O MODELO CLÁSSICO DE REGRESSÃO LINEAR

I – Metodologia da Econometria

1. Formulação da teoria ou da hipótese.
2. Especificação do modelo matemático da teoria.
3. Especificação do modelo econométrico da teoria.
4. Obtenção de dados.
5. Estimativa dos parâmetros do modelo econométrico.
6. Teste de hipótese.
7. Previsão ou predição.
8. Utilização do modelo para fins de controle ou política.

A natureza da Análise de Regressão

O termo regressão foi introduzido por Francis Galton. Ele verificou que, embora houvesse uma tendência de pais altos terem filhos altos e de pais baixos terem filhos baixos, a altura média dos filhos de pais de uma dada altura tendia a se deslocar ou "regredir" até a altura média da população como um todo. Em outras palavras, a altura dos filhos de pais extraordinariamente altos ou baixos tende a se mover para a altura média da população.

A interpretação moderna da regressão é diferente – ocupa-se do estudo da dependência de uma variável (chamada variável endógena, resposta ou dependente), em relação a uma ou mais variáveis, as variáveis explicativas (ou exógenas), com o objetivo de estimar e/ou prever a média (da população) ou valor médio de dependente em termos dos valores conhecidos ou fixos (em amostragem repetida) das explicativas.

REGRESSÃO *versus* CAUSALIDADE

É importante ressaltar que embora a análise de regressão lide com a dependência de uma variável em relação a outras variáveis, ela não implica necessariamente em causa. Uma relação estatística, por mais forte e sugestiva que seja, jamais pode estabelecer uma relação causal. As idéias sobre causa devem vir de fora da estatística, enfim, de outra teoria.

REGRESSÃO *versus* CORRELAÇÃO

A análise de regressão conceitualmente é muito diferente da análise de correlação, cujo objetivo básico é medir a intensidade ou o grau de associação linear entre duas variáveis. Por exemplo, podemos estar interessados em achar a correlação entre o hábito de fumar e o câncer no pulmão. Ou ainda, a correlação entre as pontuações em exames de estatística e de matemática.

Na análise de regressão não estamos interessados em tal medição. Em vez disso, tentamos estimar ou prever o valor médio de uma variável com base nos valores fixados de outras variáveis. Assim, podemos querer saber se é possível prever a nota média em uma prova de estatística sabendo a nota de um estudante em uma prova de matemática. O coeficiente de correlação mede a intensidade da associação (linear)

A NATUREZA E AS FONTES DE DADOS PARA ANÁLISE ECONOMÉTRICA - *O sucesso de qualquer análise econométrica depende basicamente da disponibilidade de dados apropriados.*

Capítulo 2 – O conceito de Função de Regressão Populacional (FRP) e Função de Regressão Amostral (FRA)

Função de Regressão Populacional (FRP)

(2.1) $E(Y/X_i) = f(X_i)$

A média condicional é uma função de X_i, em que $f(X_i)$ indica alguma função da variável explicativa X_i. A equação (2.1) é conhecida como função de regressão populacional (FRP) (ou equação de regressão linear) de

duas variáveis. Como uma primeira aproximação ou uma hipótese de trabalho, podemos supor que FRP E(Y/Xi) seja uma função linear de Xi, do tipo

(2.2) $E(Y/X_i) = \beta_1 + \beta_2 X_i$ onde:
β_1 – intercepto
β_2 – coeficiente de inclinação

Especificação estocástica

(2.3) $Y = \beta_1 + \beta_2 X_i + u_i$. Onde ui é uma variável aleatória não-observável que pode assumir valores positivos ou negativos, também conhecido como termo de erro estocástico ou perturbação estocástica.

Função de Regressão Amostral (FRA)

Na maioria das situações práticas é somente uma amostra de valores Y correspondentes e alguns *Xs* fixos. A nossa tarefa é estimar a FRP com base nas informações da amostra.

