Regressão linear: Apartamentos em Criciúma

Daniel Amato Zabotti

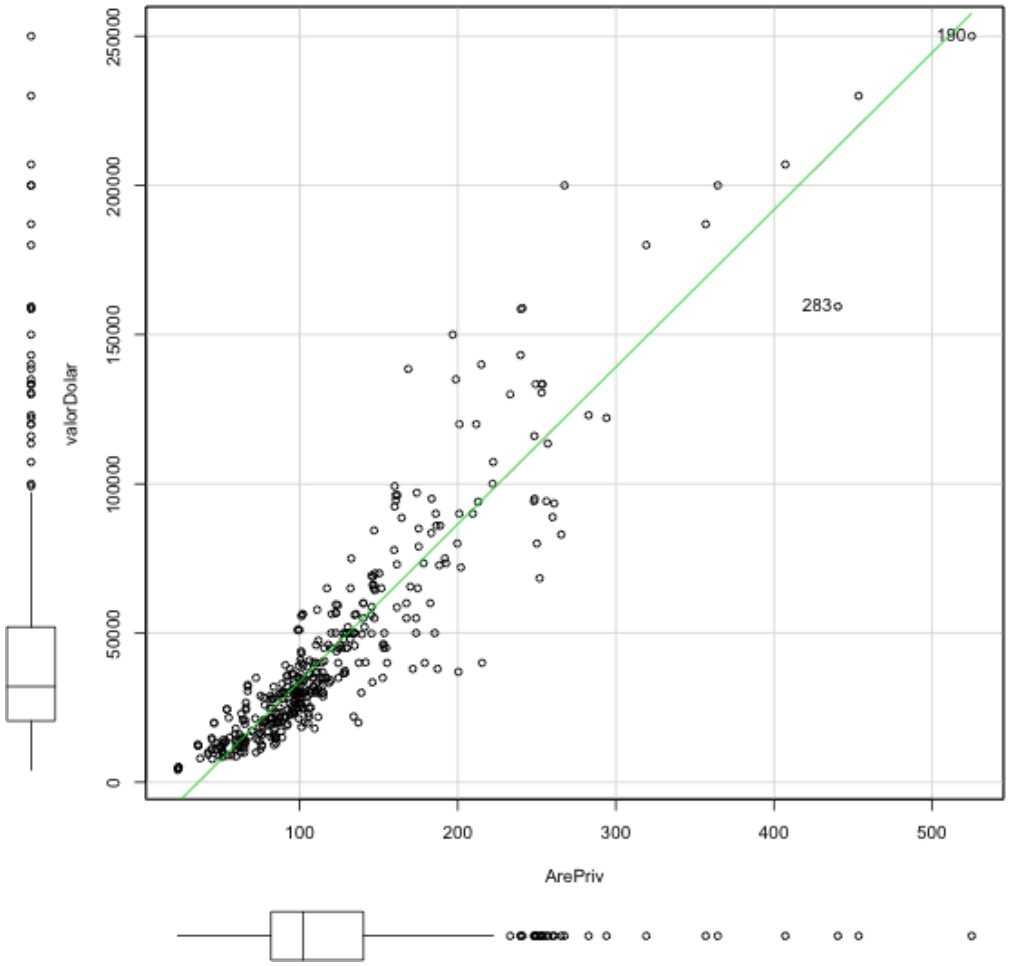
2014-04-24

```
> library(XLConnect, pos = 4)
```

```
> .Workbook <- loadWorkbook("/Users/daniel/Google Drive/UFSC-SIN/INE5649 - Técnicas Estatisticas de Predição
/Aula 24-04-2014/Apartamentos Criciuma completo.xls")
```

```
> ApartamentosCriciuma <- readWorksheet(.Workbook, "Parte_dos_dados")
```

```
> scatterplot(valorDolar ~ ArePriv, reg.line = lm, smooth = FALSE, spread = FALSE,
+     id.method = "mahal", id.n = 2, boxplots = "xy", span = 0.5, data = ApartamentosCriciuma)
```




```
190 283
190 283
```

Observando o gráfico há evidência de relação linear entre a área do imóvel e o seu valor.

Fica evidente pelo gráfico que a variância não é constante.

