

Diagnóstico em Regressão

Posted by Alfredo Rossi on April 13, 2019

Diagnóstico em Regressão

Existem situações ao se fazer uma modelagem que o objeto de estudo é a interpretabilidade do modelo, impossibilitando modelos como Suport Vector Machine, Redes Neurais e outros. Para solucionar esse problema, os modelos estatísticos tradicionais como regressão linear e modelos lineares generalizados conseguem captar o efeito das variáveis explicativas sobre a variável resposta. Contudo, cada um desses modelos possui seus pressupostos específicos e para que se possa fazer inferência nos resultados, é necessário que os pressupostos estejam sendo respeitados. Esse post tem como objetivo ensinar técnicas que vão além de retirar valores discrepantes detectados pelo boxplot (acima (abaixo) de $1,5 \times 3^{\circ}(1^{\circ})$ quartil) para melhorar o ajuste de um modelo.

Existem métodos gráficos e testes estatísticos para diagnóstico. Esses métodos tentam identificar basicamente:

- Outliers, Valores influentes, Pontos de alavanca
- Normalidade (ou outra distribuição que seja considerada nos pressupostos)
- Homocedasticidade (variância constante)
- Outros pressupostos que variam conforme o modelo

Dividiremos o post nas seguintes etapas: (1) explicaremos as técnicas mais utilizadas, sua importância, interpretação e quando deve ser utilizada; (2) aplicabilidade em um problema de regressão, utilizando uma base real do próprio software R (Iris).

Diagnóstico

Métodos de diagnóstico são utilizados para que desvios entre as observações e os valores ajustados do modelo sejam analisados e verificados o seu grau de influência sobre a análise.

Essa diferença entre o real e o previsto podem surgir por vários motivos, pela função de variância, função de ligação, ausência ou não parâmetro de dispersão, parâmetro de inflação nos zeros, ou ainda pela definição errada da escala da variável ou mesmo porque algumas observações se mostram dependentes ou possuem correlação serial. Discrepâncias pontuais podem ocorrer porque as observações estão nos limites observáveis da variável, erros de digitação, algum fator não controlado (mas relevante) influenciou a sua obtenção ou até mesmo multicolinearidade entre variáveis (correlação alta entre variáveis).

As técnicas de diagnóstico podem ser formais ou informais. As informais baseiam-se em gráficos para detectar padrões visualmente e ver o comportamento dos dados. As formais envolvem especificar o modelo sob pesquisa em uma e verificar via testes, intervalos de confiança e outras medidas específicas para cada modelo, as mais usadas são baseadas nos testes da razão de verossimilhanças e escore.

Esse post terá como foco o modelo de regressão linear na apresentação de fórmulas por simplicidade, a extensão do conteúdo acontece de forma natural para modelos lineares generalizados e outras classes de modelo acrescentando os parâmetros pertinentes.

No modelo de regressão linear, que tem forma $Y = X\beta + e$, os elementos e_i do vetor e são as diferenças entre os valores observados y_i e aqueles esperados μ_i pelo modelo. Essa diferença é chamada de resíduo e considera-se como pressuposto do modelo que os resíduos sejam independentes e que tenham distribuição normal (a distribuição esperada pode mudar conforme o modelo), cabe ressaltar que a normalidade deve ser verificada nos resíduos e não na variável resposta Y.

Os resíduos indicam a variação natural dos dados, um fator aleatório (ou não) que o modelo não capturou. Se as pressuposições do modelo são violadas, a análise será levada a resultados duvidosos e não confiáveis para inferência. Essas falhas do modelo nos pressupostos podem ser oriundas de diversos fatores como não linearidade, não-normalidade, heterocedasticidade, não-independência e isso pode ser causado por pontos atípicos (observações discrepantes), que podem influenciar, ou não, no ajuste do modelo.

