Data Scientist - Introdução à análise de regressão com R

Soraia Pereira[a], Tiago Marques[a,b]

[a] CEAUL e FCUL, Universidade de Lisboa
[b] CREEM, University of St Andrews, e Dept de Biologia Animal, FCUL

FCUL, 11 de fevereiro de 2020

House Keeping

Recursos do curso disponíveis na pasta

https://tinyurl.com/CEAULGADESCursoRM3

O curso decorre entre as 18:30 e as 22:30

Entre as 20:30 e as 20:45 faremos uma pausa para café.

Introdução

A análise de regressão é uma ferramenta estatística que inclui técnicas de modelação e análise para estimação da relação entre uma variável dependente (a variável de interesse) e uma ou mais variáveis independentes.

Em geral, os objectivos de uma análise de regressão passam por

- Identificar quais as variáveis independentes, de entre um conjunto de variáveis, que estão mais relacionadas com a variável de interesse e compreender a forma dessa relação.
- Fazer predição do valor médio de uma variável de interesse dada a observação de um conjunto de variáveis independentes.

Modelo de regressão linear simples

Considere y a variável de interesse e x uma variável independente. O modelo de regressão linear simples assume a seguinte relação:

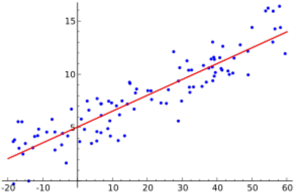
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, ..., n \tag{1}$$

onde β_0 e β_1 são parâmetros desconhecidos do modelo (a estimar) e ϵ_i é o termo residual.
Este modelo pressupõe:

- Relação linear entre a variável dependente e a variável independente
- ϵ_i com distribuição normal
- Inexistência de auto-correlação
- Homocedasticidade

Estimação pelo método dos mínimos quadrados

Os parâmetros do modelo são estimados a partir do método dos mínimos quadrados.
Seja $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, o valor estimado de y_i, e $e_i = y_i - \hat{y}_i$, o respectivo resíduo.
A ideia do método dos mínimos quadrados é determinar os valores de β_0 e β_1 que minimizam a soma dos quadrados dos resíduos, $\sum_{i=1}^{n} e_i^2$.



Uma vez definida a relação linear entre a variável de interesse e a variável independente, é possível utilizar essa relação para estimar o valor de y quando apenas o valor de x é conhecido.

Ilustração com os dados do package gapminder

Para ilustrar a aplicação de um modelo de regressão linear simples, voltamos ao exemplo dos dados do package gapminder do R. Deixamos o link de um vídeo interessante que tornou este conjunto de dados tão popular: https://www.youtube.com/watch?v=jbkSRLYSojo

Suponhamos agora que pretendemos compreender a relação entre o GDP (gdpPercap) e a esperança média de vida (lifeExp).

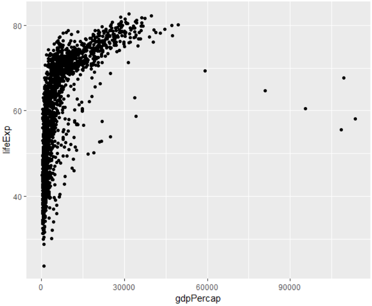
Visualização das primeiras linhas do conjunto de dados:

```
> head(gapminder)
# A tibble: 6 x 6
  country     continent  year lifeExp      pop gdpPercap
  <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
1 Afghanistan Asia       1952    28.8  8425333      779.
2 Afghanistan Asia       1957    30.3  9240934      821.
3 Afghanistan Asia       1962    32.0 10267083      853.
4 Afghanistan Asia       1967    34.0 11537966      836.
5 Afghanistan Asia       1972    36.1 13079460      740.
6 Afghanistan Asia       1977    38.4 14880372      786.
```

Gráfico de dispersão

Olhando para o gráfico de dispersão, a relação entre as duas variáveis não perece ser linear.

```
ggplot(data = gapminder, aes(x = gdpPercap, y = lifeExp))
+ geom_point()
```



Ajustamento do modelo linear simples

- No R, o ajustamento de um modelo linear pode ser feito facilmente com a função `lm`, indicando qual a variável dependente e quais as variáveis independentes (separadas por "$\sim$"):
`modelo<-lm(lifeExp ~ gdpPercap, data = gapminder)`

- A função `summary` aplicada ao modelo, mostra os resultados do ajustamento, nomeadamente as estimativas dos coeficientes do modelo e a percentagem de variabilidade de y que é explicada pelas variáveis independentes (R^2).

