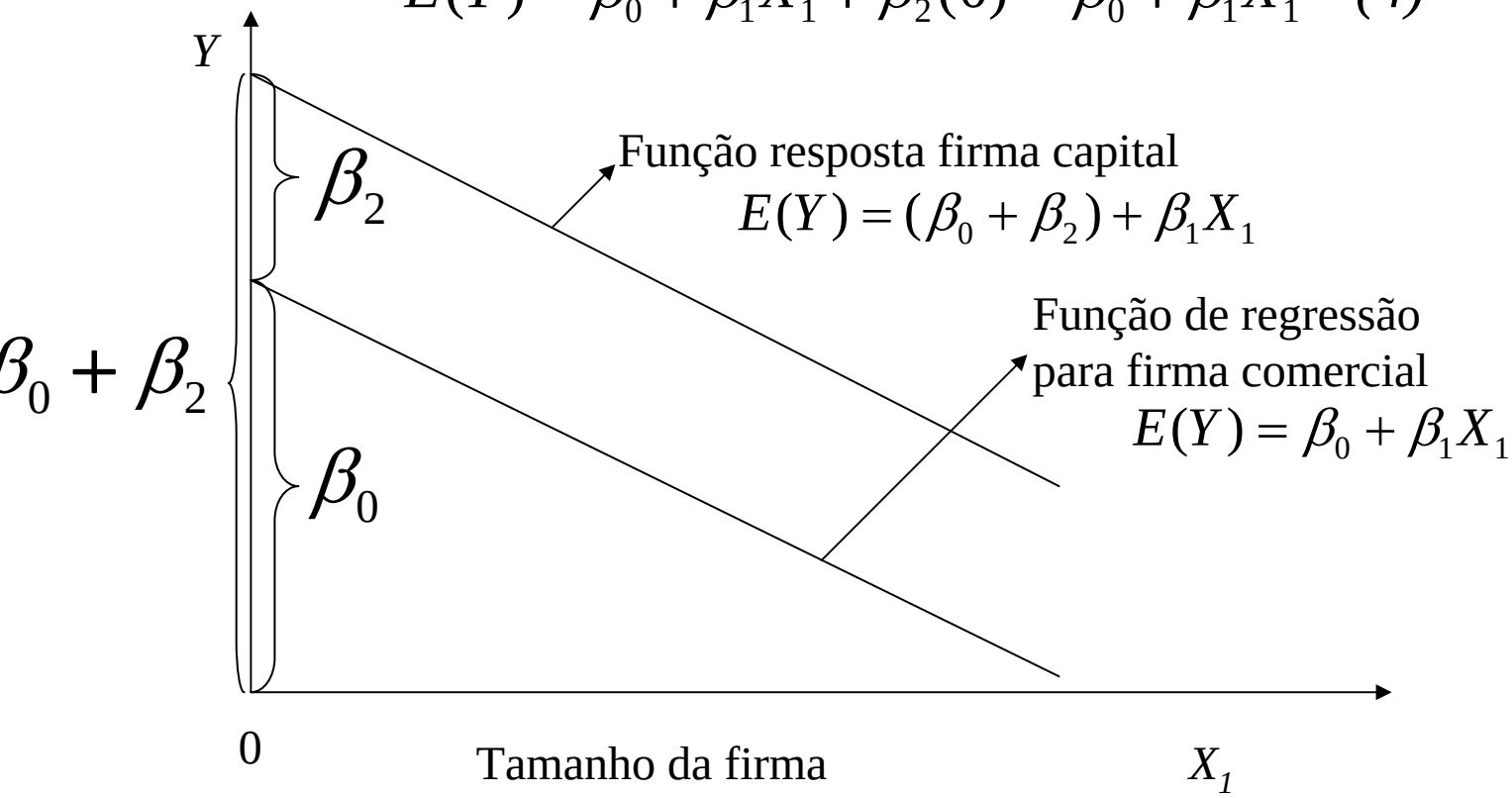
Onde: X_{i1}= *tamanho da firma* e X_{i2}=*1 se for firma de capital e* X_{i2}= *0 em outros casos.*

$$E(Y / \boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \qquad \boldsymbol{(3)}$$

