Tópico 1: Modelos de Regressão
Trabalho realizado pelas alunas Marta Barros e Marta Lobo
Pós-Graduação em Data Science & Business Intelligence

Descrição do objetivo da análise

O *dataset* escolhido para este trabalho, cujo enfoque são os modelos de regressão foi **Concrete Slump Test Data Set** disponível no arquivo UCI de Machine Learning, cedido por I-Cheng Yeh.

O concreto é um material altamente complexo, o que torna a modelação do seu comportamento uma tarefa difícil.

Assim, o objetivo da análise é perceber de que forma as variáveis de *output* SLUMP (cm), FLOW (cm) e 28-day Compressive Strength (Mpa) são afetadas pelas variáveis de *input*, neste caso, tratando-se de diferentes constituintes do concreto como cimento, escória, cinzas volantes, água, SP e agregados de construção civil em distintas quantidades por m^3 de concreto.

Descrição dos dados

Existe um total de 103 dados, sem *missing data*. Todas as variáveis deste dataset enquadram-se na categoria númerica (**Figura 1**). Estas incluem 7 variáveis de *input (*N°, Cement, Slag, Fly Ash, Water, SP, Coarse Aggr., Fine Aggr.) e 3 variáveis de *output* (Slump (cm), Flow(cm) e Compressive Strength (28-day)(Mpa)).

```
> str(slump_test_1_)
Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame':        103 obs. of  11 variables:
 $ no        : num  1 2 3 4 5 6 7 8 9 10 ...
 $ cement    : num  273 163 162 162 154 147 152 145 152 304 ...
 $ slag      : num  82 149 148 148 112 89 139 0 0 0 ...
 $ fly_ash   : num  105 191 191 190 144 115 178 227 237 140 ...
 $ water     : num  210 180 179 179 220 202 168 240 204 214 ...
 $ sp        : num  9 12 16 19 10 9 18 6 6 6 ...
 $ coar_agg  : num  904 843 840 838 923 860 944 750 785 895 ...
 $ fine_agg  : num  680 746 743 741 658 829 695 853 892 722 ...
 $ slump     : num  23 0 1 3 20 23 0 14.5 15.5 19 ...
 $ flow      : num  62 20 20 21.5 64 55 20 58.5 51 51 ...
 $ compr_stre: num  35 41.1 41.8 42.1 26.8 ...
 - attr(*, "spec")=List of 3
  ..$ cols   :List of 11
  .. ..$ No                                 : list()
  .. .. ..- attr(*, "class")= chr [1:2] "collector_double" "collector"
  .. ..$ Cement                             : list()
  .. .. ..- attr(*, "class")= chr [1:2] "collector_double" "collector"
  .. ..$ Slag                               : list()
  .. .. ..- attr(*, "class")= chr [1:2] "collector_double" "collector"
  .. ..$ Fly ash                            : list()
  .. .. ..- attr(*, "class")= chr [1:2] "collector_double" "collector"
  .. ..$ Water                              : list()
  .. .. ..- attr(*, "class")= chr [1:2] "collector_double" "collector"
  .. ..$ SP                                 : list()
  .. .. ..- attr(*, "class")= chr [1:2] "collector_double" "collector"
  .. ..$ Coarse Aggr.                       : list()
  .. .. ..- attr(*, "class")= chr [1:2] "collector_double" "collector"
  .. ..$ Fine Aggr.                         : list()
  .. .. ..- attr(*, "class")= chr [1:2] "collector_double" "collector"

 .. ..$ SLUMP(cm)                          : list()
 .. .. ..- attr(*, "class")= chr [1:2] "collector_double" "collector"
 .. ..$ FLOW(cm)                           : list()
 .. .. ..- attr(*, "class")= chr [1:2] "collector_double" "collector"
 .. ..$ Compressive Strength (28-day)(Mpa): list()
 .. .. ..- attr(*, "class")= chr [1:2] "collector_double" "collector"
 ..$ default: list()
 .. ..- attr(*, "class")= chr [1:2] "collector_guess" "collector"
 ..$ skip   : int 1
 ..- attr(*, "class")= chr "col_spec"
```

Figura 1. Descrição do tipo de variáveis.