Ajustamento do modelo linear simples

Neste caso, a covariável GDP parece ser significativa na explicação da esperança média de vida (para nível de significância de 1%). No entanto, o ajustamento do modelo linear é fraco (apenas 34% da variabilidade de y é explicada por x).

```
> modelo<-lm(lifeExp ~ gdpPercap, data = gapminder)
> summary(modelo)

Call:
lm(formula = lifeExp ~ gdpPercap, data = gapminder)

Residuals:
    Min      1Q  Median      3Q     Max 
-82.754  -7.758   2.176   8.225  18.426 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5.396e+01  3.150e-01  171.29   <2e-16 ***
gdpPercap   7.649e-04  2.579e-05   29.66   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.49 on 1702 degrees of freedom
Multiple R-squared:  0.3407,	Adjusted R-squared:  0.3403 
F-statistic: 879.6 on 1 and 1702 DF,  p-value: < 2.2e-16
```

Modelo de regressão linear múltipla

Quando o modelo de regressão linear inclui mais do que uma variável independente é chamado modelo de regressão linear múltipla.

Suponhamos que estamos interessados em compreender a relação entre a variável de interesse esperança média de vida e as variáveis GDP, população e ano.

Para adicionar variáveis independentes ao modelo, basta utilizar "+", tal como no exemplo seguinte:

`modelo<-lm(lifeExp ~ gdpPercap + pop + year, data = gapminder)`

```
> modelo2<-lm(lifeExp ~ gdpPercap + pop + year, data = gapminder)
> summary(modelo2)

Call:
lm(formula = lifeExp ~ gdpPercap + pop + year, data = gapminder)

Residuals:
    Min      1Q  Median      3Q     Max 
-67.497  -7.075   1.121   7.701  19.640 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4.115e+02  2.767e+01 -14.872  < 2e-16 ***
gdpPercap    6.729e-04  2.444e-05  27.529  < 2e-16 ***
pop          6.353e-09  2.218e-09   2.864  0.00423 ** 
year         2.354e-01  1.400e-02  16.812  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.673 on 1700 degrees of freedom
Multiple R-squared:  0.4402,	Adjusted R-squared:  0.4392 
F-statistic: 445.6 on 3 and 1700 DF,  p-value: < 2.2e-16
```

Modelo de regressão linear múltipla

Suponhamos que se pretende perceber se a interação entre GDP e ano ajuda na explicação da variável de interesse.
As interações podem ser adicionadas utilizando "*" tal como na sintax seguinte:

```
modelo<-lm(lifeExp ~ gdpPercap*year + pop , data = gapminder)
```

```
> modelo3<-lm(lifeExp ~ gdpPercap*year + pop , data = gapminder)
> summary(modelo3)

Call:
lm(formula = lifeExp ~ gdpPercap * year + pop, data = gapminder)

Residuals:
    Min      1Q  Median      3Q     Max 
-54.328  -7.210   0.908   7.925  19.871 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -3.454e+02  3.271e+01 -10.559  < 2e-16 ***
gdpPercap      -8.856e-03  2.542e-03  -3.484 0.000506 ***
year            2.019e-01  1.655e-02  12.201  < 2e-16 ***
pop             6.469e-09  2.210e-09   2.928 0.003460 ** 
gdpPercap:year  4.807e-06  1.282e-06   3.749 0.000183 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.636 on 1699 degrees of freedom
Multiple R-squared:  0.4448,	Adjusted R-squared:  0.4435 
F-statistic: 340.3 on 4 and 1699 DF,  p-value: < 2.2e-16
```

