Atividade 1 - Regressão Múltipla

Talita Lima da Silva

Agosto, 2019

1 Análise do Tamanho da Amostra

Inicialmente, realizou-se uma análise de poder do efeito da amostra estudada, por meio o *software* R, neste caso foi utilizado o teste f^2 que é usual no contexto da ANOVA e da regressão múltipla.

```
> pwr.f2.test(u = 14 , v = 86 ,f2 = NULL, sig.level = 0.05, power = 0.8)

     Multiple regression power calculation

              u = 14
              v = 86
             f2 = 0.2070028
      sig.level = 0.05
          power = 0.8
```

Segundo o Teste de Cohen (f^2), quando se obtém-se $f^2 = 0.15$, este possui um poder médio e quando $f^2 = 0.35$ o poder é definidio como alto. Considerando que o valor de f^2 encontrado para a nossa amostra foi de $f^2 = 0.2070028$, esta possui um poder de efeito (*effect size*) entre médio e alto.

2 Teste das Suposições no Modelo de Regressão

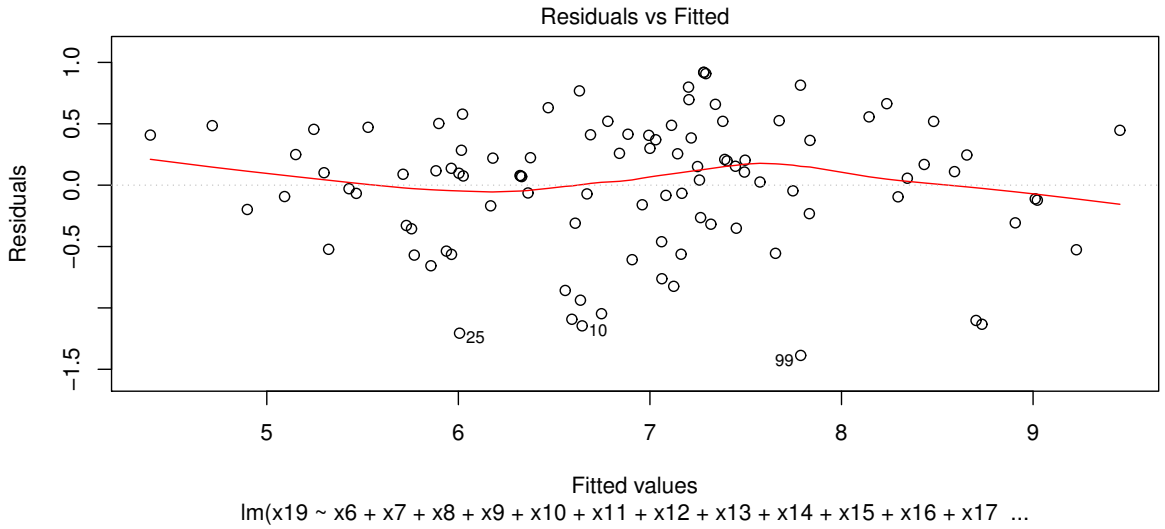
Analisada a mostra e comprovado o seu pode de efeito ($f^2 = 0.2$), montou-se a função em que a regressão linear múltipla será rodada. Nela, foi considerada a variável dependente X19 (satisfação do cliente) e as variáveis independentes X6, X7, X8, X9, X10, X11, X12, X13, X14, X15, X16, X17 e X18 como ser observado a seguir:

```
lm1<-lm(x19 ~ x6 + x7 + x8 + x9 + x10 + x11 + x12 + x13 + x14 + x15 +
x16 + x17 + x18,data = hbat)

plot(lm1)
```

Montado o modelo, obteve-se o gráfico evidenciado na Figura 1, onde é possível observar uma distribuição semelhante a de um gráco nulo, evidenciando que os resíduos ocorrem aleatoriamente, com dispersão relativamente igual em torno de zero e nenhuma tendência forte para ser maior ou menor que

zero. Do mesmo modo, nenhum padrão é encontrado para valores grandes *versus* pequenos da variável independente.

Figura 1: Distribuição das variáveis dependente e independentes



1. **Linearidade:** Segundo o autor Hair et al. (2009) a linearidade de qualquer relação bivariada pode ser identificada por meio de gráficos de resíduos, como o plotado na Figura 1. Nesse sentido, observando-se a Figura 1 é possível notar que há uma linearidade entre as variáveis analisadas.