As variáveis foram renomeadas para simplificar a análise e escrita das linhas de código. Foi também removida a variável n_o sem perda de informação, sendo que esta apenas descreve o número da observação.

```
> colnames(slump_test_1_)<-c('no','cement','slag','fly_ash','water','sp','coar_agg','fine_agg','slump','flow','compr_stre')
>
>
> head(slump_test_1_)
  no cement slag fly_ash water sp coar_agg fine_agg slump flow compr_stre
1  1    273   82     105   210  9      904      680    23 62.0      34.99
2  2    163  149     191   180 12      843      746     0 20.0      41.14
3  3    162  148     191   179 16      840      743     1 20.0      41.81
4  4    162  148     190   179 19      838      741     3 21.5      42.08
5  5    154  112     144   220 10      923      658    20 64.0      26.82
6  6    147   89     115   202  9      860      829    23 55.0      25.21

> slump_test_1_$no <- NULL
> head(slump_test_1_)
  cement slag fly_ash water sp coar_agg fine_agg slump flow compr_stre
1    273   82     105   210  9      904      680    23 62.0      34.99
2    163  149     191   180 12      843      746     0 20.0      41.14
3    162  148     191   179 16      840      743     1 20.0      41.81
4    162  148     190   179 19      838      741     3 21.5      42.08
5    154  112     144   220 10      923      658    20 64.0      26.82
6    147   89     115   202  9      860      829    23 55.0      25.21
```

Figura 2. Nova designação das variáveis e remoção da primeira coluna.

De seguida, obtivemos a estatística descritiva de cada uma das variáveis, nomeadamente, o mínimo, o máximo, a média, a mediana, o primeiro e terceiro quartil. Para visualizar estas informações mais rapidamente, efetuou-se boxplot.

```
> summary(slump_test_1_)
     cement           slag          fly_ash          water
 Min.   :137.0   Min.   :  0.00   Min.   :  0.0   Min.   :160.0
 1st Qu.:152.0   1st Qu.:  0.05   1st Qu.:115.5   1st Qu.:180.0
 Median :248.0   Median :100.00   Median :164.0   Median :196.0
 Mean   :229.9   Mean   : 77.97   Mean   :149.0   Mean   :197.2
 3rd Qu.:303.9   3rd Qu.:125.00   3rd Qu.:235.9   3rd Qu.:209.5
 Max.   :374.0   Max.   :193.00   Max.   :260.0   Max.   :240.0
       sp            coar_agg        fine_agg         slump
 Min.   : 4.40   Min.   : 708.0   Min.   :640.6   Min.   : 0.00
 1st Qu.: 6.00   1st Qu.: 819.5   1st Qu.:684.5   1st Qu.:14.50
 Median : 8.00   Median : 879.0   Median :742.7   Median :21.50
 Mean   : 8.54   Mean   : 884.0   Mean   :739.6   Mean   :18.05
 3rd Qu.:10.00   3rd Qu.: 952.8   3rd Qu.:788.0   3rd Qu.:24.00
 Max.   :19.00   Max.   :1049.9   Max.   :902.0   Max.   :29.00
      flow         compr_stre
 Min.   :20.00   Min.   :17.19
 1st Qu.:38.50   1st Qu.:30.90
 Median :54.00   Median :35.52
 Mean   :49.61   Mean   :36.04
 3rd Qu.:63.75   3rd Qu.:41.20
 Max.   :78.00   Max.   :58.53
```

Figura 3. Tabela-resumo da estatistica descritiva das dez variaveis.

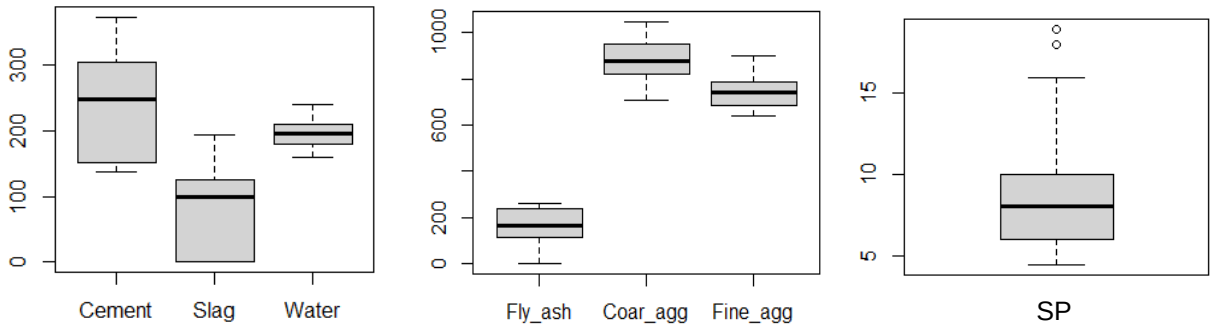

Figura 4. Boxplot representativo da distribuição, simetria e outliers de cada uma das variáveis x_n.

É possível observar simetria na representação gráfica da variável *water* (**Figura 4**), sendo que o valor da média (197.2) e da mediana (196.0) estão bastante próximos. Podemos inferir que os dados desta variável têm provavelmente um comportamento uniforme.

A variável *coar_agg* apesar de ter um pequeno enviesamento considera-se que também tem um comportamento bastante aproximado a uma distribuição uniforme.