Depois de ajustado um modelo algumas medidas básicas de pressuposições sempre devem ser verificadas como

- valores estimados (ou ajustados) $\hat{\mu}_i$;

- Resíduos ordinários $r_i = y_i - \hat{\mu}_i$ (existem outros tipos de resíduos mas partindo da mesma premissa de diferença de valores);
- Variância residual estimada (ou quadrado médio residual)
 $\hat{\sigma}^2 = s^2 = \sum(y_i - \hat{\mu}_i)^2 / (n - p)$;
- Elementos da diagonal (Leverage) da matriz de projeção $H = X(X^T X)^{-1} X^T$

A diagonal principal da matriz de projeção $H = X(X^T X)^{-1} X^T$, em que X denota a matriz modelo, também conhecido como pontos de alavaca, que receberam esse nome por terem um peso desproporcional no próprio valor ajustado. Esses pontos em geral são remotos no subespaço gerado pelas colunas da matriz X, ou seja, têm um perfil diferente das demais observações no que diz respeito aos valores das variáveis explicativas, conforme o seu afastamento das outras observações esses pontos podem exercer forte influência nas estimativas dos coeficientes da regressão.

Agora serão apresentados algumas métricas usadas para diagnóstico de forma geral.

Resíduos

- Resíduos ordinários

Os resíduos por mínimos quadrados são definidos por $r_i = y_i - \hat{\mu}_i$. Pela definição as observações do vetor do termo de erro do modelo são independentes e têm a mesma variância, os resíduos (o erro observável) obtidos com ajuste do modelo tem a seguinte variância

$$Var(R) = Var[(I - H)Y] = \sigma^2(I - H).$$

Os resíduos ordinários podem não ser adequados devido à heterogeneidade das variâncias e falta de independência. Então, foram construídas padronizações nos resíduos para minimizar esse problema.

- Resíduos estudentizados internamente (Studentized residuals)

Um estimador não viesado para a variância dos resíduos é expresso por

$\hat{Var}(r_i) = (1 - h_{ii})s^2$ Como $E(r_i) = E(Y_i - \hat{\mu}_i) = 0$, o resíduo estudentizado internamente (ou seja, retirando a média e dividindo pelo desvio padrão) é igual a

$$\boxed{rsi_i = \frac{r_i}{\sqrt{(1 - h_{ii})}}}$$

Esses resíduos são mais sensíveis do que os anteriores por considerarem variâncias distintas. Ainda assim, um valor discrepante pode mudar drasticamente a variância residual dependendo do modo como se afasta da maioria das observações. Para corrigir esse problema é feita uma pequena alteração na fórmula, que será mostrada a seguir.

- Resíduos estudentizados externamente (jackknifed residuals, RStudent)

Define-se o resíduo estudentizado externamente, como

$$rse_{(i)} = \frac{r_i}{s_{(i)}} \sqrt{(1 - h_{ii})}$$

sendo $s_{(i)}^2$ o quadrado médio residual livre da influência da observação i , ou seja, é estimado a variância sem a observação i (por isso a ideia de Jackknife).

A vantagem de usar o resíduo $rse_{(i)}$, além de ser mais robusto, é que, sob normalidade, tem distribuição t de Student com $(n - p - 1)$ graus de liberdade.

Pontos discrepantes

Pontos atípicos são caracterizadas por terem h_{ii} e/ou resíduos grandes, serem inconsistentes e/ou influentes. Uma observação inconsistente é aquela que se afasta da tendência geral das demais. Quando uma observação está distante das outras em termos das variáveis explicativas, ela pode ser, ou não, influente. Uma observação influente é aquela cuja omissão do conjunto de dados resulta em mudanças substanciais nas estatísticas de diagnóstico do modelo. Essa observação pode ser um outlier (observação aberrante), ou não. Uma observação pode ser influente de 3 maneiras:

- Ajuste geral do modelo;
- Conjunto das estimativas dos parâmetros;
- Estimativa de um determinado parâmetro;

As estatísticas mais usadas para verificar pontos atípicos são:

- Leverage: h_{ii} (já comentado);
- $rse_{(i)}$ (já comentado);
- Influência sobre o parâmetro β_i : $DFBetaS(i)$ para β_i ;
- Influência geral: DFFitS(i), D(i).

