Análise de Regressão Linear Múltipla: Variáveis Dummy

Patrícia Iwagaki Braga Ogando

28 de novembro de 2016

Sumário

1 Introdução:

Na análise de regressão, a variável dependente pode ser influenciada por variáveis qualitativas ou quantitativas.
Diferente das variáveis quantitativas, as qualitativas não são facilmente mensuradas e são caracterizadas por indicar presença ou ausência de uma qualidade ou atributo.
Essas variáveis são chamadas de *dummy*.
Um modo de quantificar esses atributos é a construção de variáveis artificiais, assumindo valores binários, ou seja 0 e 1, o que indicará respectivamente a ausência e a presença de atributo.
Porém as variáveis *dummies* não precisam necessariamente assumir os valores 0 e 1, pode ser transformada utilizando a função linear:

$$Z = a + bD \text{ , } b \neq 0 \text{ e } a, b \in \mathbb{R}$$

Atribuindo os valores 0 ou 1 para D temos:

$$D = 0,\ Z = a + b$$
$$D = 1,\ Z = a$$

Ao introduzir as variáveis *dummies* torna o modelo de regressão extremamente útil para estudos empíricos.

2 Modelo de regressão:

Seja um modelo com uma ou mais variáveis quantitativas que seja estável para todas as observações de uma dada amostra:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \text{ com } i = 1, 2, ..., n$$

Onde x_i é a variável quantitativa e $\epsilon_i \sim N(0, \sigma^2)$

Ao analisar esse modelo verificamos que apesar do x constituir uma variável importante no comportamento de Y, existe uma parcela do comportamento que não é explicado no modelo. Suponha que existam três grupos em que cada grupo possua ausência ou presença de um atributo.

Figura 3: Gráfico de dispersão com obeservações hipotéticas
Fonte: Valle e Rabelo (2002)

Analisando o gráfico anterior percebemos uma relação linear positiva, porém parece que há diferentes relações para cada grupo. Se utilizarmos apenas a variável x como independente, estaremos omitindo informações conhecidas. O mais correto nesse caso é ajustar um modelo de regressão para cada grupo. De acordo com o gráfico de dispersão os interseptos β_0 diferem, enquanto as inclinações das retas parecem ser os mesmos.

$$1^o \text{ grupo: } Y_i = \alpha_i + \beta X_i + \epsilon_i$$
$$2^o \text{ grupo: } Y_i = \alpha_2 + \beta X_i + \epsilon_i$$
$$3^o \text{ grupo: } Y_i = \alpha_3 + \beta X_i + \epsilon_i$$

Contudo a estimação dos três diferentes modelos provavelmente não terá o mesmo parâmetro β. Dessa forma a definição de regressores dummy apresenta-se como um procedimento adequado.

$$D_{2i} \begin{cases} 1, \textit{ se a observação verifica a característica que define o } 2^o \textit{ grupo} \\ 0 \end{cases}$$

$$D_{3i} \begin{cases} 1 \textit{ se a observação verifica a característica que define o } 3^o \textit{ grupo} \\ 0 \end{cases}$$

Onde D_{2i} é a diferença entre os termos independentes dos dois primeiros grupos e D_{3i} é adiferença entre os termos independentes do primeiro e do terceiro grupo.

Portanto com a introdução de regressores dummy pode-se ajustar o seguinte modelo:

$$Y_i = \alpha_1 + \gamma_2 D_{2i} + \gamma_3 D_{3i} + \epsilon_i \ , \ i = 1, ...n$$

$\epsilon_i \sim N(0, \sigma^2)$

onde $\gamma_2 = (\alpha_2 - \alpha_1)$ e $\gamma_3 = (\alpha_3 - \alpha_2)$

E portanto é possível obter uma única estimativa para o parâmetro β e simultaneamente três ordenadas, na origem distintas:

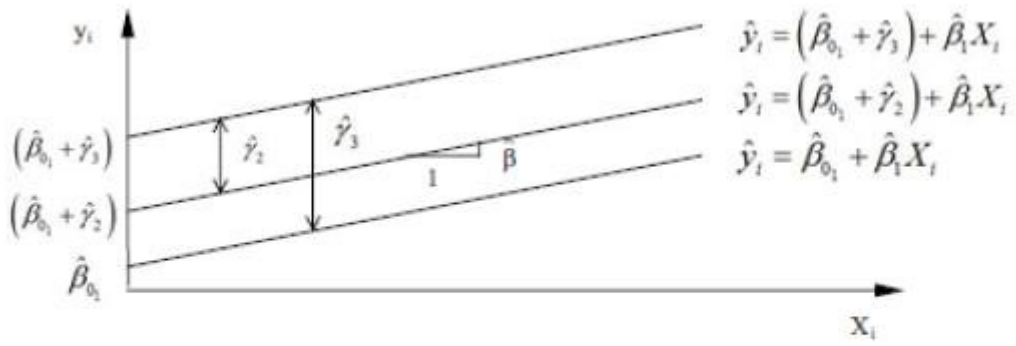


Figura 4: Gráfico da estrutura estimada do modelo 2.20
Fonte: Valle e Rabelo (2002)

Podemos separar os grupos da seguinte forma:

1º grupo: $Y_i = \alpha_1 + \beta X_i + \epsilon_i$ onde $D_{2i} = D_{3i} = 0$
2º grupo: $Y_i = (\alpha_1 + \alpha_2) + \beta X_i + \epsilon_i$ onde $D_{2i} = 1 \quad e \quad D_{3i} = 0$ 3º grupo:
$Y_i = (\alpha_1 + \alpha_3)\beta X_i + \epsilon_i$ onde $D_{2i} = 0 \quad e \quad D_{3i} = 1$

Nesse caso admitiu-se que o coeficiente de inclinação é semelhante a todos os modelos. O efeito de cada fator qualitativo é somado ao intercepto. Pode ocorrer das retas de regressão terem o mesmo intersepto com coeficientes de inclinação distinto. Podemos exemplificar de tal forma:

$$\begin{aligned} &1^{o} \text{ grupo: } Y_i = \alpha + \beta_1 X_i + \epsilon_i \\ &2^{o} \text{ grupo: } Y_i = \alpha + \beta_2 X_i + \epsilon_i \\ &3^{o} \text{ grupo: } Y_i = \alpha + \beta_3 X_i + \epsilon_i \end{aligned}$$

O objetivo nesse caso é encontrar um único modelo de tal forma que podemos produzir uma estimativa para o termo independente e três coeficientes de inclinação distintos.

Tomando:

$$D_{2i} \begin{cases} 1, \textit{ se a observação verifica a característica que define o 2º grupo} \\ 0 \end{cases}$$

$$D_{3i} \begin{cases} 1 \textit{ se a observação verifica a característica que define o 3º grupo} \\ 0 \end{cases}$$

Portanto o modelo que deve ser estimado é:

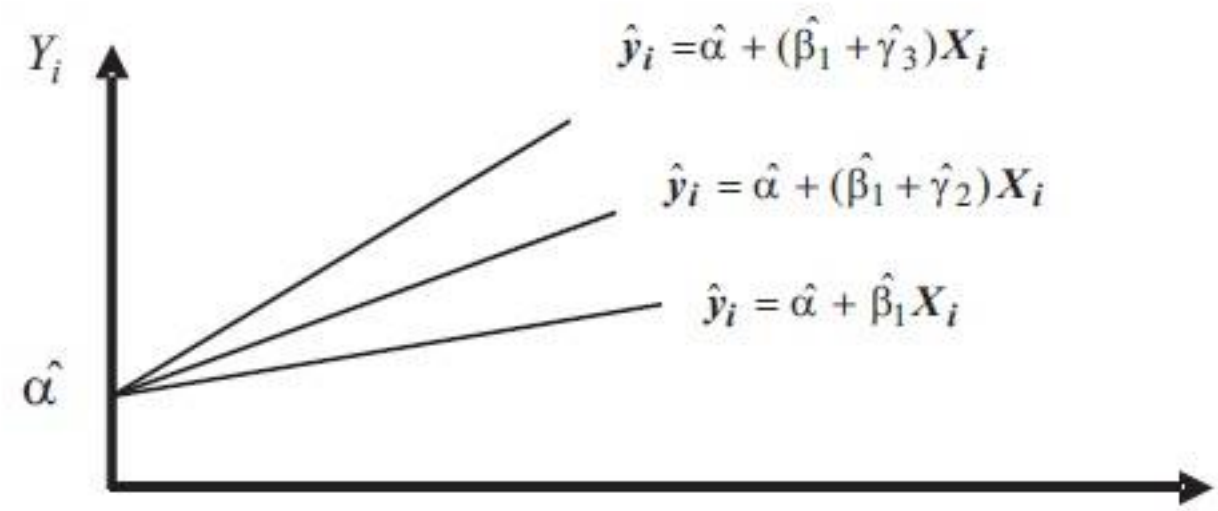
$$Y_i = \alpha + \beta_1 X_i + \gamma_2 (D_{2i} X_i) + \gamma_3 (D_{3i} X_i) + \epsilon_i \text{ com } i = 1, ..., n \text{ e } \epsilon_i \sim N(0, \sigma^2)$$

onde $\gamma_2 = (\beta_2 - \beta_1)$ e $\gamma_3 = (\beta_3 - \beta_1)$

Portanto os três grupos ficam da seguinte forma:

1º grupo: $Y_i = \alpha + \beta_1 X_i + \epsilon_i$ onde $D_{2i} = D_{3i} = 0$
2º grupo: $Y_i = \alpha + \beta_2 X_i + \epsilon_i$ onde $D_{2i} = 1 \quad e \quad D_{3i} = 0$
3º grupo: $Y_i = \alpha + \beta_3 X_i + \epsilon_i$ onde $D_{2i} = 0 \quad e \quad D_{3i} = 1$

Como mostra o gráfico a seguir:



Fonte: REBELO & VALLE (2002)
Figura 3: Estrutura geométrica do modelo (2.14).

3 Alguns exemplos e ANOVA

Vamos admitir que as retas de regressão para os distintos grupos diferem apenas no termo de intercepto, mantendo-se os mesmos coeficientes angulares, conforme pode ser observado anteriormente. Neste caso, a variável dummy é incorporada ao modelo de regressão para captar o efeito do deslocamento do intercepto como resultado de algum fator qualitativo.

Exemplo:

Através do uso de variáveis dummy busca-se identificar se existe diferença entre os salários médios recebidos por professores e professoras universitários. A hipótese implícita deste modelo é de que os professores universitários receberiam um salário maior. Neste caso, mantidos constantes todos os demais fatores, caso a diferença se confirme, pode-se especular sobre a possibilidade de haver discrimina- ção com relação ao salário pago às professoras.

Temos o seguinte modelo de regressão:

$$Y_i = \alpha + \beta D_i + \epsilon_i \text{ onde } \epsilon \sim N(0, \sigma^2) \text{ e } E(\epsilon_i) = 0$$

$$D_i \begin{cases} 1, & \textit{se for do sexo masculino} \\ 0, & \textit{se for do sexo feminino} \end{cases}$$

Testaremos a hipótese que o professor recebe mais que a professora universitária:

Salário médio de uma professora universitária: $E(Y_i/D_i = 0) = \alpha$

Salário médio de um professor universitário: $E(Y_i/D_i = 1) = \alpha + \beta$

O coeficiente de inclinação β informa em quanto o salário médio de um professor universitário difere do salário-médio de uma professora. Caso os resultados obtidos mostrem que β é estatisticamente significativo, conclui-se que, o salário de um professor, de fato, é superior ao de uma professora.