No boxplot da variável *fly_ash* e da *fine_aggr* é observada pouca variância dos valores, no entanto na *fly_ash* existe enviesamento à esquerda ou seja, existem mais dados próximos do valor mínimo do que do máximo. Já na variável *fine_aggr* acontece o oposto, a distribuição dos dados está enviesada à esquerda.

No caso da variável *slag*, esta não é simétrica e concluimos que o mínimo está incluso na distribuição, verificando também que existe enviesamento à direita.

Por fim, na variável SP apesar de não existir enviesamento, verifica-se que existe uma grande variância dos valores com maior tendência para o máximo, com outliers visiveis depois deste ponto.

Remoção de outliers

Com base nestas informações, optamos pela construção do modelo inicial de regressão linear múltipla, sem remoção de outliers. Apenas serão retirados numa segunda abordagem ao modelo em que também iremos remover variáveis que se mostrem menos significativas.

Descrição do procedimento de avaliação

Iremos prever o comportamento das variáveis contínuas Slump (cm) (**1.1**), Flow (cm)(**1.2**) e Compressive Strength (28-day)(Mpa)(**1.3**) com base nas variáveis de input (X_1,...,X_n), segundo o modelo de regressão linear múltipla.

O modelo de regressão linear múltipla define-se segundo a seguinte equação:

$$Y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Y - Variável dependente
X_i - Variavéis independentes
β_0 - Valor da constante (Intercept)
β_i - Coeficientes

Modelos lineares de regressão múltipla

1ª Versão

Para os efeitos desta versão, optamos por usar a mesma base de dados de teste e treino, pelo que 100% da variável de resposta (Y) em cada um dos modelos lineares é explicada pelas variáveis x_n.

1.1. Modelo linear para Y ~ Slump

```
> modeloslump <- lm(slump ~ cement+slag+fly_ash+water+sp+coar_agg+fine_agg, data=slump_test_1_)
>  modeloslump

Call:
lm(formula = slump ~ cement + slag + fly_ash + water + sp + coar_agg +
    fine_agg, data = slump_test_1_)

Coefficients:
(Intercept)       cement         slag      fly_ash        water           sp
 -88.525037     0.010216    -0.012966     0.006176     0.258982    -0.183954
   coar_agg     fine_agg
   0.029737     0.038584
```

$$Y \approx -1{,}048 \times 10^{2} + 6{,}276 \times 10^{-3} x_1 - 3{,}126 \times 10^{-3} x_2 + 8.701 \times 10^{-3} x_3 + 2.782 \times 10^{-1} x_4 + 1.962 \times 10^{-1} x_5 + 3.767 \times 10^{-2} x_6 + 4.172 \times 10^{-2} x_7 \ (cm)$$

1.2. Modelo linear para Y ~ Flow

```
> modeloflow<-lm(flow ~ cement+slag+fly_ash+water+sp+coar_agg+fine_agg, data=slump_test_1_)
> modeloflow

Call:
lm(formula = flow ~ cement + slag + fly_ash + water + sp + coar_agg +
    fine_agg, data = slump_test_1_)

Coefficients:
(Intercept)       cement         slag      fly_ash        water           sp
 -252.87467      0.05364     -0.00569      0.06115      0.73180      0.29833
   coar_agg     fine_agg
    0.07366      0.09402
```

$$Y \approx -252{,}875 + 0{,}054 x_1 - 0{,}006 x_2 + 0{,}061 x_3 + 0{,}732 x_4 + 0{,}298 x_5 + 0{,}074 x_6 + 0{,}094 x_7 \ (cm)$$

1.3. Modelo linear para Y ~ Compressive Strength

```
> modelocompressive<-lm(compr_stre ~ cement+slag+fly_ash+water+sp+coar_agg+fine_agg, data=slump_test_1_)
> modelocompressive

Call:
lm(formula = compr_stre ~ cement + slag + fly_ash + water + sp +
    coar_agg + fine_agg, data = slump_test_1_)

Coefficients:
(Intercept)       cement         slag      fly_ash        water           sp
  139.78150      0.06141     -0.02971      0.05053     -0.23270      0.10315
   coar_agg     fine_agg
   -0.05562     -0.03908
```

$$\mathrm{Y} \approx 125.643 + 0.066x_1 - 0.020x_2 + 0{,}055x_3 - 0{,}213x_4 + 0{,}076x_5 + 0{,}051x_6 - 0.034x_7\ (Mpa)$$

Validação dos modelos

Para validar a normalidade dos dados em cada um dos modelos, testamos cinco pressupostos, nomeadamente, linearidade das relações X-Y, normalidade dos resíduos, independência entre os erros (inexistência de autocorrelação) e variância constante dos erros (homocedasticidade).