Analogamente a FRP, que fundamenta a reta de regressão da população, podemos desenvolver o conceito de FUNÇÃO DE REGRESSÃO AMOSTRAL (FRA) para representar a reta de regressão amostral. A amostra contrapartida da Equação (2.2) pode ser escrita como

$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$, onde:

$\hat{Y}_i$ = estimador de $E(Y/X_i)$
$\hat{\beta}_1$ = estimador de β_1
$\hat{\beta}_2$ = estimador de β_2

Estimador e Estimativa

Estimador, também conhecido como uma estatística (baseado na amostra), é simplesmente uma regra, fórmula ou método que nos diz como estimar o parâmetro da população a partir das informações dadas pela amostra disponível.

Estimativa – um valor numérico particular obtido pelo estimador em uma aplicação é conhecido como uma estimativa.

Podemos expressar a FRA por sua fórmula estocástica

$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + u_i$

NOSSO PRINCIPAL OBJETIVO NA ANÁLISE DE REGRESSÃO É ESTIMAR A FRP

$Yi = \beta_1 + \beta_2 Xi + ui$

Com base na FRA

$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + u_i$

MODELO DE REGRESSÃO LINEAR SIMPLES

O MÉTODO DOS MÍNIMOS QUADRADOS ORDINÁRIOS

FRP de duas variáveis:

$Y_i = \beta_1 + \beta_2 X_i + u_i$

Entretanto a FRP não é diretamente observável. Nós a estimamos a partir de FRA:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$$
$$Y_i = \hat{Y}_i + \hat{u}_i$$

Onde:

$\hat{Y}_i$ é o valo (média condicional) estimado de Yi.

$$\hat{u}_i = Y_i - \hat{Y}_i$$
$$\hat{u}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i + \hat{u}_i$$

$\hat{u}_i$ (resíduos) são simplesmente as diferenças entre os valores Y reais e estimados.

Podemos adotar o seguinte critério:

$\sum \hat{u}_i = \sum \left(Y_i - \hat{Y}_i\right)$ seja a menor possível. Se adotarmos o critério de minimizar esta soma todos os resíduos recebem o mesmo peso na soma embora alguns resíduos estejam muito mais próximos de FRA que outros. Ou seja, os resíduos têm igual importância, independentemente de quão próximas ou dispersas as observações individuais sejam relativamente a FRA.

A soma algébrica desses resíduos é zero. Podemos evitar este problema adotando o critério dos mínimos quadrados, segundo o qual a FRA pode ser fixada de tal modo que:

$$\sum (\hat{u}_i)^2 = \sum \left(Y_i - \hat{Y}_i\right)^2$$
$$\sum (\hat{u}_i)^2 = \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)^2$$

- ⇨ Ao elevar $\hat{u}_i$ ao quadrado, este método dá maior peso a resíduos próximos da FRA do que os distantes.
- ⇨ A soma dos resíduos elevados ao quadrado é alguma função dos estimadores $\hat{\beta}_1$ e $\hat{\beta}_2$;
- ⇨ O Método dos mínimos quadrados escolhe $\hat{\beta}_1$ e $\hat{\beta}_2$ de tal maneira que, para uma dada amostra ou conjunto de dados, $\sum (\hat{u}_i)^2$ é a menor possível. Ou seja, para uma dada amostra, o método dos mínimos quadrados nos fornece estimativas únicas de β_1 β_2 que dão o menor valor possível de $\sum (\hat{u}_i)^2$.
- ⇨ Para chegar a este resultado trata-se de um exercício simples de ***cálculo diferencial.***

DERIVAÇÃO DE ESTIMATIVAS POR MÍNIMOS QUADRADOS
EQUAÇÕES NORMAIS

Passo 1:

$$\frac{\partial(\sum \hat{\mu}_i^2)}{\partial \hat{\beta}_1} = \frac{\partial(\sum (Y_i - \hat{Y}_i)^2}{\partial \hat{\beta}_1} = \frac{\partial\left[\sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2\right]}{\partial \hat{\beta}_1}$$
$$= 2\left[\sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)\right](-1)$$
$$= -2\left[\sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)\right]$$
$$= -2\sum \hat{\mu}_i^2 \text{ (}equação\text{ 3.1)}$$

$$\frac{\partial(\sum \hat{\mu}_i^2)}{\partial \hat{\beta}_2} = \frac{\partial(\sum (Y_i - \hat{Y}_i)^2}{\partial \hat{\beta}_2} = \frac{\partial\left[\sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2\right]}{\partial \hat{\beta}_2}$$
$$= 2\left[\sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)\right](-X_i)$$
$$= -2\sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) X_i$$
$$= -2\sum \hat{\mu}_i^2 X_i \text{ (}equação\text{ 3.2)}$$

