Regressão Multivariada Clássica e Regressão Logística

Utilizando o pacote *car*

Rogério Hultmann (GRR20137589), Adriane Machado (GRR20149152) e Rodrigo Paifer (GRR????????)

12 de junho de 2017

Introdução

O pacote "car" FOX et al. (2016) é um pacote do R R CORE TEAM (2015), que é uma linguagem para manipulação de dados e análises estatísticas, LANDEIRO; ECOLOGIA (2011). Fornece muitas funções que são aplicadas a um modelo de regressão ajustado, inclusive para dados de medidas repetidas.

Material e Métodos

O pacote "car" contém funções e conjuntos de dados associados ao livro An R Compation to Applied Regression, FOX; WEISBERG (2010). Que por sua vez, trata de fornecer uma ampla introdução ao R, no contexto de análise de regressão aplicada.

Dentre a ampla aplicação do pacote, temos os modelos de Regressão multivariada clássica e regressão logística. O modelo de regressão logística é similar ao modelo de regressão linear, porém neste a variável resposta é binária, portanto, assume apenas os valores de sucesso e fracasso. Outra aplicabilidade é o modelo regressão multivariada clássica, que constrói o modelo considerando as correlações LÊ et al. (2008).

Para aplicação dos dois modelos, utilizaremos os bancos de dados "Mroz" e "OBrienKaiser":

Banco de dados

Mroz

Este data frame é composto por 753 mulheres casadas, nas quais foram observadas as seguintes variáveis:

- lfp = Participação no mercado de trabalho
- k5 = Número de filhos de 5 anos ou menos.
- k618 = Número de filhos de 6 a 18 anos.
- age = anos
- wc = Esposa fez faculdade
- hc = Marido fez faculdade
- lwg = O salário esperado, para as mulheres que trabalham, o salário real e para as mulheres que não trabalham, um valor estipulado baseado na regressão de lwg nas outras variáveis.
- inc = Rendimento familiar exceto rendimento da esposa

A seguir, uma parte dos dados para visualização da composição do *data frame*

```
require("car")
head(Mroz, n=4L)
```

```
##   lfp k5 k618 age wc hc       lwg   inc
## 1 yes  1    0  32 no no 1.2101647 10.91
## 2 yes  0    2  30 no no 0.3285041 19.50
## 3 yes  1    3  35 no no 1.5141279 12.04
```

```
## 4 yes  0    3  34 no no 0.0921151  6.80
```

Iris

Este data frame é composto por dados imaginários em que 16 sujeitos do sexo feminino e masculino, nos quais foram observadas as seguintes variáveis:

- *treatment* = Tratamento A ou B
- *gender* = gênero
- *pre.1* = Pré-teste, hora 1
- *pre.2* = Pré-teste, hora 2
- *pre.3* = Pré-teste, hora 3
- *pre.4* = Pré-teste, hora 4
- *pre.5* = Pré-teste, hora 5
- *post.1* = Pós-teste, hora 1
- *post.2* = Pós-teste, hora 2
- *post.3* = Pós-teste, hora 3
- *post.4* = Pós-teste, hora 4
- *post.5* = Pós-teste, hora 5
- *fup.1* = Acompanhamento, hora 1
- *fup.2* = Acompanhamento, hora 2
- *fup.3* = Acompanhamento, hora 3
- *fup.4* = Acompanhamento, hora 4
- *fup.5* = Acompanhamento, hora 5

A seguir, uma parte dos dados para visualização da composição do *data frame*

```
require("car")
head(iris, n=4L)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
```

Aplicações

Regressão Logística

A seguinte análise será realizada com o banco de dados **Mroz**. Será ajustado um modelo linear não generalizado através da função *glm* do **R**, onde os principais argumentos da função são:

- *formula* = a definição do modelo
- *family* = a distribuição assumida pela variável resposta com a função de ligação a ser usada

```r
m1 <- glm(lfp ~ ., family=binomial, data=Mroz)
summary(m1)
```

```
##
## Call:
## glm(formula = lfp ~ ., family = binomial, data = Mroz)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1062  -1.0900   0.5978   0.9709   2.1893
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.182140   0.644375   4.938 7.88e-07 ***
## k5          -1.462913   0.197001  -7.426 1.12e-13 ***
## k618        -0.064571   0.068001  -0.950 0.342337
## age         -0.062871   0.012783  -4.918 8.73e-07 ***
## wcyes        0.807274   0.229980   3.510 0.000448 ***
## hcyes        0.111734   0.206040   0.542 0.587618
## lwg          0.604693   0.150818   4.009 6.09e-05 ***
## inc         -0.034446   0.008208  -4.196 2.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1029.75  on 752  degrees of freedom
## Residual deviance:  905.27  on 745  degrees of freedom
## AIC: 921.27
##
## Number of Fisher Scoring iterations: 4
```

Como podemos observar as variáveis k618 e hcyes, são insignificantes para na contribuição da variável lfp. Ou seja,a participação da mulher no mercado de trabalho não depende do número de filhos de 6 a 18 anos e o do marido ter feito faculdade.