Existem alguns critérios para classificação das observações como discrepantes, contudo cada autor possui o seu critério. De uma forma geral, pode-se classificar uma observação como:

- Inconsistente: ponto com $rse_{(i)} > t_{(1-\alpha)/(2n);n-p-1}$;
- com $h_{ii} > 2p/n$, no qual p é o número de parâmetros e n o tamanho da amostra. Pode ser classificado como bom, quando consistente, ou ruim, quando inconsistente;
- Outlier: ponto inconsistente com leverage pequeno, ou seja, com $rse_{(i)}$ grande e h_{ii} pequeno;
- Influente: ponto com DFFitS(i), C(i), D(i) ou DFBetaS(i) grande.

Essas medidas de influência são descritas abaixo.

- DFBeta e DFBetaS

São usados quando o coeficiente de regressão tem um significado prático. A estatística DFBeta(i) mede a alteração no vetor estimado $\hat{\beta}$ ao se retirar a i -ésima observação da análise, isto é,

$$DFBeta(i) = \hat{\beta} - \hat{\beta}_{(i)} = \frac{ri}{(1 - h_{ii})} (X^T X)^{-1} x_i$$

Portanto, é possível verificar o tamanho da influência nos parâmetros que a ausência de determinada observação pode causar

- DFFit e DFFitS

A estatística DFFit e sua versão estudentizada DFFitS medem a alteração no valor ajustado pela eliminação da observação i , é basicamente o DFBeta ajustado em x_i . São expressas como

$$DFFit(i) = x_i^T (\hat{\beta} - \hat{\beta}_{(i)}) = \hat{\mu}_i - \hat{\mu}_{(i)}$$

- Distância de Cook

É a mais utilizada, é uma medida de afastamento do vetor de estimativas resultante da eliminação da observação i . Tem uma expressão muito semelhante ao DFFitS mas que usa como estimativa da variância residual aquela obtida com todas as n observações, considera

o resíduo estudentizado internamente.

$$D_{(i)} = \frac{(\hat{\beta} - \hat{\beta}_{(i)})^T (X^T X) (\hat{\beta} - \hat{\beta}_{(i)})}{ps^2}$$

Multicolinearidade

Para finalizar, outro grande problema ocorrido nos modelos é a multicolinearidade no qual as variáveis independentes possuem relações lineares exatas ou aproximadamente exatas (correlação alta) entre elas mesmas, indicando que duas ou mais variáveis fornecem a mesma informação sobre a variável resposta. Um índice da existência da multicolinearidade é quando o R^2 é bastante alto (acima de 0,7), mas nenhum dos coeficientes da regressão é estatisticamente significativo segundo a estatística t convencional. A importância de se verificar a existência de multicolinearidade é que podem alterar as estimativas dos parâmetros e fornecer erros-padrão elevados no caso de multicolinearidade moderada ou alta.

- VIF

Além de verificar a alteração dos parâmetros e seu desvio padrão ao se retirar uma observação, existe uma medida chamada de VIF (fator de inflação da variância) que é utilizada para verificar a partir das correlações o impacto da variância em cada parâmetro. Conforme o VIF aumenta maior o índice de multicolinearidade, entretanto não existem um ponto de corte bem definido, alguns autores recomendam valores acima de 10 como limiar. O cálculo do VIF é dado pela seguinte fórmula

$$VIF_k = (1 - R_k^2)^{-1}$$

Banco de dados

O banco de dados utilizado será a base Iris que possui observações acerca de três espécies de plantas e é uma base muito utilizada e conhecida para análises iniciais por ser um banco pequeno, de fácil entendimento e ajuste do modelo.

O objetivo será modelar o comprimento da sépala usando a largura da pétala e sépala, espécie e comprimento da pétala. Depois do ajuste do modelo, será verificado cada pressuposto e se está de acordo com o esperado.

Análise de dados

Para a análise foi utilizado o banco de dados Iris utilizando o ambiente R de computação estatística.

A base de dados pode ser obtida no próprio repositório do R, é necessário apenas chamar o banco de dados. A descrição das variáveis está disponível no help do R.