A distribuição das variáveis náo é simétrica. Podemos confirmar observando os boxplots.

```
> RegLinearValorDolarAreaPriv <- lm(valorDolar ~ ArePriv, data = ApartamentosCriciuma)
> summary(RegLinearValorDolarAreaPriv)
```

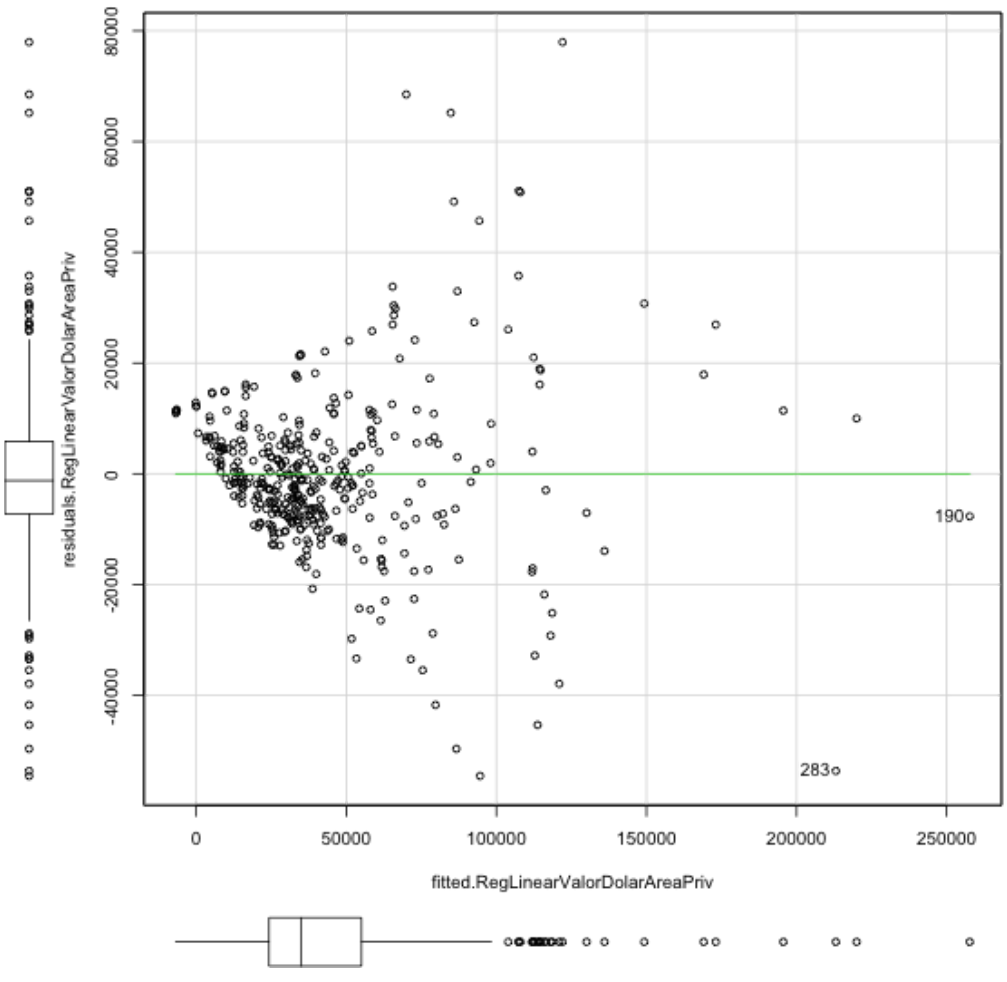
```
Call:
lm(formula = valorDolar ~ ArePriv, data = ApartamentosCriciuma)

Residuals:
   Min     1Q Median     3Q    Max
-54577  -7270  -1229   5844  77952

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -18795.8     1554.9   -12.1   <2e-16 ***
ArePriv        526.5       11.4    46.1   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14900 on 395 degrees of freedom
Multiple R-squared:  0.843, Adjusted R-squared:  0.843
F-statistic: 2.12e+03 on 1 and 395 DF,  p-value: <2e-16
```

De acordo com a regressão linear a equação é y = 526.48x -18795.79

O valor de R quadrado é 0.843. Significando que 84% da variância do valor do imóvel pode ser explicado pela área

```
> ApartamentosCriciuma$fitted.RegLinearValorDolarAreaPriv <- fitted(RegLinearValorDolarAreaPriv)
> ApartamentosCriciuma$residuals.RegLinearValorDolarAreaPriv <- residuals(RegLinearValorDolarAreaPriv)
```

```
> scatterplot(residuals.RegLinearValorDolarAreaPriv ~ fitted.RegLinearValorDolarAreaPriv,
+     reg.line = lm, smooth = FALSE, spread = FALSE, id.method = "mahal", id.n = 2,
+     boxplots = "xy", span = 0.5, data = ApartamentosCriciuma)
```



```
190 283
190 283
```

O modelo parece não estar adequado. A variabilidade em Y é grande e em X aumenta proporcionalmente.

