**Aula 4 – Lab 1**

**Modelo de Regressão em Notação Matricial**

*Módulo 1*

Qualquer sistema de equações linear pode ser traduzido em termos matriciais. Há uma maneira particular de considerar a relação entre equações e matrizes que é relevante para a análise de regressão multivariada. Neste laboratório, mostraremos como a álgebra linear é útil para a obtenção dos parâmetros de um modelo de regressão.

Mostraremos primeiro um exemplo com um par de observações de quaisquer variáveis (*x1*,y1). Um pesquisador pode se interessar em determinar a magnitude da relação entre estas duas variáveis. Se *y = f(x)*, pode-se assumir que *b* representa esta relação e seu valor seria igual a *y/x*:

 

Surge a questão, porém, se a magnitude de *b* é a mesma para outros valores das mesmas variáveis *(x,y)*. Podemos assumer que temos duas observações de cada variável: .

Novamente, pode-se pensar em obter uma medida da relação entre *x* e *y*. Podemos agora considerar que se o mesmo procedimento for repetido, pode-se encontrar valores diferentes para esta relação, digamos . Este procedimento não nos interessa porque ele não captura a relação entre *x* e *y* que seria válida para *quaisquer* valores de *x* e *y*.

Como *x* e *y* podem assumir valores distintos, podemos determinar uma função da forma que passe por ambos pares de observação

1. Escreva uma função geral que repreenta a relação linear entre estes dois pares de pontos. O que os termos *a* e *b* significam?

Agora temos uma função que captura a relação linear entre *x* e *y*. Mas ela é satisfatória? O que aconteceria se encontrássemos um terceiro par de valores para estas variáveis? Repetir o procedimento acima não é adequado, pois podemos utilizar três pares distintos: [, e ] para determinar o valor das constantes *a* e *b*. Apenas em situações muito particulares encontraríamos os mesmos valores em todas as situações. Como queremos encontrar um valor único para *a* e *b,* uma abordagem alterantiva precisa ser utilizada e é aqui que as matrizes são úteis.

Como primeiro passo, pode-se escrever as três equações que relacionam cada par de observações de *x* e *y*. Cada uma dessas relações pode ser expressa usando as mesmas equações como em:

(1)

Ou, de uma forma mais simples:

, for *i = 1,2,3*. (2)

1. Note que na equação acima indexamos as variáveis *x* e *y*, mas não os coeficientes *a* e *b*. Por que? O que isto significa?

O sistema acima pode ser novamente reescrito em termos matriciais como:

**Y** = **XB**  (3)

Onde **Y** é uma matriz 3 x 1 formada pelos valores de *y*; **B** é uma matrix 2 x1 formada pelas constantes, *a* e *b*; e, **X** é uma matriz 3 x 2 formada por um vetor-coluna de 1’s e pelos valores de *x*.

1. Escreva as três matrizes **Y**, **B** e **X** em seu caderno. Mostre que o sistema de equações (1) e a equação matricial (3) representam a mesma informação.
2. Explique porque o vetor-coluna de 1’s na matriz **X** é necessário.
3. Como encontrar o resultado dos valores de *a* e *b*? Mostre utilizando a equação matricial (3).
4. Como o sistema de equações representado acima está associado a um modelo de regressão? Discuta.

Um detalhe a respeito da estimação dos parâmetros de uma regressão via MQO é que este depende da solução de uma expressão quadrática. Então, precisamos para estimar estes parâmetros usando notação matricial entender como esta se expressa em termos quadráticos.

Uma forma de pensar na soma de valores se dá através de multiplicação de matrizes. As somas de quadrados, por exemplo, é obtida facilmente com o produto interno de matrizes. Se lembre que a soma dos quadrados dos elementos de um vetor **x** se dá por:

(4)

Por exemplo, considere o vetor **x** abaixo:

A soma dos quadrados dos elementos de **x** é igual a:

Raciocínio similar pode ser aplicado à soma de produtos de elementos em vetores **x** e **y**. O resultado desta soma equivale a relação abaixo:

(5)

Ao tomarmos uma matriz **X** qualquer, o mesmo resultado segue se calcularmos **X’X.** Os valores da matriz resultante deste produto são os mesmos obtidos por para a i-ésima linha e para a j-ésima coluna.