2. **Homoscedasticidade:** O diagnóstico desse tipo de suposição também é feito por meio de gráficos de resíduos ou testes estatísticos (Hair et al., 2009). Observando o gráfico da Figura 1 já é possível depreender que não há heteroscedasticidade entre as variáveis. Entretanto, para confirmar essa análise, foi realizado o *The Breush-Pagan test*, por meio do R, e o resultado obtido pode ser observado por meio do código utilizado abaixo:

```
> #teste para homoscedasticidade

> bptest(lm1)

studentized Breusch-Pagan test

data:  lm1
BP = 11.611, df = 13, p-value = 0.5598
```

   Como foi retornado um valor de p-value maior que um nível de significância 0,05, falha-se em rejeitar a hipótese nula de que a variância dos resíduos é constante e inferir que a homoscedasticidade está de fato presente, confirmando assim a inferência gráfica.

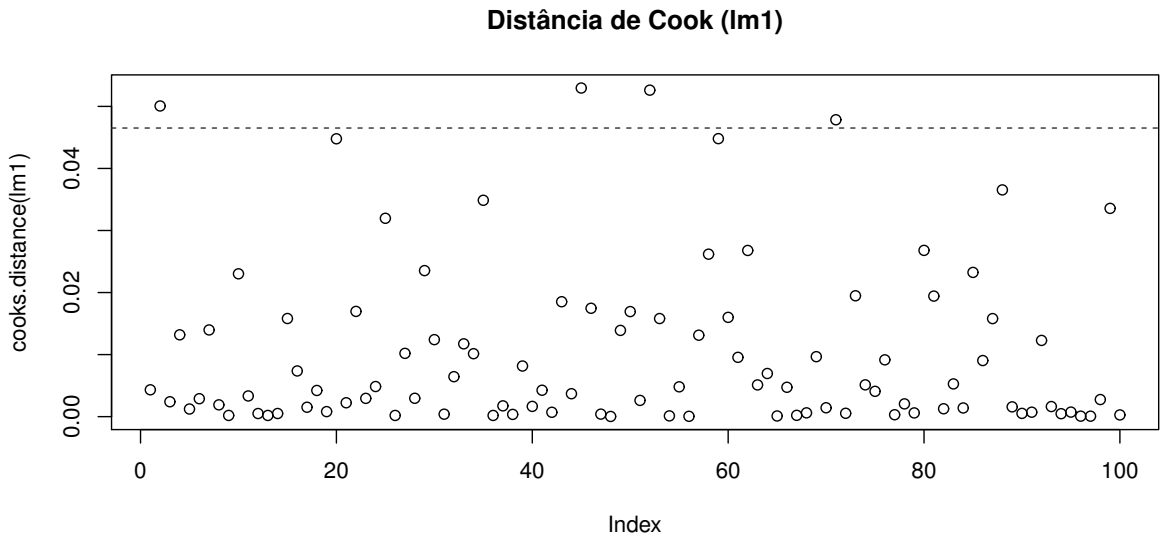
3. **Normalidade da distribuição:**

```
> shapiro.test(lm1$residuals)

Shapiro-Wilk normality test

data:  lm1$residuals
W = 0.96644, p-value = 0.01188
```

3 Detecção e Análise de *Outliers*

Para a observação de *outliers* (valores muito extremos, potencialmente errados ou fruto de alguma anormalidade) no modelo foi realizado o Teste de Distância de Cook, alguns estudiosos defendem que valores superiores a 1 são suspeitos, outros defendem que valores da distância de Cook superiores a 4/(N-k-1) já seriam suspeitos, onde N é o número de observações e k o número de variáveis explicativas. No caso do nosso modelo, o critério adotado foi o de 4/(N-k-1) e o gráfico obtido pode ser observado na Figura 2:

Figura 2: Distância de Cook do Modelo lm1



Como pode ser observado, de acordo com a Figura 2, existe um número muito baixo de *outliers*, em torno de quatro, o que não impacta nem compromete fortemente a amostra do nosso modelo e que, portanto, não precisa sofrer correção.

4 Realização da Regressão Linear Múltipla

Analisada a amostra, foi rodada a regressão linear múltipla da função **lm1** estabelecida no início do estudo, utilizando o método *stepwise* que é usado na construção de modelos para identificar um subconjunto útil de preditores. O processo adiciona sistematicamente a variável mais significativa (*forward*) ou remove a variável menos significativa durante cada etapa.(*backward*).