Primeiramente, serão lidos os dados e depois ajustado o modelo de regressão linear.

```
library(car)
head(iris)
m <- lm(Sepal.Length~., data=iris)

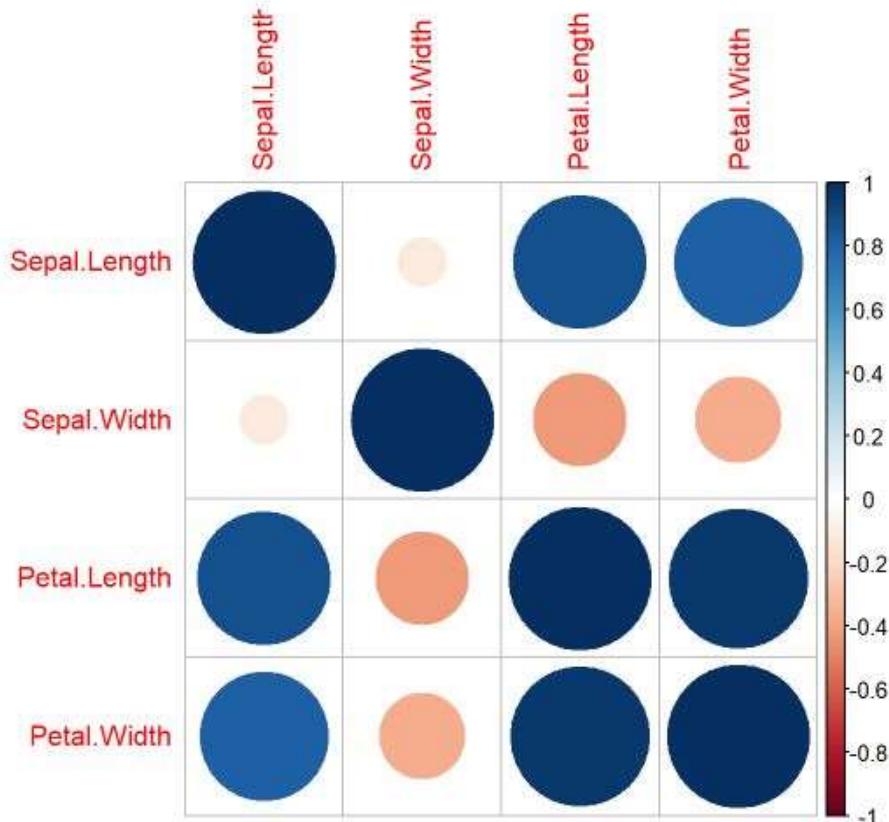
summary(m)

## 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
## 5      5.0       3.6      1.4       0.2   setosa
## 6      5.4       3.9      1.7       0.4   setosa

##
## Call:
## lm(formula = Sepal.Length ~ ., data = iris)
##
## Residuals:
##    Min     1Q   Median     3Q    Max 
## -0.79424 -0.21874  0.00899  0.20255  0.73103 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 2.17127   0.27979   7.760 1.43e-12 ***
## Sepal.Width  0.49589   0.08607   5.761 4.87e-08 ***
## Petal.Length 0.82924   0.06853  12.101 < 2e-16 ***
## Petal.Width  -0.31516   0.15120  -2.084 0.03889 *  
## Speciesversicolor -0.72356   0.24017  -3.013 0.00306 ** 
## Speciesvirginica -1.02350   0.33373  -3.067 0.00258 ** 
## ---
## Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 ' ' 1
##
## Residual standard error: 0.3068 on 144 degrees of freedom
## Multiple R-squared:  0.8673, Adjusted R-squared:  0.8627 
## F-statistic: 188.3 on 5 and 144 DF,  p-value: < 2.2e-16
```

Pelo ajuste do modelo, verificamos que as variáveis são significativas e que o mínimo e máximo dos resídos são inferiores a 1, já sendo um indício que as medidas usando resíduos studentizados devem apresentar valores baixos. Além disso, é necessário verificar a correlação entre as variáveis do modelo.

```
library(corrplot)
correlacao=cor(iris[,-5])
corrplot(correlacao)
```

