Tutorial: Regressão Linear Simples no Matlab/Octave

Prof. Dr. Guilherme de Alencar Barreto

Março /2014

Departmento de Engenharia de Teleinformática (DETI)
Programa de Pós-Graduação em Engenharia de Teleinformática (PPGETI)
Universidade Federal do Ceará (UFC), Campus do Pici, Fortaleza-CE

gbarreto@ufc.br

Objetivo - Apresentar uma sequência de comandos do Matlab/Octave para gerar dados correspondentes à simulação de um sistema linear com ruído, a respectiva estimação dos parâmetros do sistema a partir dos dados gerados e a validação do modelo de regressão encontrado através da análise dos resíduos.

1 Comandos no Matlab/Octave

- **Passo 1**: Especificar os parâmetros β_0 e β_1 do sistema linear a ser simulado, bem como a variância do ruído σ_ε^2.

  ```
  >> B0=2; B1=0.8;
  >> Vnoise=0.25;
  ```

  Comentário 1: Caso vocês desejem visualizar os valores definidos em `B0`, `B1` e `Vnoise` basta retirar o ponto-e-vírgula que aparece ao final de cada comando. Caso você já tenha digitado com o ponto-e-vírgula, basta chamar o vetor correspondente digitando seu símbolo na linha de comando, individualmente (e.g. `>> B0`) ou separados por vírgulas (`>> B0,B1`). Experimente!

- **Passo 2**: Definir a faixa de valores para x. Por exemplo, de 0 a 5, em incrementos de 0,01.

  ```
  >> x=0:0.01:5;
  >> n=length(x);
  ```

- **Passo 3**: Simular o sistema linear $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

  ```
  >> for i=1:n, ...
  y(i)=B0+B1*x(i)+normrnd(0,sqrt(Vnoise)); ...
  end
  ```

Comentário 2: O comando `normrnd` gera um número aleatório normalmente distribuído de média 0 e desvio-padrão igual $\sqrt{Vnoise}$. Note que especificamos anteriormente a variância do ruído. Para obter o desvio-padrão, basta extrair a raiz quadrada da variância. Para mais detalhes do comando `normrnd`, digite `>> help normrnd`.

- **Passo 4**: Usar os dados gerados para implementar as seguintes equações de estimação de β_0 e β_1:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \tag{1}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \tag{2}$$

```
>> mx=mean(x); my=mean(y); Sxy=dot(x,y); Sx=sum(x); Sy=sum(y);
Sx2=sum(x.^2); S2x=Sx^2;
>> num=Sxy - Sx*Sy/n;
>> den=Sx2 - S2x/n;
>> B1h=num/den
>> B0h=my-B1h*mx;
>> B0h, B1h
B0h=
  2.0201
B1h=
  0.7875
```

Comentário 3: Note que na sua simulação os valores de $\hat{\beta}_0$ e $\hat{\beta}_1$ não serão os mesmos que os mostrados acima, pois os valores do ruído gerados pelo comando `normrnd` serão diferentes cada vez que você rodar a simulação.

- **Passo 5**: Calcular os valores de $\hat{y}_i$ usando a seguinte reta de regressão:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{3}$$

```
>> for i=1:n, ...
yh(i)=B0h+B1h*x(i); ...
end
```

- **Passo 6**: Calcular os resíduos (i.e. erros de predição) resultantes, $e_i = y_i - \hat{y}_i$, bem como estimar a variância do ruído ($\hat{\sigma}_\varepsilon^2$).

```
>> residuos = y - yh;
>> Vnoiseh= sum(residuos.^2)/(n-2)
Vnoiseh=
       0.2516
```

- **Passo 7**: Avaliar *qualitativamente* a forma da distribuição dos resíduos usando histograma.

```
>> histfit(residuos)
```

Comentário 4: Quanto mais semelhante à distribuição Normal melhor.

- **Passo 8**: Avaliar *quantitativamente* a gaussianidade da distribuição dos resíduos através de um teste de hipótese. Neste exemplo, vamos usar o teste de Kolmogorov-Smirnov. Para isso, temos antes que normalizar a variância dos resíduos para 1. Se o resultado for $H = 0$, a hipótese nula de que a distribuição dos resíduos é $N(0, 1)$ deve ser aceita. Se $H = 1$, tal hipótese deve ser rejeitada (ou como preferem dizer os Estatísticos, não há informação suficiente para aceitar a hipótese nula.)

```
>> residuos_norm=residuos/std(residuos);
>> H=kstest(residuos_norm)
H=
       0
```

- **Passo 9**: Avaliar a distribuição acumulada (FDA) dos resíduos normalizados versus a FDA da distribuição normal padronizada (i. e. $N(0, 1)$) através do comando `cdfplot`.

```
>> figure; cdfplot(residuos_norm); hold on
>> z=randn(n,1); cdfplot(z); hold off
```

  Comentário 5: O resultado está mostrado na figura abaixo. Esta análise comparativa das FDAs dos resíduos normalizados e da distribuição Normal padronizada $N(0, 1)$ é basicamente o que o teste de Kolmogorov-Smirnov faz de modo quantitativo.

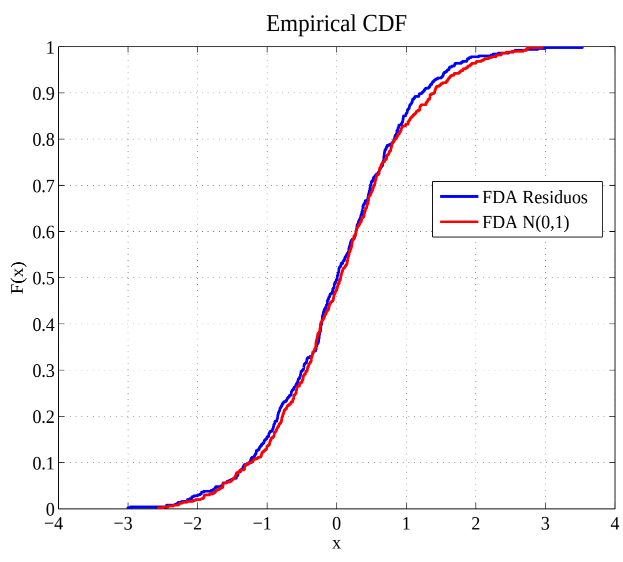

Figura 1: FDAs empíricas: Resíduos normalizados versus amostra de uma distribuição $N(0, 1)$.