Modelo de regressão linear múltipla

As variáveis categóricas são incluídas da mesma forma no preditor linear.
Note-se que a sua estrutura deve ser `factor`.
Exemplo: adicionar variável "continent":

```
modelo<-lm(lifeExp ~ gdpPercap + pop + year + continent ,
data = gapminder)
```

```
> modelo4<-lm(lifeExp ~ gdpPercap + pop +year + continent, data = gapminder)
> summary(modelo4)

Call:
lm(formula = lifeExp ~ gdpPercap + pop + year + continent, data = gapminder)

Residuals:
     Min       1Q   Median       3Q      Max
-28.4051  -4.0550   0.2317   4.5073  20.0217

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -5.185e+02  1.989e+01 -26.062   <2e-16 ***
gdpPercap          2.985e-04  2.002e-05  14.908   <2e-16 ***
pop                1.791e-09  1.634e-09   1.096    0.273
year               2.863e-01  1.006e-02  28.469   <2e-16 ***
continentAmericas  1.429e+01  4.946e-01  28.898   <2e-16 ***
continentAsia      9.375e+00  4.719e-01  19.869   <2e-16 ***
continentEurope    1.936e+01  5.182e-01  37.361   <2e-16 ***
continentOceania   2.056e+01  1.469e+00  13.995   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.883 on 1696 degrees of freedom
Multiple R-squared:  0.7172,	Adjusted R-squared:  0.716
F-statistic: 614.5 on 7 and 1696 DF,  p-value: < 2.2e-16
```

Análise dos resíduos

A análise dos resíduos é muito importante para verificar os pressupostos de um modelo de regressão linear.

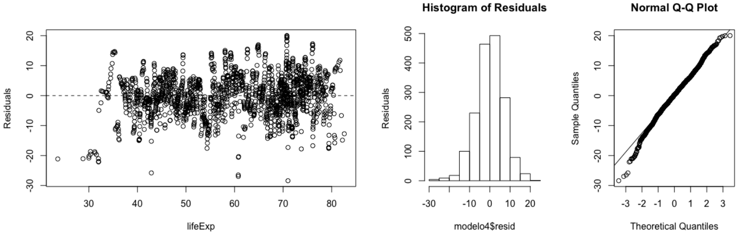
Os resíduos devem respeitar as seguintes condições:

- Seguem uma distribuição normal.
- Têm média zero.
- Têm variância constante.
- São independentes.

Análise dos resíduos

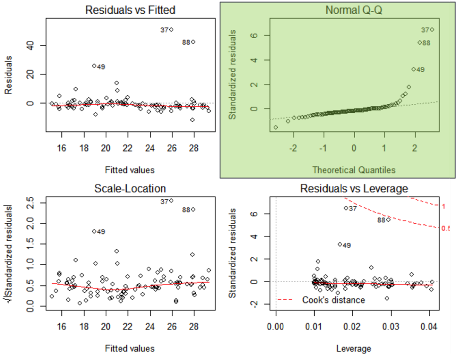
A visualização gráfica é uma ferramenta fundamental na análise aos resíduos. Uma forma de a fazer é através de `plot(modelo4)`. Alternativamente, podemos utilizar a seguinte sintax:

```
> plot(modelo4$resid~gapminder$lifeExp[order(gapminder$lifeExp)],
+      main="",
+      xlab="lifeExp", ylab="Residuals")
> abline(h=0,lty=2)
> hist(modelo4$resid, main="Histogram of Residuals",
+      ylab="Residuals")
> qqnorm(modelo4$resid)
> qqline(modelo4$resid)
```



