Modelos de Regressão Múltipla

Modelo geral

Um modelo de regressão múltipla é expresso como:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

em que:

y_i : valores da variável resposta, i = 1, 2,..., n observações;

x_{ki} : valores das variáveis explicativas, k = 1, 2,..., K variáveis;

β_k : parâmetros do modelo;

ε_i : erro aleatório.

A equação estimada para este modelo é definida como:

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki}$$

em que:

b_k : coeficientes estimados.

Variável dummy

Em algumas situações é necessário introduzir, como variável preditora (independente), uma variável categórica no modelo de regressão linear simples ou múltiplo, como por exemplo, local (urbano ou rural), área (preservada ou degradada), etc, podendo ter mais que duas categorias. Essa variável terá que ser codificada, utilizando somente códigos 0 e 1, assim chamada variável dummy.

O número de variáveis dummy no modelo será sempre igual ao número de categorias da variável preditora original menos 1. Por exemplo:

- Para a variável preditora 'local' que assume valores - urbano ou rural, então têm-se a variável dummy local_dummy assumindo 0 para rural e 1 para urbano; também, poderia ser utilizado 1 para rural e 0 para urbano. Uma indicação é que a categoria que assume o valor 0 seja a categoria de referência.
- Para a variável preditora 'grau de escolaridade que assume valores – ensino fundamental, ensino médio, ensino superior, então têm-se as variáveis dummy: escola1 e escola2, assim definido:
  - escola1=0 e escola2=0 para ensino fundamental;
  - escola1=1 e escola2=0 para ensino médio;
  - escola1=0 e escola2=1 para ensino superior.

Exercício:

1) Utilizando o banco de dados ARVORE2, ajuste um modelo de regressão linear simples para predizer a altura das árvores em função do diâmetro. Veja essa relação no diagrama de dispersão. Interprete os resultados.

Relembrando Modelos de Regressão Linear Simples – Curso Básico do Software R

1.1 Ajustar a equação de regressão. Interpretá-la.

1.2 Encontrar e interpretar a significância da equação.

1.3 Encontrar e interpretar o coeficiente de determinação.

1.4 Analisar graficamente os resíduos.

1.5 Testar a normalidade dos resíduos.

Adicionalmente - Curso Avançado do Software R

1.6 Analisar pontos outliers nos resíduos.

Para análise dos valores outliers nos resíduos (residuals standard e residuals studentized), utilizam-se os seguintes comandos:

Sintaxe no software R:

```
> rstudent(regressao)
> rstandard(regressao)
```

E o gráfico para verificar valores outliers nos resíduos:

Sintaxe no software R:

```
> plot(rstudent(regressao))
> plot(rstandard(regressao))
```

Aqueles valores maiores que |2| são possíveis outliers. Incluir uma linha y =2 e y=-2, para facilitar a visualização de outliers.

1.7 Analisar pontos influentes nos resíduos.

Para análise dos valores influentes, utiliza-se:

Sintaxe no software R:

```
> dffits(regressao)
```

Aqueles valores maiores que 2*(p/n)^(1/2) são possíveis pontos influentes. Em que, p = número de parâmetros do modelo e n = tamanho da amostra. O gráfico para detectar pontos influentes pode ser elaborado pelo comando:

Sintaxe no software R:

```
> plot(dffits(regressao))
```

Aqueles valores maiores, em módulo, são possíveis influentes. Incluir linhas para facilitar a visualização de pontos influentes.

Ainda, pode-se utilizar o comando **plot(regressao)** elabora diferentes gráficos para o diagnóstico do modelo.

2) Ajuste um segundo modelo de regressão linear simples para predizer a altura das árvores em função da espécie. Veja essa relação no diagrama de dispersão. Interprete os resultados.