Linearidade nas relações X-Y

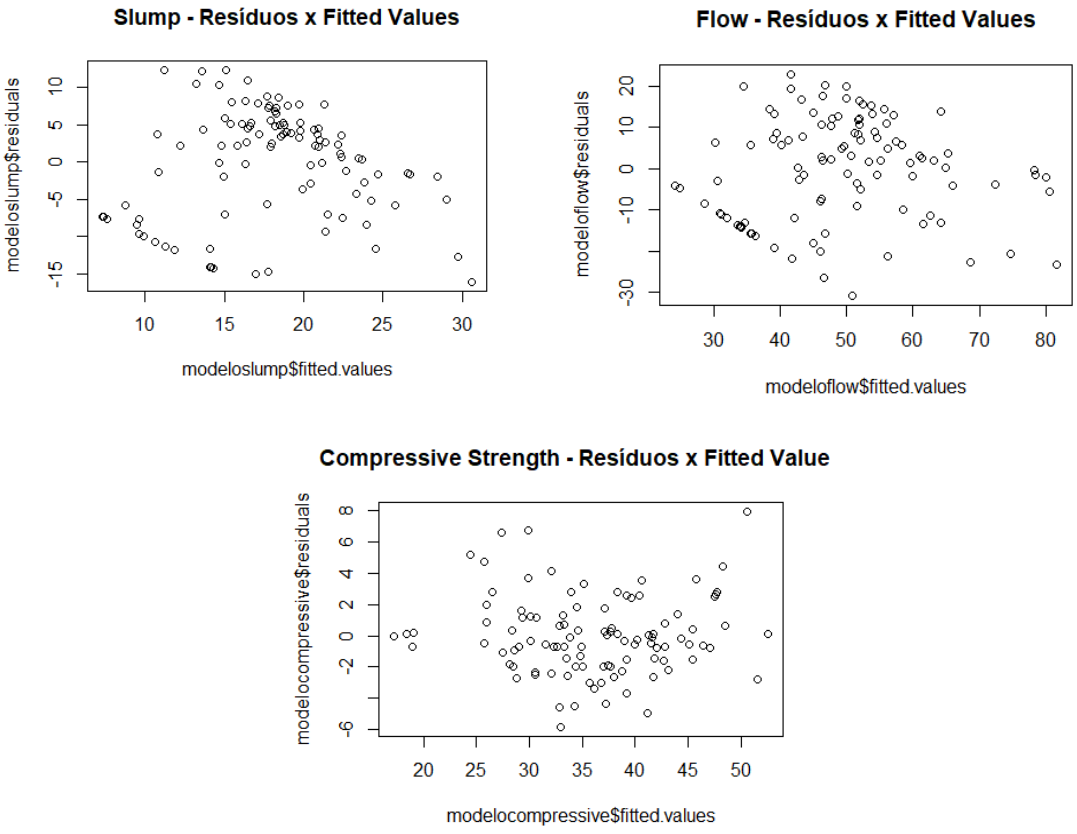

Figura 5. Análise gráfica dos resíduos de três modelos lineares: *modeloslump*, *modeloflow*, *modelocompressive*.

Através da **Figura 5**, podemos observar o desvio entre as observações e os valores ajustados do modelo. A diferença entre o real e o previsto, isto é, análise residual, pode ser analisada através da dispersão dos dados acima e abaixo da linha Y=0.

Concluimos que não existem padrões na curva de erros para os três modelos criados.

Normalidade dos resíduos

```
> ks.test(modeloslump$residuals, "pnorm", 0, sd(modeloslump$residuals))

        One-sample Kolmogorov-Smirnov test

data:  modeloslump$residuals
D = 0.14504, p-value = 0.02623
alternative hypothesis: two-sided

> ks.test(modeloflow$residuals, "pnorm", 0, sd(modeloflow$residuals))

        One-sample Kolmogorov-Smirnov test

data:  modeloflow$residuals
D = 0.069786, p-value = 0.6975
alternative hypothesis: two-sided

> ks.test(modelocompressive$residuals, "pnorm", 0, sd(modelocompressive$residuals))

        One-sample Kolmogorov-Smirnov test

data:  modelocompressive$residuals
D = 0.10086, p-value = 0.2456
alternative hypothesis: two-sided
```

Figura 6. Teste de Kolmogorov-Smirnov aplicado aos três modelos.

Efetuou-se o teste Kolmogorov-Smirnov, onde se obteve um p-value que nos permite aceitar a hipótese nula (H_0) da inexistência de diferença entre a distribuição dos resíduos e a distribuição normal com média 0 no caso do *modelflow* e *modelcompressive*.