Exemplo

- Para entender o significado dos coeficientes de regressão neste modelo, considere, primeiramente, o caso da firma comercial. Para esta firma $X_2=0$ e a função de resposta fica:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2(0) = \beta_0 + \beta_1 X_1 \quad (4)$$



Exemplo

- Para a firma de capital, $X_2=1$ a função de resposta (3) é dada por:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2(1) = (\beta_0 + \beta_2) + \beta_1 X_1 \quad (5)$$

- Também temos a equação de uma reta, com mesmo coeficiente angular, β_1, mas com intercepto Y dado por $(\beta_0+\beta_2)$. Esta função de resposta também está indicada na figura anterior.
- Na equação (3), o tempo médio passado antes da inovação ser adotada, $E(Y)$, é uma função linear do tamanho da firma (X_1), com o mesmo coeficiente angular, β_1, para ambas as firmas. O parâmetro β_2 indica quanto maior (ou menor) é a função de resposta para a firma de capitais do que a firma comercial, para qualquer tamanho da firma. Portanto, β_2 é um diferencial do efeito do tipo de firma.

Comentários

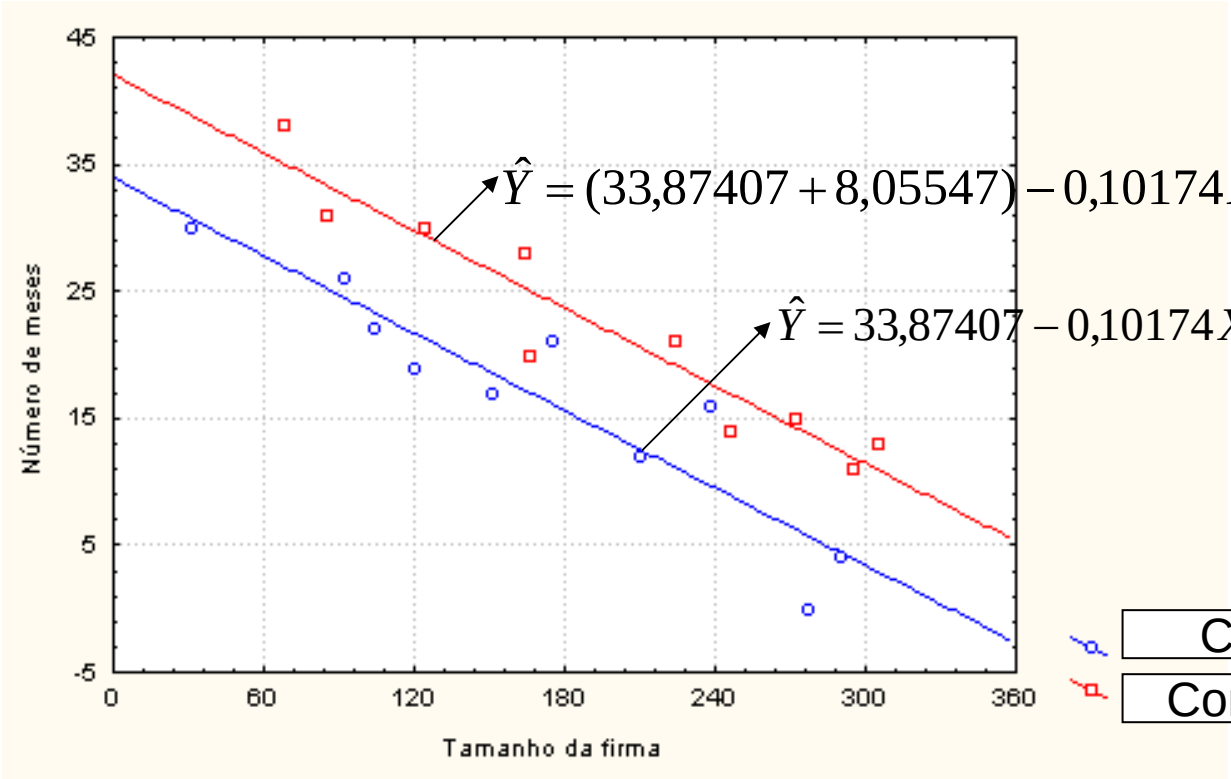
- Visto que existem duas regressões, poderíamos fazer duas regressões separadamente.
- Contudo uma abordagem única é preferível porque o analista tem somente uma equação final para trabalhar. Além disso, faz sentido quando é assumido que as linhas tem a mesma inclinação.
- Essa abordagem também apresenta uma estimativa única para a variância dos erros σ^2 e permite mais variabilidade usando mais graus de liberdade.