3) Ajuste um terceiro modelo de regressão múltipla para predizer a altura das árvores em função do diâmetro e da espécie. Interprete os resultados.

```
> modelom=lm(altura_m~diametro_cm+especie)

> modelom

Call:
lm(formula = altura_m ~ diametro_cm + especie)

Coefficients:
(Intercept)  diametro_cm      especie
   12.32845      0.05758     -1.42281
```

Modelo:

$$Y = 12{,}328 + 0{,}0576x1 - 1{,}423x2$$

Ou

$$\text{Altura} = 12{,}328 + 0{,}0576\text{diâmetro} - 1{,}423\text{espécie}$$

Verificando a significância de cada coeficiente do modelo de regressão múltipla:

```
> summary(modelom)

Call:
lm(formula = altura_m ~ diametro_cm + especie)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2688 -0.7663 -0.1236  0.8132  2.8727

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.69592    0.38639  32.857  < 2e-16 ***
diametro_cm  0.05713    0.00445  12.837  < 2e-16 ***
especie     -1.62517    0.24459  -6.644 1.52e-09 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.185 on 102 degrees of freedom
Multiple R-squared:  0.6995,   Adjusted R-squared:  0.6937
F-statistic: 118.7 on 2 and 102 DF,  p-value: < 2.2e-16
```

Verificar a significância do modelo completo.

Verificar o coeficiente de determinação do modelo.

Realizar análise dos resíduos.

- gráfico dos resíduos com cada variável preditora
- resíduos padronizados para verificar outlier
- verificar pontos infuentes

A interpretação dos termos de regressão é um pouco mais complicada. Em geral, um modelo com múltiplos preditores indica a diferença média na variável desfecho quando mudamos o valor de uma variável e mantemos a outra constante (Kaszubowski, 2016).

Nesse caso, entre árvores de mesmo diâmetro (x1), a diferença média esperada da altura (y) para a espécie *Syphoneugena reitzii* em relação a espécie Sebastiania commersoniana é de cerca de 1,42m a menos (pois b3=-1,42).

Da mesma forma, árvores da mesma espécie têm, em média, 0,05758m (pois b2=0,05758) a mais a cada 1 cm de diâmetro.

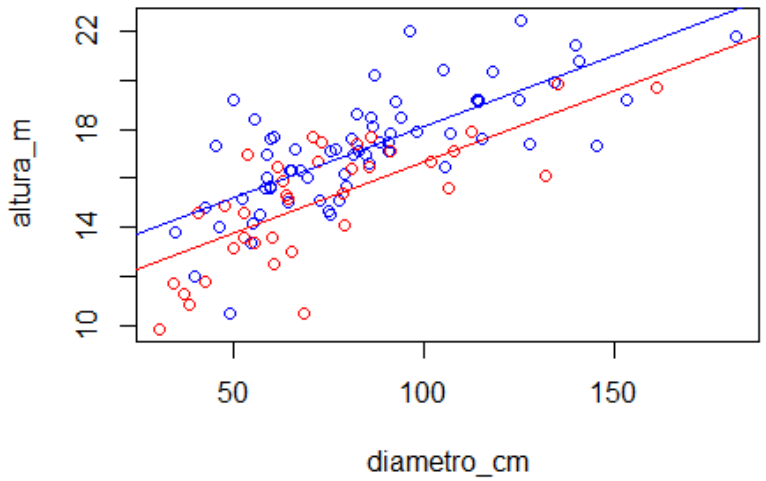
Como envolvem mais variáveis, não é possível resolver o modelo inteiro num único gráfico. Como alternativa, pode-se plotar a reta para cada espécie (variável categórica).

Primeiro, os pontos são plotados. O argumento type='n' indica que não é para acrescentar nenhum ponto ao gráfico.

Em seguida, os pontos são acrescentados separadamente, com a função points, a qual acrescenta pontos ao gráfico, sendo que o colchetes [espécie==0] seleciona somente os casos desejados.