No entanto, no que se refere ao *modeloslump*, obtivemos um p-value de 0.026, que não nos permite aceitar H_0, logo é violado o princípio de normalidade dos resíduos.

QQ plot

```
> plot(modeloflow, 2)
> plot(modelocompressive, 2)
```

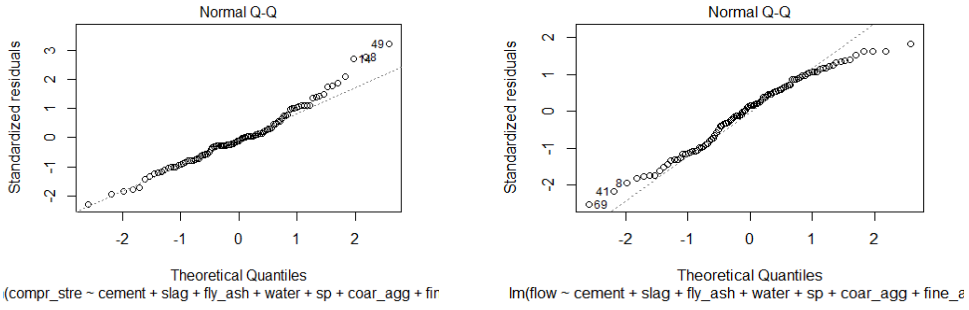

Figura 7. QQ plot do *modeloflow* e *modelocompressive*, que respeitam a normalidade dos resíduos estandarizados.

Para assumir a normalidade da distribuição dos resíduos, os valores devem alinhar-se na reta de previsão com os quantis teóricos. Deste modo, verificamos uma boa correspondência dos dados entre -1 e 1.

Independência entre os erros

Para os três modelos, verifica-se um p-value que permite aceitar a hipótese nula, pelo que se pode afirmar uma variância constante nos erros (homocedasticidade).

```
> Box.test(modeloslump$residuals, type="Ljung")

        Box-Ljung test

data:  modeloslump$residuals
X-squared = 0.86205, df = 1, p-value = 0.3532

> Box.test(modeloflow$residuals, type="Ljung")

        Box-Ljung test

data:  modeloflow$residuals
X-squared = 0.016567, df = 1, p-value = 0.8976

> Box.test(modelocompressive$residuals, type="Ljung")

        Box-Ljung test

data:  modelocompressive$residuals
X-squared = 1.7506, df = 1, p-value = 0.1858
```

Figura 8. Box-test aplicado aos três modelos.

Variância constante nos erros

Verifica-se a inexistência de autocorrelação (dados independentes) pela aceitação da hipótese nula em todos os casos.

```
> lmtest::bptest(modeloslump)

        studentized Breusch-Pagan test

data:  modeloslump
BP = 10.009, df = 7, p-value = 0.1881

> lmtest::bptest(modeloflow)

        studentized Breusch-Pagan test

data:  modeloflow
BP = 12.192, df = 7, p-value = 0.09443

> lmtest::bptest(modelocompressive)

        studentized Breusch-Pagan test

data:  modelocompressive
BP = 8.791, df = 7, p-value = 0.268
```

Figura 9. Teste de Breusch-Pagan aplicado aos três modelos.

Validação estatística dos coeficientes

1.1. Slump (cm)

```
> summary(modeloslump)

Call:
lm(formula = slump ~ cement + slag + fly_ash + water + sp + coar_agg +
    fine_agg, data = slump_test_1_)

Residuals:
    Min      1Q  Median      3Q     Max
-16.125  -5.726   2.184   5.064  12.380

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -88.525037 203.303168  -0.435    0.664
cement        0.010216   0.065256   0.157    0.876
slag         -0.012966   0.090819  -0.143    0.887
fly_ash       0.006176   0.066216   0.093    0.926
water         0.258982   0.204900   1.264    0.209
sp           -0.183954   0.384827  -0.478    0.634
coar_agg      0.029737   0.078458   0.379    0.706
fine_agg      0.038584   0.082415   0.468    0.641

Residual standard error: 7.459 on 95 degrees of freedom
Multiple R-squared:  0.3233,    Adjusted R-squared:  0.2734
F-statistic: 6.484 on 7 and 95 DF,  p-value: 2.98e-06
```

1.2. Flow (cm)

```
> summary(modeloflow)

Call:
lm(formula = flow ~ cement + slag + fly_ash + water + sp + coar_agg +
    fine_agg, data = slump_test_1_)

Residuals:
    Min      1Q  Median      3Q     Max
-30.880 -10.428   1.815   9.601  22.953

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -252.87467  350.06649  -0.722   0.4718
cement         0.05364    0.11236   0.477   0.6342
slag          -0.00569    0.15638  -0.036   0.9710
fly_ash        0.06115    0.11402   0.536   0.5930
water          0.73180    0.35282   2.074   0.0408 *
sp             0.29833    0.66263   0.450   0.6536
coar_agg       0.07366    0.13510   0.545   0.5869
fine_agg       0.09402    0.14191   0.663   0.5092
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.84 on 95 degrees of freedom
Multiple R-squared:  0.5022,    Adjusted R-squared:  0.4656
F-statistic: 13.69 on 7 and 95 DF,  p-value: 3.915e-12
```