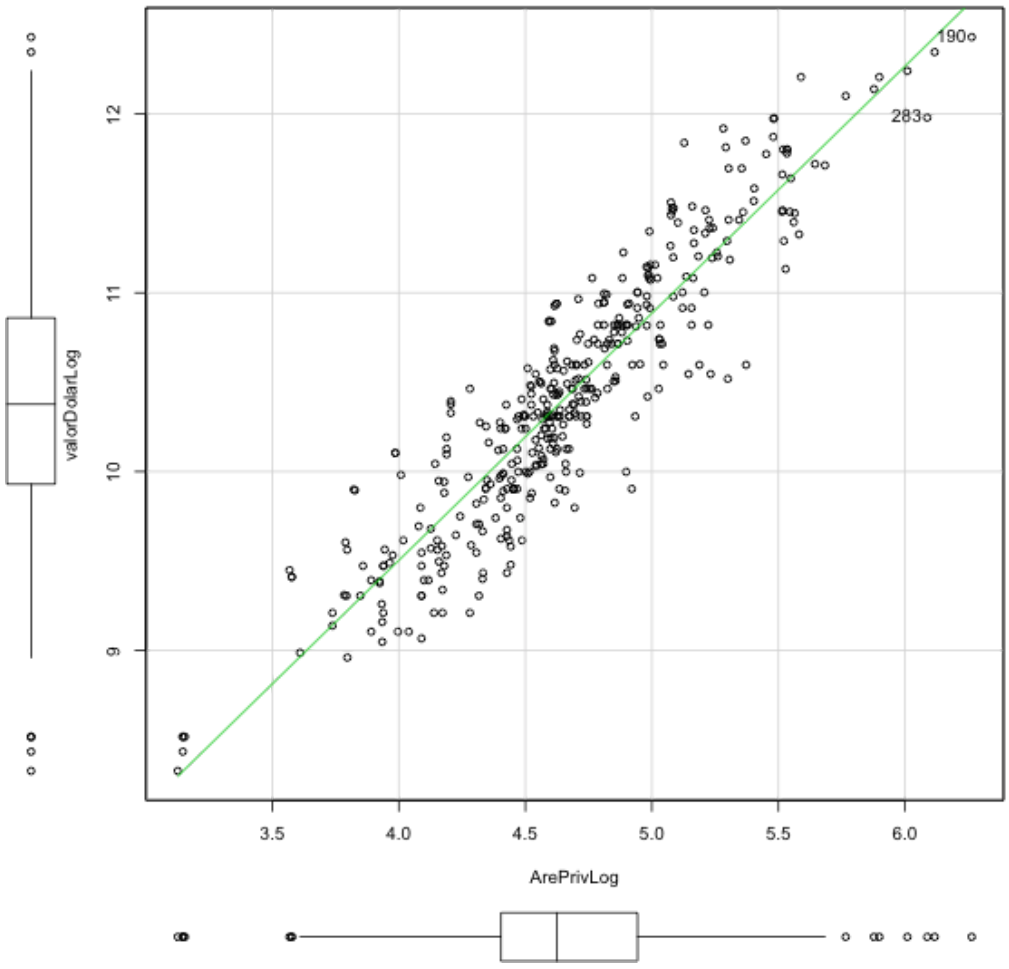
Podemos aplicar a transformação logarítmica em ambas as variáveis, afim de preservar a relação linear já observada

```
> ApartamentosCriciuma$ArePrivLog <- with(ApartamentosCriciuma, log(ArePriv))
```

```
> ApartamentosCriciuma$valorDolarLog <- with(ApartamentosCriciuma, log(valorDolar))
```

O novo gráfico, com as variáveis transformadas fica

```
> scatterplot(valorDolarLog ~ ArePrivLog, reg.line = lm, smooth = FALSE, spread = FALSE,
+     id.method = "mahal", id.n = 2, boxplots = "xy", span = 0.5, data = ApartamentosCriciuma)
```




```
190 283
190 283
```

Agora a relação fica mais harmoniosa.

```
> RegLinearLogDolarAreaPriv <- lm(valorDolarLog ~ ArePrivLog, data = ApartamentosCriciuma)
> summary(RegLinearLogDolarAreaPriv)
```