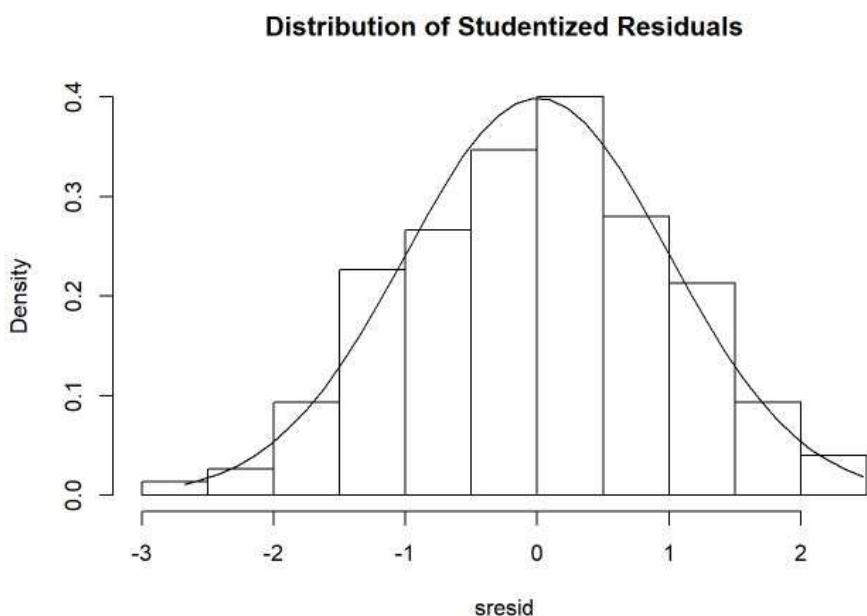
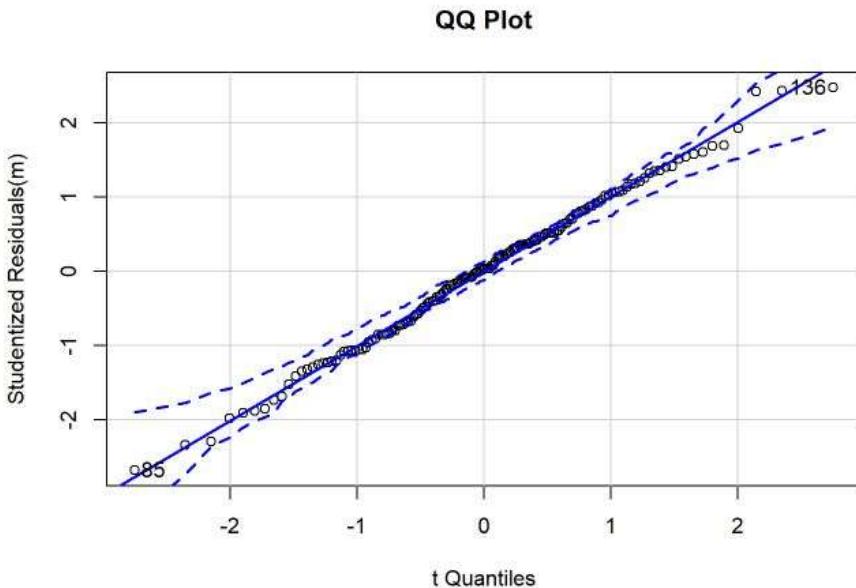


Por meio do gráfico de correlação entre as variáveis explicativas é possível verificar que existem algumas variáveis relacionadas e que podem causar multicolineariedade. Essa análise será feita mais adiante.