Passo 2: Equações Normais. Para achar o ponto de mínimo igualamos as equações do passo 1 a zero:

$$-2\sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$
$$\sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$
$$\sum Y_i - \sum \hat{\beta}_1 - \sum \hat{\beta}_2 X_i = 0$$
$$\sum Y_i - n\hat{\beta}_1 - \hat{\beta}_2 \sum X_i = 0$$
$$(equação 3.3) \sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i$$

$$-2\sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) X_i = 0$$
$$\sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) X_i = 0$$
$$\sum \left(Y_i X_i - \hat{\beta}_1 X_i - \hat{\beta}_2 X_i^2\right) = 0$$
$$\sum Y_i X_i - \sum \hat{\beta}_1 X_i - \sum \hat{\beta}_2 X_i^2 = 0$$
$$\sum Y_i X_i - \hat{\beta}_1 \sum X_i - \hat{\beta}_2 \sum X_i^2 = 0$$
$$(equação 3.4) \sum Y_i X_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2$$

Resolvendo as equações normais simultaneamente, obtemos:

$$(equação 3.3) \sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i$$

$$n\hat{\beta}_1 = \sum Y_i - \hat{\beta}_2 \sum X_i$$

$$(equação 3.3.1) \hat{\beta}_1 = \frac{\sum Y_i}{n} - \hat{\beta}_2 \frac{\sum X_i}{n}$$

ou

$$(equação 3.3.1) \hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}$$

$$(equação 3.4) \sum X_i Y_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2$$

$$\sum X_i Y_i = \left[\frac{\sum Y_i}{n} - \hat{\beta}_2 \frac{\sum X_i}{n} \right] \sum X_i + \hat{\beta}_2 \sum X_i^2$$

$$\sum X_i Y_i = \frac{\sum X_i \sum Y_i}{n} - \hat{\beta}_2 \frac{\left(\sum X_i\right)^2}{n} + \hat{\beta}_2 \sum X_i^2$$

$$\sum X_i Y_i = \frac{\sum X_i \sum Y_i}{n} + \hat{\beta}_2 \left[\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n} \right]$$

$$\hat{\beta}_2 \left[\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n} \right] = \sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}$$

$$\hat{\beta}_2 \left[\frac{n\sum X_i^2 - \left(\sum X_i\right)^2}{n} \right] = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n}$$

$$\hat{\beta}_2 \left\{ (1/n) \left[n\sum X_i^2 - \left(\sum X_i\right)^2 \right] \right\} = (1/n) \left[n\sum X_i Y_i - \sum X_i \sum Y_i \right]$$

$$\hat{\beta}_2 = \frac{(1/n) \left[n\sum X_i Y_i - \sum X_i \sum Y_i \right]}{(1/n) \left[n\sum X_i^2 - \left(\sum X_i\right)^2 \right]}$$

$$\hat{\beta}_2 = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2}$$

$$\hat{\beta}_2 = \frac{\sum \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(X_i - \overline{X}\right)^2}$$

ou

$$\hat{\beta}_2 = \frac{\sum \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(X_i - \overline{X}\right)^2}$$

$$\hat{\beta}_2 = \frac{\sum (x_i)(y_i)}{\sum (x_i)^2}$$

Onde:

$$x_i = \left(X_i - \overline{X}\right)$$

$$y_i = \left(Y_i - \overline{Y}\right)$$

$$(x_i)^2 = \left(X_i - \overline{X}\right)$$

O Modelo Clássico de Regressão Linear: As Hipóteses Subjacentes ao Método dos Mínimos Quadrados

O nosso objetivo não é somente obter $\hat{\beta}_1$ e $\hat{\beta}_2$, mas também fazer inferências sobre os verdadeiros β_1 e β_2. Por exemplo, gostaríamos de saber quão próximo $\hat{Y}_i$ é do verdadeiro $E(Y/X_i)$. Para tanto, devemos não apenas especificar a forma funcional do modelo, como, mas também formular certas hipóteses sobre o modo pelo qual $\hat{Y}_i$ são gerados.