Por fim, acrescentamos as retas de regressão para cada resposta a variável independente espécie. Usamos a função 'coef' para extrair os coeficientes de interesse.

```
> plot(diametro_cm,altura_m)
> plot(diametro_cm,altura_m,type='n')              # Gera o gráfico sem pontos

> points(diametro_cm[especie==0],altura_m[especie==0],col='blue') # Acrescenta os pontos
> points(diametro_cm[especie==1],altura_m[especie==1],col='red')

> abline(coef(modelom)[1], coef(modelo)[2], col='blue')     # Acrescenta as linhas
> abline(coef(modelom)[1]+coef(modelom)[3], coef(modelom)[2], col='red')
```



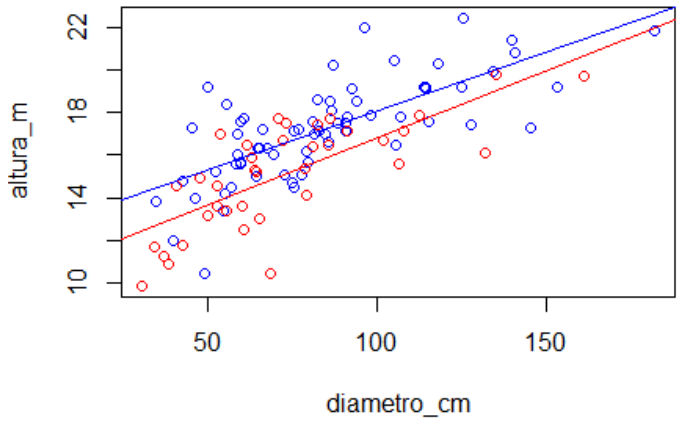
Interação entre variáveis preditoras

Quando suspeita-se que os coeficientes de inclinação podem variar entre as categorias da variável preditora então aconselha-se testar a interação entre as duas variáveis. No software R utiliza-se ':' para indicar a interação entre as duas variáveis. Se a interação for significativa (P<0,05), então conclui-se que os coeficientes de inclinação diferem entre si.

O modelo de regressão múltipla apresentando anteriormente pressupõe que a inclinação da reta de regressão é igual para os dois grupos considerados, espécie *Syphoneugena reitzii* e espécie Sebastiania commersoniana espécie. Se existem motivos para acreditar que a inclinação pode variar de um grupo para o outro, pode-se acrescentar um termo de interação (interação entre variáveis) (Kaszubowski, 2016).

A interação, neste caso, nada mais é do que o acréscimo de uma nova variável preditora ao modelo. Essa nova variável preditora é o produto das duas variáveis que já constam no modelo. Para acrescentar um termo de interação no R, basta utilizar dois pontos ':' entre o nome das duas variáveis para as quais se deseja criar o termo de interação.

```
> modelom=lm(altura_m~diametro_cm+especie+diametro_cm:especie)
> modelom

Call:
lm(formula = altura_m ~ diametro_cm + especie + diametro_cm:especie)

Coefficients:
        (Intercept)          diametro_cm
           11.54805              0.07137
            especie  diametro_cm:especie
            0.78630             -0.03158


> summary(modelom)

Call:
lm(formula = altura_m ~ diametro_cm + especie + diametro_cm:especie)

Residuals:
     Min       1Q   Median       3Q      Max
-2.24597 -0.85455  0.06317  0.75516  2.55607

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         11.548047   0.475441  24.289  < 2e-16
diametro_cm          0.071375   0.005655  12.620  < 2e-16
especie              0.786303   0.683038   1.151 0.252375
diametro_cm:especie -0.031579   0.008421  -3.750 0.000295

(Intercept)         ***
diametro_cm         ***
especie
diametro_cm:especie ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.116 on 101 degrees of freedom
Multiple R-squared:  0.7363,   Adjusted R-squared:  0.7284
F-statistic: 93.99 on 3 and 101 DF,  p-value: < 2.2e-16


> plot(diametro_cm,altura_m,type='n')

> points(diametro_cm[especie==0],altura_m[especie==0],col='blue')
> points(diametro_cm[especie==1],altura_m[especie==1],col='red')

> abline(coef(modelom)[1],coef(modelom)[2], col='blue')
> abline(coef(modelom)[1]+coef(modelom)[3],coef(modelom)[2]+coef(modelom)[4],
col='red')
```



Métodos seleção de variáveis na regressão múltipla

Full model – Modelo completo

Sintaxe no software R para um modelo de regressão múltipla com três variáveis preditivas:

```
> regressao=lm(y~x1+x2+x3)
> summary(regressao)
```

Existem três métodos de seleção de variáveis para modelos de regressão múltipla: backward, forward e stepwise.