Primeiramente, será verificado a normalidade dos resíduos studentizados por meio do qqplot e histograma.

```
qqPlot(m, main="QQ Plot") #qq plot for studentized resid

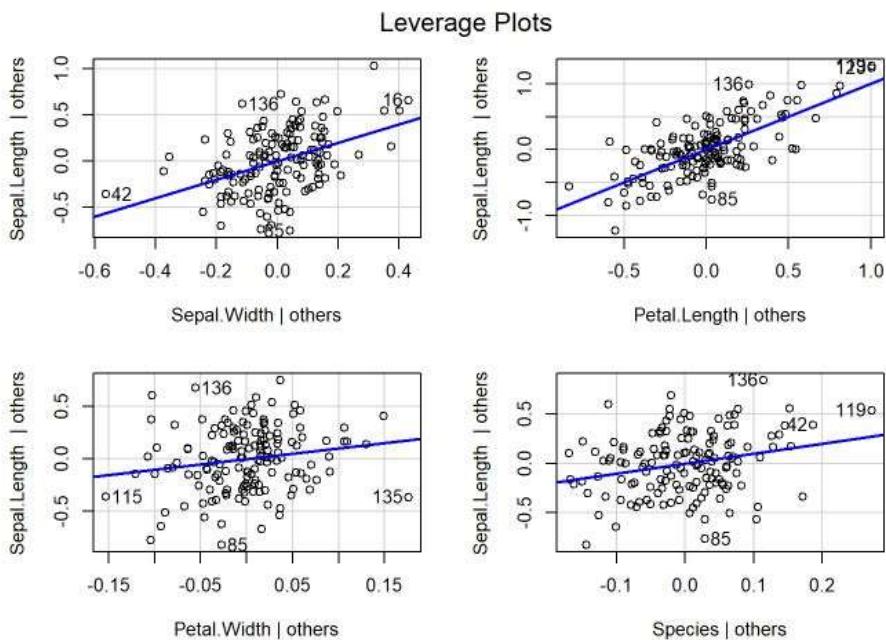
# distribution of studentized residuals
library(MASS)
sresid <- studres(m)
hist(sresid, freq=FALSE,
     main="Distribution of Studentized Residuals")
xm<-seq(min(sresid),max(sresid),length=40)
ym<-dnorm(xm)
lines(xm, ym)
```



Pelo QQplot, não encontramos indícios de que não existe normalidade nos resíduos porque o comportamento está bem linear e dentro das faixas. As faixas tracejadas são o intervalo de confiança enquanto que o eixo x é o quantil teórico enquanto que o eixo vertical é o empírico. Pelo histograma, é possível ver que os resíduos estão simétricos em torno de 0 e que uma distribuição normal está se ajustando bem aos dados.

O gráfico do leveragePlots é um gráfico de alavancagem para um efeito fixo que mostra o impacto da adição desse efeito ao modelo, dados os outros efeitos já existentes no modelo. Ou seja, são mostrados os resíduos da regressão usando determinada variável desconsiderando as demais, a reta azul é a regressão e quanto mais distante a observação maior a falta de ajuste.

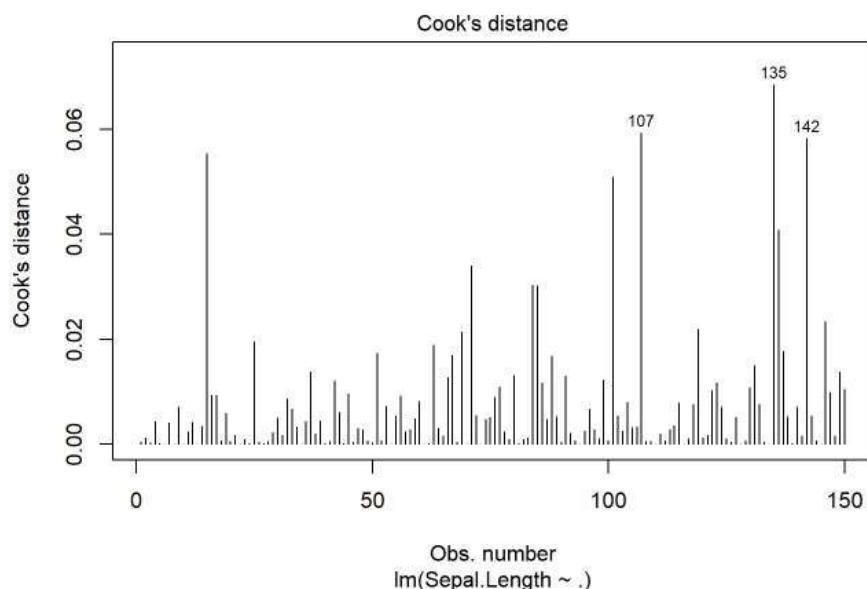
```
leveragePlots(m) # leverage plots
```



Pelos gráficos, é possível analisar que os pontos estão bem próximos da reta e que as distâncias praticamente não ultrapassam 1 nos resíduos, indicando que tem um leverage controlado.