1.3. Compressive strength (28-day)(Mpa)

```
> summary(modelocompressive)

Call:
lm(formula = compr_stre ~ cement + slag + fly_ash + water + sp +
    coar_agg + fine_agg, data = slump_test_1_)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8411 -1.7063 -0.2831  1.2986  7.9424

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 139.78150   71.10128   1.966  0.05222 .
cement        0.06141    0.02282   2.691  0.00842 **
slag         -0.02971    0.03176  -0.935  0.35200
fly_ash       0.05053    0.02316   2.182  0.03159 *
water        -0.23270    0.07166  -3.247  0.00161 **
sp            0.10315    0.13459   0.766  0.44532
coar_agg     -0.05562    0.02744  -2.027  0.04546 *
fine_agg     -0.03908    0.02882  -1.356  0.17833
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.609 on 95 degrees of freedom
Multiple R-squared:  0.8968,    Adjusted R-squared:  0.8892
F-statistic:   118 on 7 and 95 DF,  p-value: < 2.2e-16
```

Crítica aos modelos lineares obtidos

1ª Versão

De acordo com os resultados obtidos em Pr(>|t|), R^2 ajustado, F-statistic e p-value associado concluímos que:

- O modelo 1.1. deverá ser sujeito ao método de backward elimination, onde concecutivamente serão removidas as variáveis Fly_Ash, Slag, Cement, Coar_agg, SP e Fine_agg para uma melhor adaptação do modelo linear aos dados.
- O modelo 1.2. deverá ser sujeito ao método de backward elimination, onde concecutivamente serão removidas as variáveis Slag e SP, pela mesma razão acima mencionada.
- O modelo 1.3. revela-se muito satisfatório (R^2 ajustado = 0.8892), contudo iremos testar um novo modelo onde se excluí a variável Slag.

Modelos lineares com tratamento

2ª Versão

Descrição do procedimento de treino e teste do modelo

Com base na análise exploratória efetuada, decidimos desta vez dividir o dataset numa base de dados de treino e numa de teste, originando 85 e 18 resultados, respetivamente.

Por outras palavras, 80% do comportamento de y, isto é, das variáveis Slump, Flow e Compressive Strenght serão explicadas pelos 20% dos dados obtidos das variáveis x (Cement, Slag, Fly_Ash, Water, SP, Coar_Agg, Fine_Agg), segundo o princípio de Pareto.

```
> set.seed(123)
> split <- sample(2, nrow(slump_test_1_), replace = TRUE, prob = c(0.8, 0.2))
> training_set <- slump_test_1_[split == 1, ]
> test_set <- slump_test_1_[split == 2, ]
> nrow(training_set)
[1] 85
> nrow(test_set)
[1] 18
```

Figura 10. Split (80-20) do dataset original.

Remoção de outliers

A partir da análise prévia do boxplot da distribuição das variáveis observa-se que apenas a variável SP possui outliers.

```
> boxplot(slump_test_1_$sp, main = "SP" )
>
> sp_outliers<-which(slump_test_1_$sp %in% boxplot(slump_test_1_$sp)$out)
>
> sp_outliers <- unique(sp_outliers)
> slump_test_1_[sp_outliers,]
  cement slag fly_ash water sp coar_agg fine_agg slump flow compr_stre
4    162  148     190   179 19      838      741     3 21.5      42.08
7    152  139     178   168 18      944      695     0 20.0      38.86
```

Figura 11. Determinação dos outliers.

Foram identificadas as observações número 4 e 7 como outliers. Antes de aplicarmos o novo modelo de regressão linear múltipla, estes valores serão removidos.