Um outro modelo para o exemplo: diferentes intercepto e inclinação

Dados do exemplo de seguros					
	(1)	(2)	(3)	(4)	(5)
Firma	Número de meses	Tamanho da firma	Tipo de firma	Variável codificada	
i	Y_i	X_{i1}		X_{i2}	$X_{i1}X_{i2}$
1	17	151	Comercial	0	0
2	26	92	Comercial	0	0
3	21	175	Comercial	0	0
4	30	31	Comercial	0	0
5	22	104	Comercial	0	0
6	0	277	Comercial	0	0
7	12	210	Comercial	0	0
8	19	120	Comercial	0	0
9	4	290	Comercial	0	0
10	16	238	Comercial	0	0
11	28	164	Capitais	1	164
12	15	272	Capitais	1	272
13	11	295	Capitais	1	295
14	38	68	Capitais	1	68
15	31	85	Capitais	1	85
16	21	224	Capitais	1	224
17	20	166	Capitais	1	166
18	13	305	Capitais	1	305
19	30	124	Capitais	1	124
20	14	246	Capitais	1	246

Resultados do exemplo

- O modelo $$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \quad (6)$$
- Reta estimada única $$\hat{Y} = 33{,}8741 - 0{,}1017 X_1 + 8{,}0555 X_2 \quad (7)$$



A figura contém a função de resposta ajustada para cada tipo de firma, juntamente com os valores observados

Resultados do exemplo

- O economista está mais interessado no tipo de firma (X_2) sobre o tempo necessário para a inovação ser adotada e, assim, deseja construir um intervalo de confiança para β_2. Com o auxílio de um programa obtemos o valor de *t*=2,109815 com 17 graus de liberdade e α *=0,05.* Usando os resultados da saída do programa, o intervalo de confiança é dado por:

$$8{,}05547 \pm 2{,}110(1{,}45911)$$

$$4{,}98 \leq \beta_2 \leq 11{,}13$$

- Portanto, temos 95% de confiança que mudando de firma comercial para firma capital cresce a velocidade média de inovação entre 4,98 e 11,13 meses.

Modelo contendo o efeito da interação

- Considere o modelo com intercepto e inclinação diferentes

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i \qquad (9)$$

- Onde

$$X_{i1} = tamanho \quad da \quad firma$$

$$X_{i2} = \begin{cases} 1 & para \quad firma \quad de \quad capitais \\ 0 & outros \quad casos \end{cases}$$

- Equação estimada

$$E(Y_i / \mathbf{X}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} \qquad (10)$$

Interação

(produto cruzado)