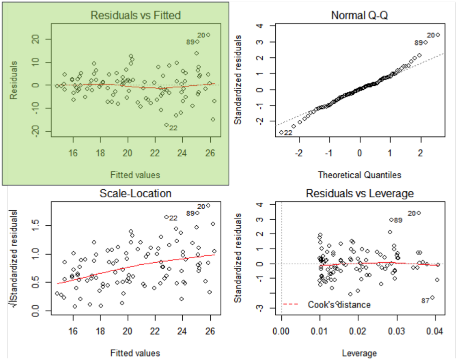
Exemplos de resíduos problemáticos

Não Normalidade dos resíduos



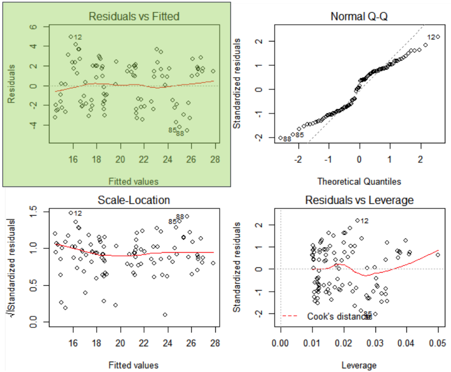
Exemplos de resíduos problemáticos

Heterocedasticidade dos resíduos



Exemplos de resíduos problemáticos

Não independência dos resíduos



Modelo de regressão linear generalizado

- O modelo de regressão linear supõe que a variável de interesse segue uma distribuição normal, mas muitas vezes não é o caso em problemas da vida real.
- Os modelos lineares generalizados são uma classe de modelos que permite lidar com variáveis de interesse que seguem uma distribuição pertencente à família exponencial. São exemplos de distribuições pertencentes a esta família a Poisson, Binomial, Gamma e Normal.
- Nesta classe de modelos, a relação entre o valor médio da variável resposta e o preditor linear não é necessariamente a identidade. Esta relação é definida através da chamada função de ligação, e a sua forma depende da distribuição da variável de interesse.

Ilustração com os dados do package `titanic`

Considere-se o conjunto de dados acerca da sobrevivência dos passageiros do Titanic, disponível no package `titanic` do R.
Das variáveis disponíveis, iremos focar-nos na sobrevivência (0-não sobreviveu, 1-sobreviveu), classe do passageiro, sexo e idade.
O objetivo é identificar quais as variáveis que melhor explicam a sobrevivência dos passageiros.

```
> library(titanic)
> data("titanic_train")
> names(titanic_train)
 [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"         "Age"
 [7] "SibSp"       "Parch"       "Ticket"      "Fare"        "Cabin"       "Embarked"
> titanic_data<-titanic_train[,c(2,3,5,6)]
> head(titanic_data)
  Survived Pclass    Sex Age
1        0      3   male  22
2        1      1 female  38
3        1      3 female  26
4        1      1 female  35
5        0      3   male  35
6        0      3   male  NA
```

Ilustração com os dados do package `titanic`

Note-se que a variável de interesse é binária (toma apenas valor 0 e 1), e portanto um modelo de regressão linear não é certamente adequado para o ajustamento.
Neste caso, o modelo indicado é o modelo de regressão logística, um caso particular dos modelos lineares generalizados.

```
> modelo<-glm(Survived ~ Pclass + Sex + Age, family=binomial(link="logit"), data=titanic_data)
> summary(modelo)

Call:
glm(formula = Survived ~ Pclass + Sex + Age, family = binomial(link = "logit"),
    data = titanic_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7270  -0.6799  -0.3947   0.6483   2.4668

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.056006   0.502128  10.069  < 2e-16 ***
Pclass      -1.288545   0.139259  -9.253  < 2e-16 ***
Sexmale     -2.522131   0.207283 -12.168  < 2e-16 ***
Age         -0.036929   0.007628  -4.841 1.29e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 964.52  on 713  degrees of freedom
Residual deviance: 647.29  on 710  degrees of freedom
  (177 observations deleted due to missingness)
AIC: 655.29

Number of Fisher Scoring iterations: 5
```