Na sequenêcia, será feito o gráfico da distância de Cook

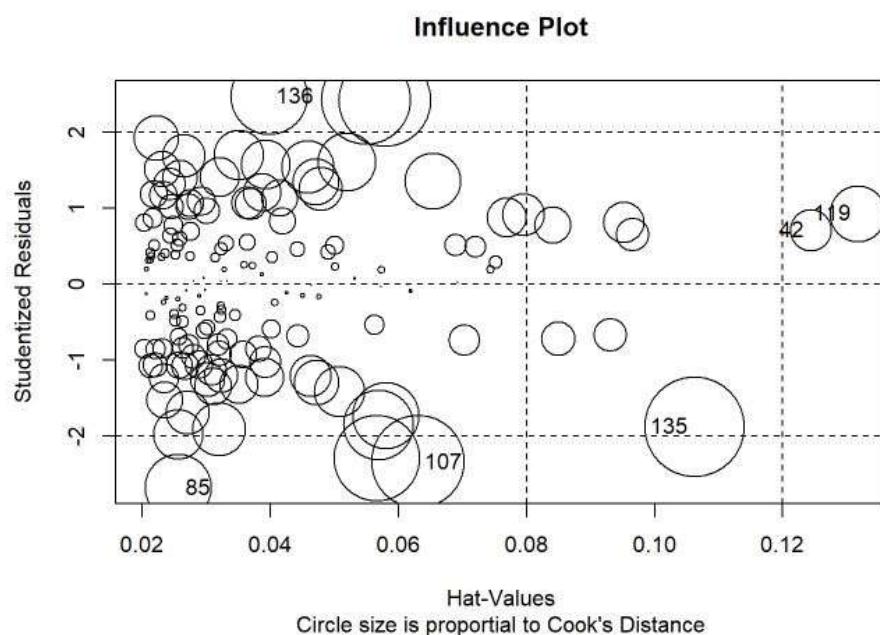
```
# Influential Observations
# Cook's D plot
# identify D values > 4/(n-k-1)
cutoff <- 4/((nrow(mtcars)-length(m$coefficients)-2))
plot(m, which=4, cook.levels=cutoff)
```



Pelo gráfico não existem indícios de alguma observação (número da observação no eixo horizontal) que esteja acima de 0,50 (considerando uma distância grande), pelo contrário, todas estão abaixo de 0,10.

Dado que não existem valores influêntes, o próximo gráfico verifica os pontos de leverage, outliers e discrepantes conjuntamente.

```
# Influence Plot  
influencePlot(m, id.method="identify", main="Influence Plot", sub="Circle size is proportional to Cook's Dist
```



As linhas acima de 2 e abaixo de -2 indicam os resíduos studentizados mais elevados enquanto que as linhas verticais indicam valores de leverage mais elevados e o tamanho do círculo é proporcional a distância de Cook. Esse gráfico faz a união de diversas medidas e é um ótimo indicativo para quais observações podem se destacar das demais. Nesse estudo, mesmo as observações 42,119 e 135 estarem um pouco afastadas, os valores absolutos de cada medida indicaram que não são observações influentes e que devem atrapalhar de forma geral o ajuste do modelo.

Com relação ao pressuposto da homocedasticidade (variância constante), foi realizado o teste de Breuch-Pagan e NCV que possuem como hipótese nula a homocedasticidade dos resíduos e se baseiam na verossimilhança. Ao realizar os testes, ambos falham em rejeitar a hipótese nula ao nível de 5%, ou seja, não existem indícios de que o pressuposto de homocedasticidade não é respeitado.

```
# Evaluate homoscedasticity  
lmtest::bptest(m)  
# non-constant error variance test  
ncvTest(m)
```

```

## 
## studentized Breusch-Pagan test
##
## data: m
## BP = 7.3844, df = 5, p-value = 0.1936

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 2.715134, Df = 1, p = 0.099401

```

Pela matriz de correlação foi verificado algumas relações entre as variáveis, para verificar o grau de significância disso foi calculado o VIF para cada variável. Pelos resultados, algumas variáveis como comprimento e largura da pétala possuem um valor de VIF que podem influenciar no ajuste, por isso seria interessante refazer o modelo retirando elas e refazer as medidas de diagnóstico para verificar se tem uma alteração brusca nos valores.

```

# Evaluate Collinearity
vif(m) # variance inflation factors
vif(m) > 10 # indício