Modelo contendo o efeito da interação

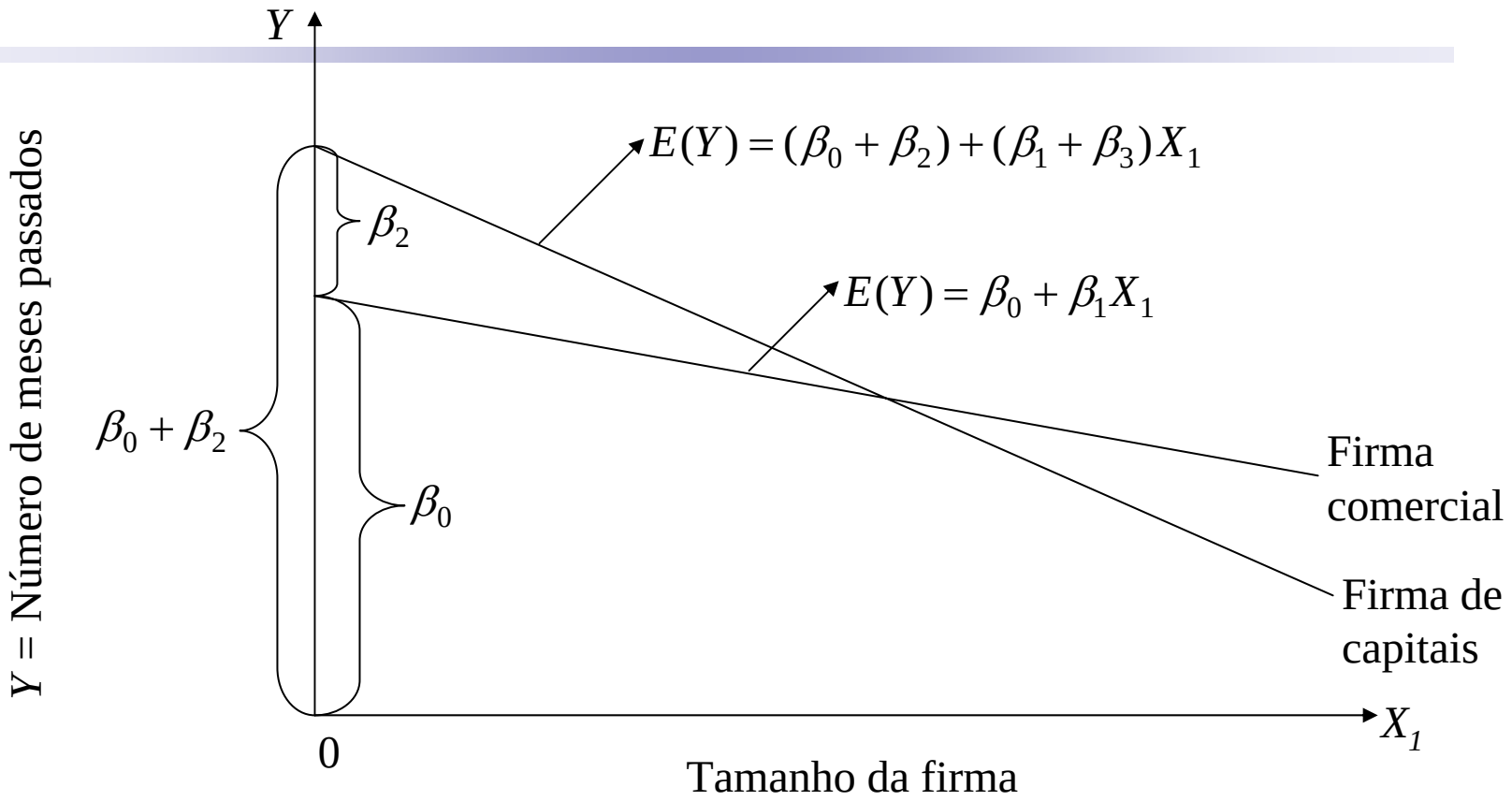
- Significado dos parâmetros
  - *Para firma de comércio:* $X_2=0$ e assim, $X_1X_2=0$, assim, a função de resposta (10) fica:

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2(0) + \beta_3(0) = \beta_0 + \beta_1 X_1 \quad (11)$$

  - *Para firma de capitais:* Para firma de capitais temos $X_2=1$ e assim, $X_1X_2=X_1$, assim, a função de resposta (10) fica:

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2(1) + \beta_3 X_{i1} \quad (12)$$

$$E(Y_i) = \underbrace{(\beta_0 + \beta_2)}_{\text{Coef. linear}} + \underbrace{(\beta_1 + \beta_3)}_{\text{Coef. angular}} X_{i1} \quad (13)$$



Nós vimos que β_2 indica o quanto é maior (ou menor) o intercepto da função de resposta para a classe com o código 1 (firma de capitais) do que a classe com o código 0 (comercial). Da mesma forma, β_3 indica quanto é maior (ou menor) o coeficiente angular da função resposta para a classe com código 1 do que a classe com código 0. Como tanto o intercepto como o coeficiente angular são diferentes para as duas classes no modelo de regressão (9), não podemos fazer a afirmação acima para β_2 para qualquer nível de X_1.