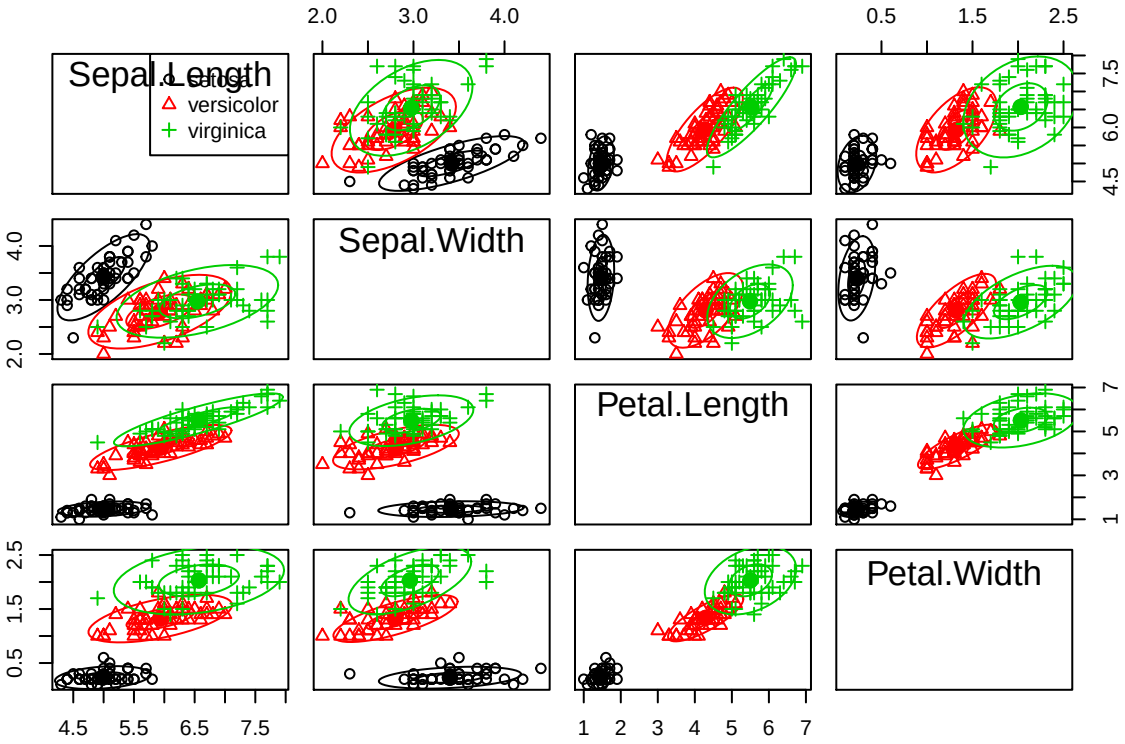
Regressão multivariada clássica

Os modelos lineares multivariados são ajustados no R com a função **lm**. O procedimento é simples: o lado esquerdo do modelo é a matriz de respostas, com cada coluna representando uma variável de resposta e cada linha uma observação. O lado direito do modelo e todos os outros argumentos para a função são os mesmos que para um modelo linear univariado.

A função **anova** é capaz de manipular os modelos lineares multivariados. Utilizaremos como exemplo os dados **Iris**