No método de *setpwise* utilizado no R foi solicitado que ocorressem as duas verificações, tanto a *backward* quanto a *forward* para dar mais robustez ao resultado obtido.

```
> step(lm1,direction = "both")
Start:  AIC=-100.79
x19 ~ x6 + x7 + x8 + x9 + x10 + x11 + x12 + x13 + x14 + x15 +
    x16 + x17 + x18

       Df Sum of Sq    RSS      AIC
- x15   1    0.0018 27.585 -102.788
- x18   1    0.0753 27.659 -102.522
- x8    1    0.0943 27.678 -102.453
- x10   1    0.0962 27.680 -102.447
- x14   1    0.2329 27.817 -101.954
- x17   1    0.2445 27.828 -101.912
- x13   1    0.3110 27.895 -101.674
<none>              27.584 -100.795
- x11   1    0.5922 28.176 -100.670
- x16   1    0.6009 28.185 -100.640
- x9    1    0.7113 28.295 -100.249
- x7    1    3.5800 31.164  -90.592
- x6    1   16.4466 44.030  -56.029
- x12   1   21.3289 48.913  -45.514

Step:  AIC=-102.79
x19 ~ x6 + x7 + x8 + x9 + x10 + x11 + x12 + x13 + x14 + x16 +
    x17 + x18

       Df Sum of Sq    RSS      AIC
- x18   1    0.0739 27.659 -104.521
- x10   1    0.0970 27.682 -104.437
- x8    1    0.1001 27.685 -104.426
- x14   1    0.2432 27.829 -103.911
- x17   1    0.2434 27.829 -103.910
- x13   1    0.3104 27.896 -103.670
<none>              27.585 -102.788
- x11   1    0.5928 28.178 -102.662
- x16   1    0.6036 28.189 -102.624
- x9    1    0.7183 28.304 -102.218
+ x15   1    0.0018 27.584 -100.795
- x7    1    3.5953 31.181  -92.537
- x6    1   16.5158 44.101  -57.868
- x12   1   21.3702 48.956  -47.426

Step:  AIC=-104.52
x19 ~ x6 + x7 + x8 + x9 + x10 + x11 + x12 + x13 + x14 + x16 +
    x17
```

```
       Df Sum of Sq    RSS      AIC
- x10   1    0.0864 27.746 -106.209
- x8    1    0.0933 27.753 -106.184
- x14   1    0.2228 27.882 -105.718
- x13   1    0.2779 27.937 -105.521
- x17   1    0.3946 28.054 -105.104
<none>              27.659 -104.521
- x16   1    0.5643 28.224 -104.501
- x9    1    0.6840 28.343 -104.078
+ x18   1    0.0739 27.585 -102.788
+ x15   1    0.0003 27.659 -102.522
- x11   1    2.3002 29.960  -98.532
- x7    1    3.5222 31.181  -94.535
- x6    1   16.5666 44.226  -59.586
- x12   1   22.1356 49.795  -47.726

Step:  AIC=-106.21
x19 ~ x6 + x7 + x8 + x9 + x11 + x12 + x13 + x14 + x16 + x17

       Df Sum of Sq    RSS      AIC
- x8    1    0.1039 27.850 -107.835
- x14   1    0.2313 27.977 -107.379
- x13   1    0.2545 28.000 -107.296
- x17   1    0.3291 28.075 -107.030
<none>              27.746 -106.209
- x16   1    0.5893 28.335 -106.107
- x9    1    0.7408 28.486 -105.574
+ x10   1    0.0864 27.659 -104.521
+ x18   1    0.0633 27.682 -104.437
+ x15   1    0.0007 27.745 -104.212
- x11   1    2.2247 29.970 -100.496
- x7    1    3.5274 31.273  -96.241
- x6    1   16.5294 44.275  -61.475
- x12   1   23.8218 51.567  -46.228

Step:  AIC=-107.84
x19 ~ x6 + x7 + x9 + x11 + x12 + x13 + x14 + x16 + x17

       Df Sum of Sq    RSS      AIC
- x14   1    0.1357 27.985 -109.349
- x13   1    0.3022 28.152 -108.756
- x17   1    0.3113 28.161 -108.724
- x16   1    0.5288 28.378 -107.954
<none>              27.850 -107.835
- x9    1    0.8650 28.715 -106.777
+ x8    1    0.1039 27.746 -106.209
```

```
+ x10   1    0.0970 27.753 -106.184
+ x18   1    0.0561 27.793 -106.037
+ x15   1    0.0055 27.844 -105.855