Modelo contendo o efeito da interação

- Pela figura observamos que o efeito do tipo de firma, no modelo (9), depende do tamanho da mesma. Para firmas pequenas, as companhias de comércio adotam a inovação mais rapidamente, porém, para firmas maiores, as companhias de capitais adotam a inovação antes do que as de comércio. É o efeito da interação.

- Teste de significância (dois modelos são idênticos)

$$H_0 : \beta_2 = \beta_3 = 0$$

$$H_a : \textit{pelo menos um é diferente de zero}$$

  - Se H_0 não é rejeitada, isto implica que um modelo único pode explicar o relacionamento entre a velocidade de inovação e o tamanho da firma.

- Teste de significância da interação

$$H_0 : \beta_3 = 0$$

$$H_a : \beta_3 \neq 0$$

Testando a significância de dois modelos

- Hipóteses:

$$H_0 : \beta_2 = \beta_3 = 0$$

$$H_a : \textit{pelo menos um é diferente de zero}$$

- Estatística do teste: soma extra de quadrados
- Para o exemplo

$$F^* = \frac{SS_M(X_2, X_1X_2 | X_1)/2}{MS_R}$$

$$SS_M(X_2, X_1X_2 \mid X_1) = SS_M(X_2 \mid X_1) + SS_M(X_1X_2 \mid X_1, X_2)$$
$$= 316.245973 + 0.005708$$
$$= 316{,}251681$$

$$F^* = \frac{316{,}251681}{2} \div \frac{176{,}38096}{16} = \frac{158{,}1258405}{11{,}02381} = 14{,}3440$$

- P(F>14,3440)=0,000270. Portanto, rejeita-se a hipótese nula.

Modelos mais complexos

- Exemplo: vamos considerar a regressão da durabilidade de uma ferramenta (*Y*), sobre a velocidade (X_1) e o modelo da ferramenta, onde, esta é uma variável qualitativa com 4 classes (M1, M2, M3, M4). Para trabalhar com esta variável precisamos definir as seguintes variáveis indicadoras:

$$X_2 = \begin{cases} 1 & para\ \ o\ modelo\ \ 1 \\ 0 & outros\ \ casos \end{cases}$$

$$X_3 = \begin{cases} 1 & para\ \ o\ modelo\ \ 2 \\ 0 & outros\ \ casos \end{cases}$$

$$X_4 = \begin{cases} 1 & para\ \ o\ modelo\ \ 3 \\ 0 & outros\ \ casos \end{cases}$$

Comparação entre modelos

- Modelo completo: Considere um modelo de regressão linear simples com m níveis. O modelo completo é dado por m regressões computadas separadamente. A soma de quadrados de resíduos é a soma de somas de quadrados de resíduos dos modelos.

$$y = \beta_{0m} + \beta_{1m} x + \varepsilon \qquad m = 1,2,...,M$$

- É interessante comparar o modelo completo com modelos reduzidos:

$$F^* = \frac{SS_R(MR) - SS_R(MC)/(df_{MR} - df_{MC})}{SS_R(MC)/df_{MC}}$$

$$SS_R(MC) = \sum_{i=1}^{M} SS_R(m) \qquad df_{MC} = \sum_{i=1}^{M} n_m - 2 = n - 2M$$

  - Se F^* é pequeno, o modelo reduzido é satisfatório. F^* será menor que $F_{\alpha,dfRM-dfMC,dfMC}$.

Modelos reduzidos

- Modelo em que todos as M inclinações são iguais porém os interceptos são diferentes. Seja D uma variável indicadora de níveis

$$y = \beta_{0m} + \beta_{1m} x + \beta_2 D_1 + \ldots + \beta_m D_{m-1} + \varepsilon$$

- Modelo em que todos as M inclinações são diferentes porém os interceptos são iguais.

$$y = \beta_0 + \beta_1 x + \beta_2 Z_1 + \ldots + \beta_m Z_{m-1} + \varepsilon$$

$$Z_k = xD_k$$

- Modelo em que todas as M inclinações e os M interceptos são iguais.

$$y = \beta_0 + \beta_1 x + \varepsilon$$