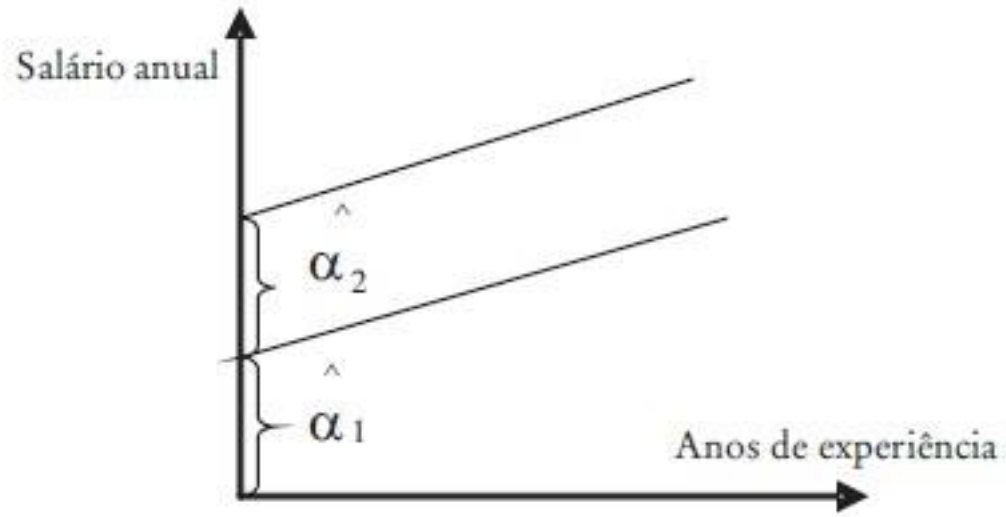


Figura 6: Funções salários em relação aos anos de experiência.

Para estimar os salários das professoras:

$$Y_i = (\alpha_1 + \alpha_2) + \beta X_i + \epsilon_i$$

O vetor das estimativas dos parâmetros neste caso é dado por:
$B = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{bmatrix}$ [] Falta-se avaliar se o valor estimado para o parâmetro $\widehat{b}$ é estatisticamente significativo, para tanto, calcula-se a análise de variância de regressão:

Tabela 1: Análise de Variância.

CV	GL	SQ	QM
Regressão	$K-1$	$\hat{b}' X'y - n\bar{Y}^2$	$\hat{b}' X'y - n\bar{Y}^2 \Big/ K-1$
Resíduo	$n-K$	$y'y - \hat{b}' X'y$	$y'y - \hat{b}' X'y \Big/ n-K$
Total	$n-1$	$y'y - n\bar{Y}^2$	

A partir da Análise de Variância é possível fazer um teste F, para calcular a significância do parâmetro e verificar se os professores universitários recebem mais que as professoras.

4 Aplicação:

A aplicação está em um arquivo separado!

5 Conclusão:

A introdução das variáveis dummy na análise de regressão é um instrumento importante que amplia o poder de análise dos modelos. Isso se deve ao fato de que elas permitem a o uso de variáveis que não podem ser medidas quantitativas. O trabalho teve por objetivo apresentar, situações em que as variáveis dummy podem ser inseridas, no caso em que estas são consideradas variáveis independentes.

6 Referências Bibliográficas:

- Análise de Variância e Análise de Regressão com variáveis Dummy:Mais Semelhanças do que Diferenças. Revista de Estatística, Vol. I, pp. 49-86, 2002.
- Dualidades entre Análise de Covariância e Análise de Regressão com variáveis dummy. Revista de Estatística. 2º quadrimestre de 2002, pp. 65-86.
- http://www.anpad.org.br/admin/pdf/ADI-D768.pdf.
- http://rpubs.com/adriano/analreg01.
- REBELO, E; VALLE, P.O. O uso de regressores dummy na especificação de modelos com parâmetros Variáveis. Revista de Estatística, 3ºquadrimestre de 2002, pp. 17-40.