- x11   1    2.1322 29.982 -102.458
- x7    1    3.4608 31.310  -98.122
- x6    1   16.6259 44.475  -63.023
- x12   1   23.7497 51.599  -48.166

Step:  AIC=-109.35
x19 ~ x6 + x7 + x9 + x11 + x12 + x13 + x16 + x17

       Df Sum of Sq    RSS      AIC
- x13   1    0.2590 28.244 -110.428
- x17   1    0.4011 28.386 -109.926
- x16   1    0.4454 28.431 -109.770
<none>              27.985 -109.349
- x9    1    0.8755 28.861 -108.268
+ x14   1    0.1357 27.850 -107.835
+ x10   1    0.0916 27.894 -107.677
+ x18   1    0.0425 27.943 -107.501
+ x15   1    0.0084 27.977 -107.379
+ x8    1    0.0084 27.977 -107.379
- x11   1    2.1044 30.090 -104.099
- x7    1    3.3852 31.371  -99.930
- x6    1   17.3949 45.380  -63.009
- x12   1   23.8422 51.828  -49.725

Step:  AIC=-110.43
x19 ~ x6 + x7 + x9 + x11 + x12 + x16 + x17

       Df Sum of Sq    RSS      AIC
- x17   1    0.2637 28.508 -111.498
- x16   1    0.5454 28.790 -110.515
<none>              28.244 -110.428
+ x13   1    0.2590 27.985 -109.349
- x9    1    0.9528 29.197 -109.110
+ x14   1    0.0926 28.152 -108.756
+ x10   1    0.0698 28.175 -108.675
+ x18   1    0.0210 28.223 -108.502
+ x15   1    0.0091 28.235 -108.460
+ x8    1    0.0000 28.244 -108.428
- x11   1    2.2643 30.509 -104.716
- x7    1    3.4567 31.701 -100.882
- x6    1   17.6717 45.916  -63.836
- x12   1   23.5935 51.838  -51.705

Step:  AIC=-111.5
```

```
x19 ~ x6 + x7 + x9 + x11 + x12 + x16

       Df Sum of Sq    RSS      AIC
<none>              28.508 -111.498
+ x17   1    0.2637 28.244 -110.428
- x16   1    0.9140 29.422 -110.342
+ x18   1    0.1801 28.328 -110.132
+ x14   1    0.1662 28.342 -110.083
+ x13   1    0.1217 28.386 -109.926
+ x10   1    0.0171 28.491 -109.558
+ x8    1    0.0135 28.495 -109.546
+ x15   1    0.0001 28.508 -109.499
- x9    1    2.3970 30.905 -105.425
- x11   1    2.6247 31.133 -104.691
- x7    1    3.3996 31.908 -102.232
- x6    1   18.6802 47.188  -63.102
- x12   1   24.1101 52.618  -52.211

Call:
lm(formula = x19 ~ x6 + x7 + x9 + x11 + x12 + x16, data = hbat)

Coefficients:
(Intercept)      x6           x7           x9          x11
 -1.2690       0.3650      -0.4364       0.2258       0.1766

  x12           x16
 0.7817       0.1591

> modelo1<-lm(formula = x19 ~ x6 + x7 + x9 + x11 + x12 + x16, data = hbat)
> summary(modelo1)

Call:
lm(formula = x19 ~ x6 + x7 + x9 + x11 + x12 + x16, data = hbat)

Residuals:
<Labelled double>: X19 - Satisfaction
     Min       1Q   Median       3Q      Max
-1.40616 -0.32428  0.03067  0.38672  0.97362

Labels:
 value                label
     0 Not At All Satisfied
    10 Completely Satisfied

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.26902    0.49935  -2.541  0.01270 *
```

```
x6           0.36499    0.04676   7.806 8.59e-12 ***
x7          -0.43635    0.13103  -3.330  0.00125 **
x9           0.22577    0.08074   2.796  0.00628 **
x11          0.17655    0.06034   2.926  0.00431 **
x12          0.78167    0.08814   8.869 5.06e-14 ***
x16          0.15911    0.09215   1.727  0.08753 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5537 on 93 degrees of freedom
Multiple R-squared:  0.7973,Adjusted R-squared:  0.7842
F-statistic: 60.96 on 6 and 93 DF,  p-value: < 2.2e-16
```