Como exemplo, assuma a seguinte matriz:

Neste caso, X’X é calculado por:

Note que o valor de é a soma dos quadrados da primeira coluna da matriz original **X**, e o mesmo ocorre com os valores de em relação à segunda coluna da matriz **X**. Entretanto, o valor de é igual ao valor de , que é igual à soma do produto da primeira coluna multiplicado pela segunda. É interessante notar que os valores da diagonal da matriz resultante mostram os valores obtidos por .

1. Usando um software, use a matriz ***X*** para calcular ***X’X***.
2. Defina ***A*** *=* ***X’X***. Encontre a inversa de ***A****,* ou seja,.
3. Assuma que ***Y*** seja uma matriz *3x1* tal como:

Considere ***B***como a matriz equivalente a ***X’Y***. Determine ***B***.

1. Chame a matriz que é equivalente ao produto . Encontre . Interprete os valores obtidos ali

*Módulo II*

Vamos utilizar as técnicas da álgebra linear do Módulo I para resolver um problema mais complexo.

A tabela abaixo registra a renda média per capita e o percentual de trabalhadores empregados pelo governo dos EUA em sete estados diferentes[[1]](#footnote-1):

|  |  |  |
| --- | --- | --- |
| Estado | Renda Per Capita | % Funcionários Públicos |
| Alabama | $ 24,028 | 19.2% |
| Florida | $ 30,446 | 14.5% |
| Georgia | $ 29,442 | 16.4% |
| Mississippi | $ 23,448 | 21.8% |
| Carolina do Norte | $ 28,235 | 17.3% |
| Carolina do Sul | $ 26,132 | 18.2% |
| Tennessee | $ 28,455 | 15.5% |

1. Examine os dados através de um gráfico de dispersão. Há alguma relação específica entre o percentual de funcionários públicos em um estado e a renda per capita entre estes estados? Explique.
2. Agora considere que queremos representar a porcentagem de funcionários públicos (*g*) como uma função linear da renda per capita (*i*). Reescreva os dados indicados na tabela acima como um sistema de equações.
3. Transforme o sistema acima em termos matriciais. Por favor, construa tantas matrizes quantas forem necessárias de tal forma que esteja expresso o sistema em termos matriciais. Especifique as dimensões de todas as matrizes.

Agora, seguindo os passos do modulo I, encontre a matriz

1. Estime os parâmetros deste modelo utilizando os comandos do R. Compare os resultados obtidos na matriz .
2. Escreva a equação estimada e interprete seus valores. Estão de acordo com sua análise no item a) acima?

Podemos agora avançar na avaliação de uma importante hipótese do modelo de regressão. Para isto, precisamos primeiro encontrar os resíduos do modelo.

1. Estime os resíduos para o modelo acima. Pode utilizar tanto os comandos do software, como fazer na “mão”.
2. Utilizando álgebra matricial indicada no Módulo I, encontre a soma dos quadrados dos resíduos. Este dado é uma estimativa do que? Explique.

Uma importante hipótese do modelo de MQO é dada pela expressão abaixo:

1. O que ela significa?

Uma maneira de verificarmos se este é o caso está em mostrar que

1. O que a igualdade acima indica?
2. Construa um gráfico de dispersão que nos ajude a verifica se a hipótese se sustenta. Discuta os resultados do gráfico.

O próprio modelo de regressão pode nos ajudar a indicar se a hipótese se sustenta.

1. Como isto pode ser feito? Obtenha um resultado e novamente, discuta se a hipótese é válida ou não.

1. Este exemplo foi utilizado em Moore & Siegel *A Mathematics Course for Political and Social Research*, p. 403. [↑](#footnote-ref-1)