Correlação

Verificou-se a correlação existente entre variáveis, de onde se depreende quais as variáveis mais preponderantes (próximo de 1, mas nunca auto-correlacionadas) a selecionar para o modelo.

```
> cor(training_set[,c(2,3,4,5,6,7,8,9)])
               cement         slag    fly_ash        water           sp
cement      1.00000000 -0.222486688 -0.5073494  0.268794590 -0.01060553
slag       -0.22248669  1.000000000 -0.2786484  0.007904724  0.30721689
fly_ash    -0.50734945 -0.278648435  1.0000000 -0.282362070 -0.18627086
water       0.26879459  0.007904724 -0.2823621  1.000000000 -0.08593223
sp         -0.01060553  0.307216889 -0.1862709 -0.085932226  1.00000000
coarse_agg -0.34295908 -0.332319967  0.2365237 -0.604640437 -0.18459862
fine_aggr   0.06776705 -0.133094893 -0.3783242  0.097516162  0.07273156
slump       0.16159863 -0.287830957 -0.1395011  0.510483468 -0.19221655
            coarse_agg   fine_aggr       slump
cement     -0.3429591  0.06776705  0.1615986
slag       -0.3323200 -0.13309489 -0.2878310
fly_ash     0.2365237 -0.37832421 -0.1395011
water      -0.6046404  0.09751616  0.5104835
sp         -0.1845986  0.07273156 -0.1922165
coarse_agg  1.0000000 -0.43013146 -0.1970394
fine_aggr  -0.4301315  1.00000000  0.2066685
slump      -0.1970394  0.20666849  1.0000000
```

Figura 12. Determinação da correlação entre as diversas variáveis.

Modelos atualizados com base na crítica aos originais

Foram posteriormente elaborados modelos de regressão linear múltipla corrigidos para as variáveis de output *slump*, *flow* e *compr_stre*, verificando-se agora Pr(>|t|) mais significativos.

```
> summary(modelslumpcorfine)

Call:
lm(formula = slump ~ slag + water, data = slump_test_1_)

Residuals:
    Min      1Q  Median      3Q     Max
-16.871  -5.326   2.493   5.428  11.356

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -18.09945    7.31394  -2.475  0.01502 *
slag         -0.03933    0.01219  -3.227  0.00169 **
water         0.19889    0.03646   5.455 3.56e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.439 on 100 degrees of freedom
Multiple R-squared:  0.2915,    Adjusted R-squared:  0.2773
F-statistic: 20.57 on 2 and 100 DF,  p-value: 3.294e-08
```

```
> summary(modeloflowcorsp)

Call:
lm(formula = flow ~ cement + fly_ash + water + coar_agg + fine_agg,
    data = slump_test_1_)

Residuals:
    Min      1Q  Median      3Q     Max
-31.893 -10.125   1.773   9.559  23.914

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -249.50866   48.90884  -5.102 1.67e-06 ***
cement         0.05366    0.01979   2.712 0.007909 **
fly_ash        0.06101    0.01859   3.281 0.001436 **
water          0.72313    0.08426   8.582 1.53e-13 ***
coar_agg       0.07291    0.02266   3.217 0.001760 **
fine_agg       0.09554    0.02573   3.714 0.000341 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.74 on 97 degrees of freedom
Multiple R-squared:  0.5003,    Adjusted R-squared:  0.4745
F-statistic: 19.42 on 5 and 97 DF,  p-value: 2.36e-13
```

```
> summary(modelocompressivecor)

Call:
lm(formula = compr_stre ~ cement + slag + fly_ash + water + coar_agg +
    fine_agg, data = slump_test_1_)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8507 -1.7931 -0.1958  1.1138  7.7033

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 177.11354   51.68426   3.427 0.000901 ***
cement        0.04970    0.01692   2.938 0.004135 **
slag         -0.04519    0.02446  -1.847 0.067782 .
fly_ash       0.03859    0.01710   2.257 0.026291 *
water        -0.27055    0.05181  -5.222 1.03e-06 ***
coar_agg     -0.06986    0.02015  -3.468 0.000786 ***
fine_agg     -0.05358    0.02170  -2.469 0.015337 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.603 on 96 degrees of freedom
Multiple R-squared:  0.8962,    Adjusted R-squared:  0.8897
F-statistic: 138.1 on 6 and 96 DF,  p-value: < 2.2e-16
```

Avaliação da normalidade

```
> Box.test(regressor_um$residuals, type="Ljung")

        Box-Ljung test

data:  regressor_um$residuals
X-squared = 1.1676, df = 1, p-value = 0.2799
```

Figura 13. Box-test aplicado aos três modelos corrigidos.

Para validar o modelo linear de regressão múltipla, comprovamos também a inexistência de padrões na curva de erros, bem como verificamos a normalidade dos resíduos e todos os restantes passos acima detalhados nos primeiros modelos. Pode assumir-se normalidade com base nos resultados.