Precisão ou Erros-padrão das Estimativas por Mínimos Quadrados

As estimativas por mínimos quadrados são uma função dos dados da amostra. Mas como os dados provavelmente variam de amostra para amostra, as estimativas variarão. Conseqüentemente, o que se necessita é de alguma medida de "confiabilidade" ou precisão dos estimadores $\hat{\beta}_1$ e $\hat{\beta}_2$. Na estatística, a precisão de uma estimativa é medida por seu erro-padrão (ep).

Data as hipóteses os erros-padrão das estimativas por MQO podem ser obtidos:

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2}$$

$$ep(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum x_i^2}}$$

$$\text{var}(\hat{\beta}_1) = \frac{\sum X_i^2}{n\sum x_i^2}\sigma^2$$

$$ep(\hat{\beta}_1) = \sqrt{\frac{\sum X_i^2}{n\sum x_i^2}}\sigma$$

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2}$$

Em que ***var*** = variância e ***ep*** = erro-padrão, e σ^2 é a variância constante ou homoscedástica de u*i* da Hipótese 4 (ver capítulo 2 e 3).

Propriedades dos Estimadores de Mínimos Quadrados: O Teorema de Gauss-Markov.

Dada as hipóteses do modelo clássico de regressão linear, as estimativas por mínimos quadrados possuem algumas propriedades ideais ou ótimas. Estas propriedades estão contidas no conhecido teorema de Gauss-Markov. Para entender este teorema, precisa considerar a **propriedade do melhor estimador linear não-viesado** para um estimador.

O estimador de MQO $\hat{\beta}_2$, é um melhor estimador linear não-viesado (MELNV) de $\hat{\beta}_2$, caso seja válido o seguinte:

1. É linear, isto é, uma função linear de uma variável aleatória, tal como a variável dependente Y no modelo de regressão.
2. É não-viesado, isto é, seu valor médio ou esperado, E($\hat{\beta}_2$), é igual ao valor verdadeiro, $\hat{\beta}_2$.
3. Tem mínima variância na classe de todos esses estimadores lineares não-viesados; um estimador não-viesado com a menor variância é conhecido como um estimador eficiente.

O Coeficiente de Determinação r^2: Uma Medida do Grau de Ajuste.

O **coeficiente de determinação** r^2 (caso de duas variáveis) ou R^2 (regressão múltipla) é uma medida sintética que diz quão bem a reta de regressão da amostra se ajusta aos dados.

$$r^2 = \frac{\sum\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum\left(Y_i - \bar{Y}\right)^2} = \frac{SQE}{SQT}$$

Seja:
SQT = SQE + SQR

SQT = Soma dos Quadrados Total (SQT)
SQE = Soma dos Quadrados Explicada (SQE)
SQR = Soma dos Quadrados dos Resíduos (SQR)

Alternativamente podemos escrever:

$$r^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum\left(Y_i - \bar{Y}\right)^2}$$
$$= 1 - \frac{SQR}{SQT}$$

A quantidade r^2 assim definida é conhecida como coeficiente de determinação (da amostra) e é a medida utilizada do grau de ajuste de uma reta de regressão. Traduzindo, r^2 mede a proporção ou a porcentagem da variação total em Y explicada pelo modelo de regressão.

Duas propriedades de r2 podem ser destacadas:

1. É uma quantidade não-negativa.
2. Seus limites são $0<=r^2<=1$. Um r^2 igual a 1 significa um perfeito ajuste, isto é $\hat{Y}_i = Y_i$ para todo i. Por outro lado, um r^2 igual a zero significa que não há nenhuma relação entre o regredido e o regressor.