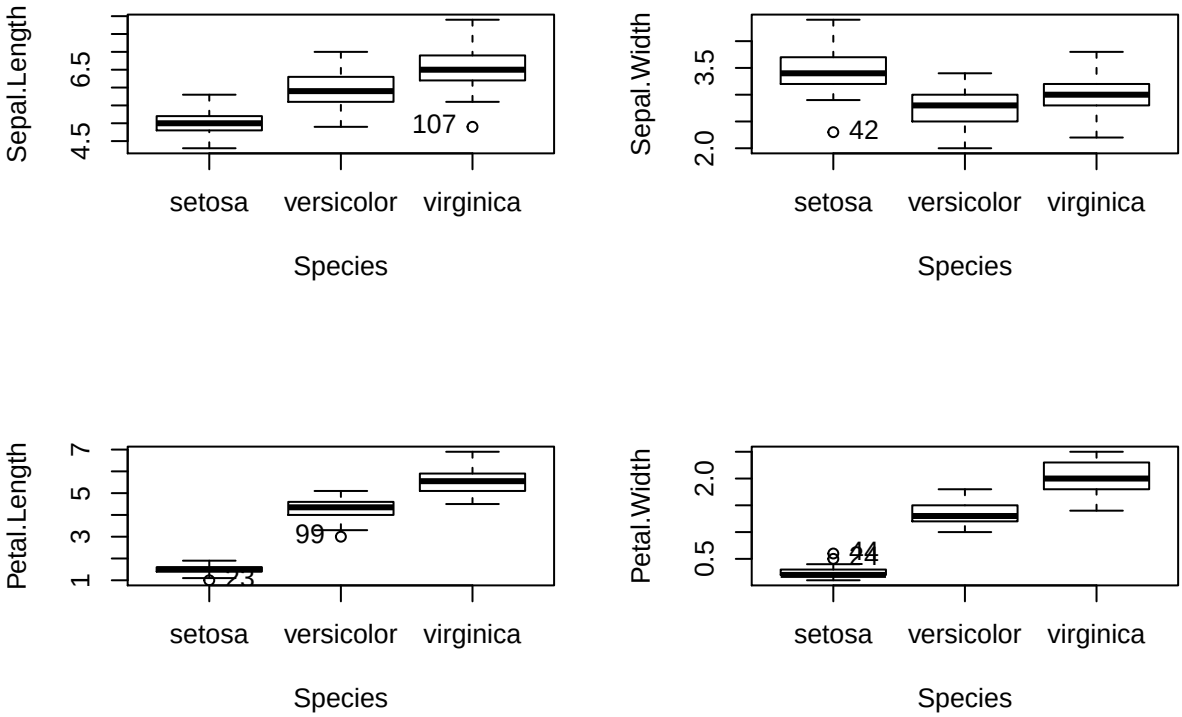
A figura a seguir mostram diagramas de dispersão das quatro medidas, mostrando elipses de confiaça de 50% e 95% dentro das espécies.

```
scatterplotMatrix(~ Sepal.Length + Sepal.Width + Petal.Length     + Petal.Width | Species, data=iris, sm
```



Boxplots para análise descritiva:

```
 par(mfrow=c(2, 2))
for (response in c("Sepal.Length", "Sepal.Width", "Petal.Length", "Petal.Width"))
Boxplot(iris[, response] ~ Species, data=iris, ylab=response)
```



Os gráficos apresentados indicam que *versicolor* e *virginica* são mais parecidos um com o outro do que com *setosa*. Além disso, as elipses de confiança sugerem que a suposição de matriz de covariâncias constantes dentro do grupo é problemática. Procedemos então para uma ANOVA para testar a hipótese nula: as médias das quatro respostas são idênticas nas três espécies de íris.

```
mod.iris <- lm(cbind(Sepal.Length, Sepal.Width, Petal.Length, Petal.Width)~ Species, data=iris)
anova(mod.iris)
```

```
## Analysis of Variance Table
##
##              Df  Pillai approx F num Df den Df    Pr(>F)
## (Intercept)   1 0.99313   5203.9      4    144 < 2.2e-16 ***
## Species       2 1.19190     53.5      8    290 < 2.2e-16 ***
## Residuals   147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A hipótese nula é claramente rejeitada.

Referências

FOX, J.; WEISBERG, S. **An r companion to applied regression**. SAGE Publications, 2010.

FOX, J.; WEISBERG, S.; ADLER, D.; et al. Package "car". **Companion to Applied Regression. R Package version**, p. 2–1, 2016.

LANDEIRO, V. L.; ECOLOGIA, C. DE P. EM. Introdução ao uso do programa r. **Manaus: Instituto Nacional de Pesquisas da Amazônia**, 2011.

LÊ, S.; JOSSE, J.; HUSSON, F.; OTHERS. FactoMineR: An r package for multivariate analysis. **Journal of statistical software**, v. 25, n. 1, p. 1–18, 2008. Foundation for Open Access Statistics.

R CORE TEAM. **R: A language and environment for statistical computing**. Vienna, Austria: R Foundation for Statistical Computing, 2015.