Sintaxe no software R:

```
> regressao=step(lm(y~x1+x2+x3),direction = 'método')
```

backward
forward
stepwise

Procedimento backward

Considera todas as variáveis inicialmente, testando posteriormente, a permanência de cada uma no modelo. Se p ≤ 15%, permanece no modelo (saiu do modelo não entra mais) (Riboldi, 2005).

Passo 1) Ajustar o modelo completo de m variáveis e obter SQR^{c}_{eg} e $\hat{\sigma}^2$;
Passo 2) Para cada uma das m variáveis do modelo completo do passo 1, considerar o modelo reduzido – retirando esta variável – e calcular SQR^{r}_{eg} para obter o valor da estatística (slide 24);
Passo 3) Achar o mínimo dos m valores da estatística obtidos no passo 2, denotado por F_{min};
Passo 4) Seja F_{out} o valor da distribuição F com 1 e (n-m-1) gl;
Se $F_{min} > F_{out}$ -> interromper o processo e optar pelo modelo completo desta etapa;
Se $F_{min} < F_{out}$ -> voltar ao passo 1, iniciando nova etapa em que o modelo completo tem (m-1) variáveis – dada a eliminação da variável cuja estatística é igual a F_{min}.

Procedimento forward

Inclui uma variável de cada vez, se p ≤ 20%, entra no modelo. Este método não testa a permanência da variável (entrou no modelo não sai mais) (Riboldi, 2005).

Passo 1) Ajustar o modelo reduzido de m variáveis e obter SQR^{c}_{eg}

Passo 2) Para cada variável não pertencente ao modelo do passo 1, considerar o modelo completo com adição desta variável extra e calcular SQR_{eg}^{r} e $\hat{\sigma}^2$ para obter o valor da estatística (slide 26);
Passo 3) Achar o máximo dos valores da estatística obtidos no passo 2, denotado por F_{max};
Passo 4) Seja F_{in} o valor da distribuição F com 1 e (n-m) gl;

Se $F_{max} > F_{in}$ -> voltar ao passo 1, iniciando nova etapa em que o modelo reduzido tem (m+1) variáveis – dada a inclusão da variável cuja estatística é igual a F_{max}.

Se $F_{max} < F_{in}$ -> interromper o processo e optar pelo modelo reduzido desta etapa;

Procedimento stepwise

Inclui as variáveis passo-a-passo e testa a permanência (as variáveis podem entrar e sair do modelo) (Riboldi, 2005).

Passo 1) Ajustar o modelo reduzido de m variáveis e obter SQR_{eg}^{r};
Passo 2) Para cada variável não pertencente ao modelo do passo 1, considerar o modelo completo - com adição desta variável extra - e calcular SQR_{eg}^{c} e $\hat{\sigma}^2$ para obter o valor da estatística (slide 26);
Passo 3) Achar o máximo dos valores da estatística obtidos no passo 2, denotado por F_{max};
Passo 4) Seja F_{in} o valor da distribuição F com 1 e (n-m) gl;