Comparação do erro entre modelo original e corrigido

Para o cálculo do erro podem ser usadas diferentes métricas, como sejam ME, MSE e MAPE. É possível constatar que para os três modelos o erro associado é muito similar comparando o modelo linear original (Versão 1) com o corrigido (Versão 2), o que indica que este último foi bem trabalhado e representativo do dataset original.

```
> slump_predi=predict(modeloslump,newdata=test_set[,-c(1)])
> slump_predi2=predict(modelslumpcorfine,newdata=test_set[,-c(1,2,4,6,7,8)])
> ME_1 <- mean(test_set$slump - slump_predi)
> ME_2 <- mean(test_set$slump - slump_predi2)
> MSE_1 <- mean((test_set$slump - slump_predi)^2)
> MSE_2 <- mean((test_set$slump - slump_predi2)^2)
> MAPE_1 <- mean(abs(test_set$slump - slump_predi)/test_set$slump)
> MAPE_2 <- mean(abs(test_set$slump - slump_predi2)/test_set$slump)
> data.frame("Métrica" = c("ME", "MSE", "MAPE"), "Model Slump" = c(ME_1, MSE_1, MAPE_1),
+            "Model Slump Corrected" = c(ME_2, MSE_2, MAPE_2))
  Métrica Model.Slump Model.Slump.Corrected
1      ME   1.0818490             1.0099449
2     MSE  43.8379775            46.1124646
3    MAPE   0.3661554             0.4281634

> flow_predi=predict(modeloflow,newdata=test_set[,-c(1)])
> flow_predi2=predict(modeloflowcorsp,newdata=test_set[,-c(1,3,6)])
> ME_Flow <- mean(test_set$flow - flow_predi)
> ME_Flow <- mean(test_set$flow - flow_predi2)
> ME_Flow <- mean(test_set$flow - flow_predi)
> ME_Flow2 <- mean(test_set$flow - flow_predi2)
> MSE_Flow1 <- mean((test_set$flow - flow_predi)^2)
> MSE_Flow2 <- mean((test_set$flow - flow_predi2)^2)
> MAPE_Flow1 <- mean(abs(test_set$flow - flow_predi)/test_set$flow)
> MAPE_Flow2 <- mean(abs(test_set$flow - flow_predi2)/test_set$flow)
> data.frame("Métrica" = c("ME_Flow", "MSE_Flow1", "MAPE_Flow1"), "Model Flow" = c(ME_Flow, MSE_Flow1, MAPE_1),
+            "Model Slump Corrected" = c(ME_Flow2, MSE_Flow2, MAPE_Flow2))

> data.frame("Métrica" = c("ME_Flow", "MSE_Flow1", "MAPE_Flow1"), "Model Flow" = c(ME_Flow, MSE_Flow1, MAPE_Flow1),"Model Flow Co
rrected" = c(ME_Flow2, MSE_Flow2, MAPE_Flow2))
     Métrica  Model.Flow Model.Flow.Corrected
1    ME_Flow   2.0836525            2.3258355
2  MSE_Flow1 107.1882371          107.2608845
3 MAPE_Flow1   0.1725558            0.1709023
>

> ME_Compressive1 <- mean(test_set$compr_stre - compressive_predi)
> ME_Compressive2 <- mean(test_set$compr_stre - compressive_predi2)
> MSE_Compressive1 <- mean((test_set$compr_stre - compressive_predi)^2)
> MSE_Compressive2 <- mean((test_set$compr_stre - compressive_predi2)^2)
> MAPE_Compressive1 <- mean(abs(test_set$compr_stre - compressive_predi)/test_set$compr_stre)
> MAPE_Compressive2 <- mean(abs(test_set$compr_stre - compressive_predi2)/test_set$compr_stre)
> data.frame("Métrica" = c("ME_Compressive1", "MSE_Compressive1", "MAPE_Flow1"), "Model Flow" = c(ME_Compressive1, MSE_Compressiv
e1, MAPE_Compressive1),"Model Flow Corrected" = c(ME_Compressive2, MSE_Compressive2, MAPE_Compressive2))
          Métrica Model.Flow Model.Flow.Corrected
1  ME_Compressive1 0.44298918           0.44298918
2 MSE_Compressive1 7.95156218           7.95156218
3       MAPE_Flow1 0.06016843           0.06016843
>
```

Figura 14. Comparação do erro obtido entre modelo inicial e corrigido.

Conclusão:

Após a construção dos modelos de regressão linear múltipla das variáveis slump, flow e compressive strength conseguimos concluir que:

- As variáveis slag e water são as mais significativas para o nosso modelo slump, isto é, sabemos que estas tem mais preponderância para uma previsão linear do comportamento de queda do concreto (slump).

- No modelo flow, a combinação das variáveis cement, fly_ash, water e coarse aggregation é a mais significativa pelo que estes fatores devem ser considerados para a composição do cimento ter a fluidez (flow) desejada.

- Já na compressive strength, observa-se uma previsão eficaz da rigidez do concreto considerando todas as variáveis deste dataset, com uma participação menor da variável slag.

Existe uma questão fundamental na interpretação dos modelos que é o número de observações reduzido. Assim, era ideal aumentar o nosso dataset para uma previsão mais certeira do comportamento das variáveis de output.