Validação Cruzada

Rachid Muleia

A validação cruzada é uma técnica que nos permite comparar valores previstos com valores observados. Em dados espaciais esta técnica pode nos ajudar a decidir qual modelo de semi-variograma escolher ou qual método de predição dá melhores resultados. A idéia básica da validação cruzada é a seguinte: Omite-se o ponto i do conjunto de dados e interpola-se o mesmo usando os restantes $n-1$ pontos. Portanto, podemos comparar o valor previsto com o valor verdadeiro na localização. Outra maneira é dividir o conjunto de dados em duas partes. A primeira parte será usada para modelizar o variograma. A localização espacial da outra parte do conjunto de dados será a nossa malha. Depois de prever os valores, podemos compará-los com os valores observados nesses locais.

Para ilustração da *validação cruzada* iremos usar a base de dados *meuse river*

```
library(sp)
data("meuse")
lead_zinc.df=meuse[, c(1,2,5,6)] # vamos apenas trabalhar com as variaveis
chumbo e zinco
head(lead_zinc.df) #visualiza as primeiras 6 linhas da base de dados
      x y lead zinc
1 181072 333611 299 1022
2 181025 333558 277 1141
3 181165 333537 199 640
4 181298 333484 116 257
5 181307 333330 117 269
6 181390 333260 137 281
```

Vamos dividir aleatoriamente os dados em duas partes. Serão utilizadas 100 observações para modelização e 55 para predição. Aqui estão os comandos:

```
choose100 <- sample(1:155, 100)
part_model <- lead_zinc.df[choose100, ]
part_valid <- lead_zinc.df[-choose100, ]
```

Nota: Observe que esta é uma seleção aleatória e cada vez que executamos esses comandos, obteremos amostras diferentes.

Para estimar o semivariograma vamos usar a base de dados part_model

```
g=gstat(id="log_lead", formula = log(lead)~1, locations = ~x+y,  data=part_model)
q=variogram(g) # calcula o semivariograma experiemntal
plot(q) # representacao grafica do semivariograma experimental

v.fit <- fit.variogram(q, vgm(1, "Sph", 800, 1))
plot(q, v.fit) # representacao do semivaiograma teorico junto ao semivariograma experimental
```

Agora que temos as estimativas da krigagem, vamos calcular as diferencas entre os valores previstos e observados

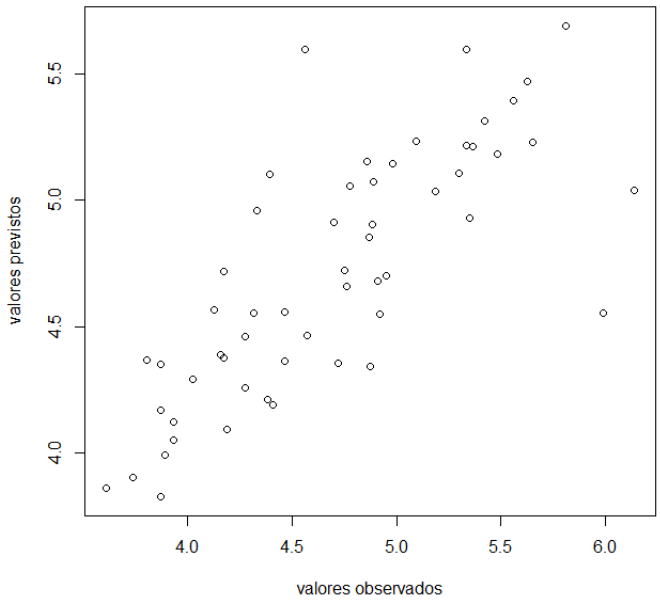
```
difference <- log(part_valid$lead) - part_valid_pr$log_lead.pred
summary(difference)
```

Assim que temos as diferencas entre os valos previstos e observados podemos calular , o erro médio da estimativa, erro quadrático médio, erro quatrático médio padronizado, e a o coeficiente de correlação.

```
> mean(difference ) #erro medio
[1] -0.01452672
> mean(difference^2) #erro quadratico medio
[1] 0.1553702
> mean(difference^2/part_valid_pr$log_lead.var) #erro quadratico medio padronizado
[1] 0.7394981
cor(part_valid_pr$log_lead.pred,log(part_valid$lead)) # correlacao entre valores
previstos e observados
[1] 0.7726411
```

Além de calcularmos a correlação podemos também representar os valores previstos e observados num diagrama de dispersão

```
plot(log(part_valid$lead),part_valid_pr$log_lead.pred, xlab="valores observados",
ylab="valores previstos")
```



Uma forma mais automatizada de fazer a validação cruzada é usar o comando `krige.cv`

```
cv_pr <- krige.cv(log(lead)~1, data=lead_zinc.df, locations=~x+y,
                  model=v.fit, nfold=nrow(lead_zinc.df))
```

```
> summary(cv_pr)
   var1.pred        var1.var         observed        residual
 Min.   :3.685   Min.   :0.1240   Min.   :3.611   Min.   :-1.0465998
 1st Qu.:4.375   1st Qu.:0.1565   1st Qu.:4.284   1st Qu.:-0.2146486
 Median :4.786   Median :0.1723   Median :4.812   Median :-0.0233480
 Mean   :4.808   Mean   :0.1813   Mean   :4.807   Mean   :-0.0008775
 3rd Qu.:5.202   3rd Qu.:0.1931   3rd Qu.:5.333   3rd Qu.: 0.1845019
 Max.   :6.073   Max.   :0.4912   Max.   :6.483   Max.   : 1.6235280
     zscore             fold             x               y
 Min.   :-2.4714   Min.   :  1.0   Min.   :178605   Min.   :329714
 1st Qu.:-0.4799   1st Qu.: 39.5   1st Qu.:179371   1st Qu.:330762
 Median :-0.0593   Median : 78.0   Median :179991   Median :331633
 Mean   :-0.0008   Mean   : 78.0   Mean   :180005   Mean   :331635
 3rd Qu.: 0.4344   3rd Qu.:116.5   3rd Qu.:180630   3rd Qu.:332463
 Max.   : 3.8577   Max.   :155.0   Max.   :181390   Max.   :333611
```

Se quisermos comparar dois semivariogramas (exponencial e esférico), ou escolher entre métodos de predição (por exemplo, krigagem ordinária, krigagem universal, método de interpolação inversa ponderada a distância , etc.), ou entre diferentes tipos de semivariogramas experimentais(clássicos ou robustos) ou diferentes Pesos, podemos usar o erro médio, erro quadrático médio, ou ainda a soma dos quadrados dos resíduos. Em geral seleccionamos o metódo que tiver menor soma dos quadrados dos resíduos.

$$\text{PRESS} = \sum_{i=1}^{n} (z(s_i) - \hat{z}(s_i))^2$$

PRESS(Prediction sum of squares)

Faça o mesmo exercício usando a variável zinco.