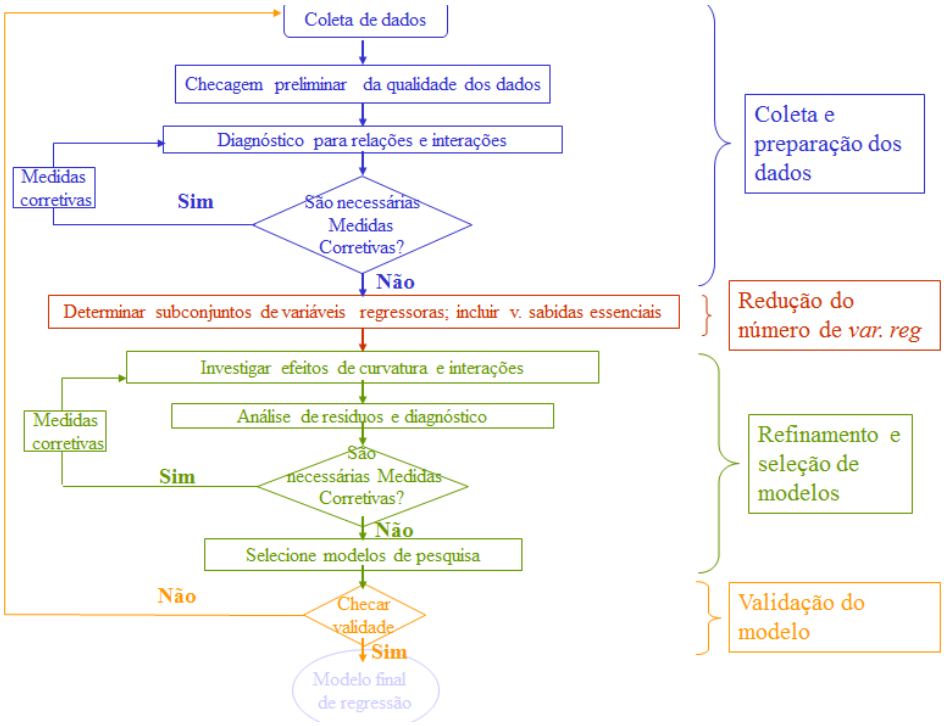
Se $F_{max} > F_{in}$ -> passar ao passo 5, com modelo completo composto por (m+1) variáveis – as m variáveis do modelo do passo 1 e a variável cuja estatística é igual a F_{max}.

Se $F_{max} < F_{in}$ -> passar ao passo 5, com modelo completo igual ao modelo do passo 1 ou encerrar o processo se no passo 8 da etapa anterior, nenhuma variável tiver sido eliminada;
Passo 5) Ajustar o modelo completo de k variáveis – sendo k igual a m ou (m+1), e obter SQR_{eg}^{c} e $\hat{\sigma}^2$
Passo 6) Para cada uma das k variáveis do modelo completo do passo 5, considerar o modelo reduzido – retirando esta variável – e calcular SQR_{eg}^{r} para obter o valor da estatística;
Passo 7) Achar o mínimo dos k valores da estatística obtidos no passo 6, denotado por F_{min};
Passo 8) Seja F_{out} o valor da distribuição F com 1 e (n-k-1) gl;

Se $F_{min} > F_{out}$ -> não eliminar nenhuma variável e voltar ao passo 1, iniciando nova etapa com modelo reduzido com k variáveis ou encerrar o processo de no passo 4 nenhuma variável tiver sido anexada;

Se $F_{min} < F_{out}$ -> eliminar a variável cuja estatística é igual a F_{min} e voltar ao passo 1 iniciando nova etapa com modelo reduzido com (k-1) variáveis.

Figura 1 – Modelagem estatística



Fonte: Riboldi, 2005.

Roteiro para o diagnóstico do modelo de regressão múltipla ajustado

Identificação de observações destoantes para Y

- Resíduo studentizado externamente – r_student (studentized residual with currente observation deleted)
- Resíduo studentizado internamente – student (studentized residual)
- Identificação de observações destoantes com base nos resíduos – residual

Identificação de observações destoantes para X

- Matriz H
- Alavanca (Leverage = h_i)

Identificação de casos de influência

- DFFITS (standard influence of observation on predict values)
- Distância de Cook (_cookd)
- DFBetas

Verificação da existência de multicolinearidade (correlação entre os X's)

- Matriz de correlação das variáveis
- Análise de estrutura k (condition index)
- Fator de inflação de variância – VIF (variance inflation)
- Teste de Durbin-Watson pra atutocorrelação

Referências

HOFFMANN, R.; VIEIRA, S. **Análise de Regressão.** Uma introdução à Econometria. Hucitec: São Paulo, SP. 1998;

RIBOLDI, J. **Modelos Lineares**: notas de aula. Programa de Pós-Graduação em Epidemiologia. Faculdade de Medicina. UFRGS, 2005.