```

```

##          GVIF Df GVIF^(1/(2*Df))
## Sepal.Width  2.227466  1      1.492470
## Petal.Length 23.161648  1      4.812655
## Petal.Width  21.021401  1      4.584910
## Species       40.039177  2      2.515482

##          GVIF   Df GVIF^(1/(2*Df))
## Sepal.Width FALSE FALSE           FALSE
## Petal.Length TRUE FALSE           FALSE
## Petal.Width  TRUE FALSE           FALSE
## Species      TRUE FALSE           FALSE

```

Para verificar independência dos resíduos foi ajustado o teste de Durbin Watson que possui hipótese nula de ausência de correlação serial e sua estatística do teste é construída somente com os resíduos. Ao nível de 5% de significância não encontramos indícios para rejeitar a hipótese de correlação, portanto é mais um pressuposto atendido.

```

# Test for Autocorrelated Errors
durbinWatsonTest(m)

```

```

## lag Autocorrelation D-W Statistic p-value
##  1      0.01099141    1.965705   0.708
## Alternative hypothesis: rho != 0

```

Para finalizar, a função GVLMA fornece estimativas do ajuste do modelo e também verifica automaticamente alguns pressupostos. Ela pode ser útil para uma validação rápida, contudo existem outros pressupostos que são necessários verificar conforme o modelo utilizado.

```
# Global test of model assumptions
library(gvlma)
gvmmodel <- gvlma(m)
summary(gvmmodel)
```

```
## Call:
## lm(formula = Sepal.Length ~ ., data = iris)
##
## Residuals:
##   Min     1Q Median     3Q    Max 
## -0.79424 -0.21874  0.00899  0.20255  0.73103 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 2.17127   0.27979  7.760 1.43e-12 ***
## Sepal.Width  0.49589   0.08607  5.761 4.87e-08 ***
## Petal.Length 0.82924   0.06853 12.101 < 2e-16 ***
## Petal.Width  -0.31516   0.15120 -2.084  0.03889 *  
## Speciesversicolor -0.72356   0.24017 -3.013  0.00306 ** 
## Speciesvirginica -1.02350   0.33373 -3.067  0.00258 ** 
## ---
## Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1
##
## Residual standard error: 0.3068 on 144 degrees of freedom
## Multiple R-squared:  0.8673, Adjusted R-squared:  0.8627 
## F-statistic: 188.3 on 5 and 144 DF,  p-value: < 2.2e-16
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance = 0.05
##
## Call:
## gvlma(x = m)
##
##          Value p-value      Decision  
## Global Stat 3.29923 0.5091 Assumptions acceptable.
## Skewness     0.08977 0.7645 Assumptions acceptable.
## Kurtosis     0.48196 0.4875 Assumptions acceptable.
## Link Function 0.89840 0.3432 Assumptions acceptable.
## Heteroscedasticity 1.82910 0.1762 Assumptions acceptable.
```

De forma geral os resultados de diagnóstico foram adequados e apenas a medida do VIF que deveria ser verificada com a retirada de uma variável do modelo.

As técnicas aqui expostas não esgotam o assunto e são apenas algumas das mais utilizadas, existem diversos outros testes e gráficos alternativos para se diagnosticar um modelo. A parte de diagnóstico é imprescindível para validação de um modelo que possua como objetivo a interpretabilidade e entendimento dos resultados.

Referências

Dobson, Annette J., 1945- An introduction to generalized linear models / Annette J. Dobson and Adrian G. Barnett. – 3rd ed.

Cordeiro, Gauss Moutinho; Demétrio, Clarice Garcia Borges; Moral, Rafael de Andrade.
Modelos Lineares Generalizados e Extensões. Piracicaba, setembro de 2016.

Gilberto A. Paula, MODELOS DE REGRESSÃO com apoio computacional.Instituto de
Matemática e Estatística Universidade de São Paulo.

← **PREVIOUS POST** (/2018/09/29/MLG/)

NEXT POST → (/2019/04/21/SORTING-ALGORITHMS/)

We were unable to load Disqus. If you are a moderator please see our [troubleshooting guide](#).



(/feed.xml)



(<https://www.facebook.com/lamfounb>)



(<https://github.com/lamfo-unb>)



(<mailto:lamfo@unb.br>)

Copyright © LAMFO - UNB 2021