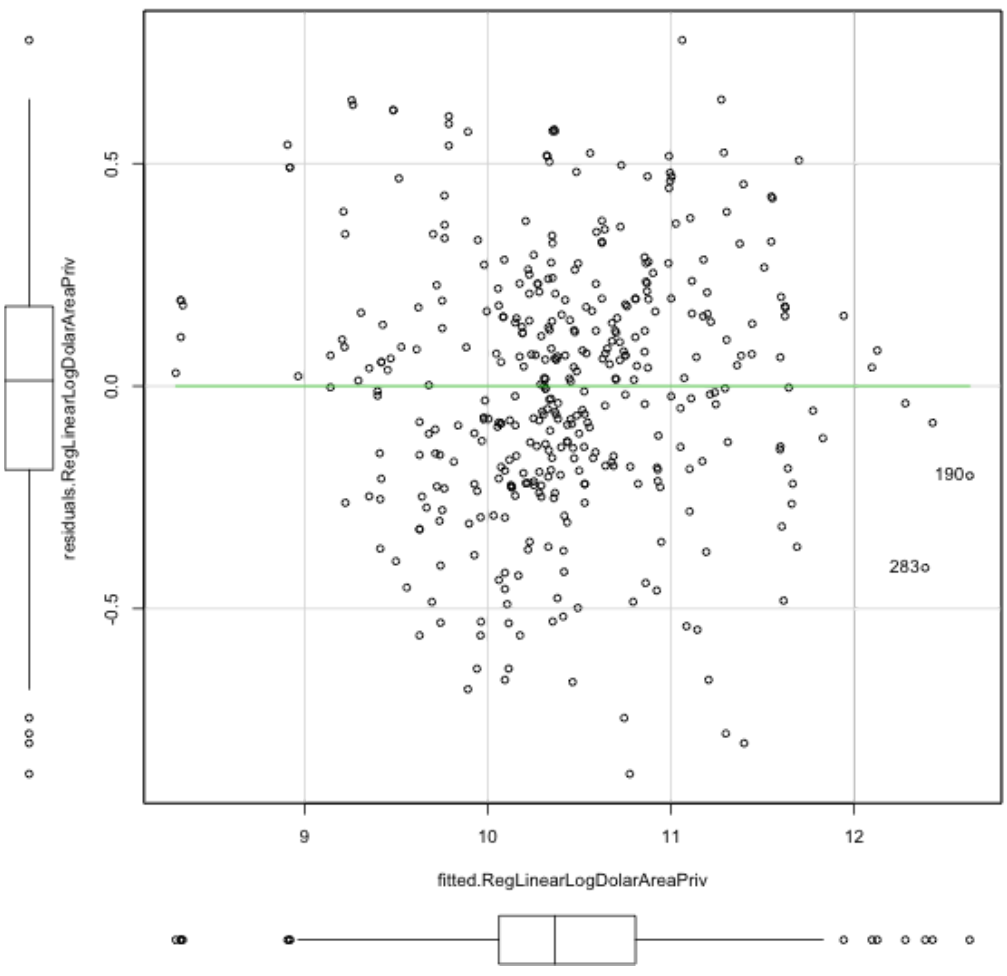
```
Call:
lm(formula = valorDolarLog ~ ArePrivLog, data = ApartamentosCriciuma)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8715 -0.1881  0.0123  0.1793  0.7773

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.9770     0.1386    28.7   <2e-16 ***
ArePrivLog    1.3816     0.0296    46.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.29 on 395 degrees of freedom
Multiple R-squared:  0.847, Adjusted R-squared:  0.846
F-statistic: 2.18e+03 on 1 and 395 DF,  p-value: <2e-16
```

A nova fórmula de regressão é y = 1.38x + 3.98

O valor de R quadrado é 0.843. Significando que 84% da variância do valor do imóvel pode ser explicado pela área

```
> ApartamentosCriciuma$fitted.RegLinearLogDolarAreaPriv <- fitted(RegLinearLogDolarAreaPriv)
> ApartamentosCriciuma$residuals.RegLinearLogDolarAreaPriv <- residuals(RegLinearLogDolarAreaPriv)
```

```
> scatterplot(residuals.RegLinearLogDolarAreaPriv ~ fitted.RegLinearLogDolarAreaPriv,
+     reg.line = lm, smooth = FALSE, spread = FALSE, id.method = "mahal", id.n = 2,
+     boxplots = "xy", span = 0.5, data = ApartamentosCriciuma)
```



```
190 283
190 283
```

Agora a distribuição dos erros ficou mais aleatória, indicando que o modelo está mais adequado.