5 Interpretação dos Resultados do Modelo

Inicialmente, a partir do modelo rodado foram retornadas 6 variáveis (de 13) que melhor explicam o modelo proposto: X_6 (qualidade do produto), X_7 (comércio eletrônico), X_9 (solução de reclamação), X_{11} (linha do produto), X_{12} (imagem da equipe) e X_{16} (encomenda e cobrança). Ou seja, estas são as variáveis que impactam mais diretamente na variável dependente X_{19} (satisfação do cliente) aqui estudada.

Portanto, a partir do modelo:

$$X_{19} = -1.26902 + (0.36499X_6) + (-0.43635X_7) + (0.22577X_9) + (0.17655X_{11}) + (0.78167X_{12}) + (0.15911X_{16})$$

É possível inferir que a variável X_7 (comércio eletrônico) influencia negativamente X19, a cada uma unidade de variação que ela sofra. Alternativamente, a variável X_{12} é a que influência mais positivamente a variável X19 a cada unidade de variação que ocorra nela.

Além disso, foi obtido um coeficiente de determinação de R^2 = 0.8039 e um o R^2 Ajustado = 0.7742. Sabe-se que o coeficiente de determinação corresponde a quanto o modelo consegue explicar os valores observados, ou seja, quanto maior o R^2, mais explicativo é o modelo e melhor ele se ajusta à amostra. Portanto, pode-se dizer que que o modelo lm1 consegue explicar bem as variáveis independentes selecionadas.

6 Validação do Modelo

A fim de validar o modelo encontrado, foi rodada um segunda regressão linear utilizando uma amostra aleatória menor, dentro da própria amostra. Desse modo, foi sorteada uma amostra de tamanho 50 (que corresponde a 50% do total) e realizado o método de *stepwise* novamente, como pode ser observado a seguir:

```
> sample_50<-hbat[sample(nrow(hbat),50),]
> lm2<-lm(x19 ~ x6 + x7 + x9 + x11 + x12 + x16, data = sample_50)
> step(lm2, direction = "both")
Start:  AIC=-44.91
x19 ~ x6 + x7 + x9 + x11 + x12 + x16
```

```
       Df Sum of Sq    RSS     AIC
- x16   1    0.5254 15.916 -45.236
- x11   1    0.6237 16.014 -44.928
<none>              15.390 -44.914
- x7    1    0.8981 16.288 -44.078
- x9    1    2.0159 17.406 -40.760
- x6    1    7.5481 22.938 -26.961
- x12   1   10.8208 26.211 -20.292

Step:  AIC=-45.24
x19 ~ x6 + x7 + x9 + x11 + x12

       Df Sum of Sq    RSS     AIC
<none>              15.916 -45.236
+ x16   1    0.5254 15.390 -44.914
- x7    1    0.8130 16.729 -44.745
- x11   1    0.8757 16.791 -44.558
- x9    1    4.1169 20.033 -35.733
- x6    1    7.7000 23.616 -27.506
- x12   1   10.8404 26.756 -21.263

Call:
lm(formula = x19 ~ x6 + x7 + x9 + x11 + x12, data = sample_50)

Coefficients:
(Intercept)           x6           x7           x9          x11          x12
    -1.3570       0.3630      -0.3228       0.3590       0.1574       0.7316

> lm3<-(lm(x19 ~ x6 + x7 + x9 + x11 + x12, data = hbat))
> summary(lm3)

Call:
lm(formula = x19 ~ x6 + x7 + x9 + x11 + x12, data = hbat)

Residuals:
<Labelled double>: X19 - Satisfaction
     Min       1Q   Median       3Q      Max
-1.40116 -0.30067  0.05776  0.41636  0.91403

Labels:
 value                label
     0 Not At All Satisfied
    10 Completely Satisfied

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -1.15106    0.49984  -2.303  0.02349 *
x6           0.36900    0.04719   7.820 7.61e-12 ***
x7          -0.41714    0.13192  -3.162  0.00211 **
x9           0.31896    0.06068   5.256 9.16e-07 ***
x11          0.17435    0.06095   2.860  0.00521 **
x12          0.77513    0.08898   8.711 1.01e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5595 on 94 degrees of freedom
Multiple R-squared:  0.7908,Adjusted R-squared:  0.7797
F-statistic: 71.06 on 5 and 94 DF,  p-value: < 2.2e-16
```

Realizando o mesmo teste para uma amostra menor, foi possível observar que a variável X_{16} foi retirada do modelo. Portanto, para uma amostra menor ele possui uma influência baixa ou quase insignificante sobre a variável X_{19}. Além disso, é possível observar que o valor entre de o R^2 = 0.7908 e o R^2 Ajustado = 0.7797 sofre menos variação, ou seja, pouca perda no poder preditivo, o que indica uma falta de superajuste

Dessa forma, o modelo final após etapa de validação é

$$X_{19} = -1.15106 + (0.36900X_6) + (-0.41714X_7) + (0.31896X_9) + (0.17435X_{11}) + (0.77513X_{12})$$

Referências

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., and Tatham, R. L. (2009). *Análise multivariada de dados*. Bookman Editora.