01

Gráfico e Modelo Linear

Transcrição

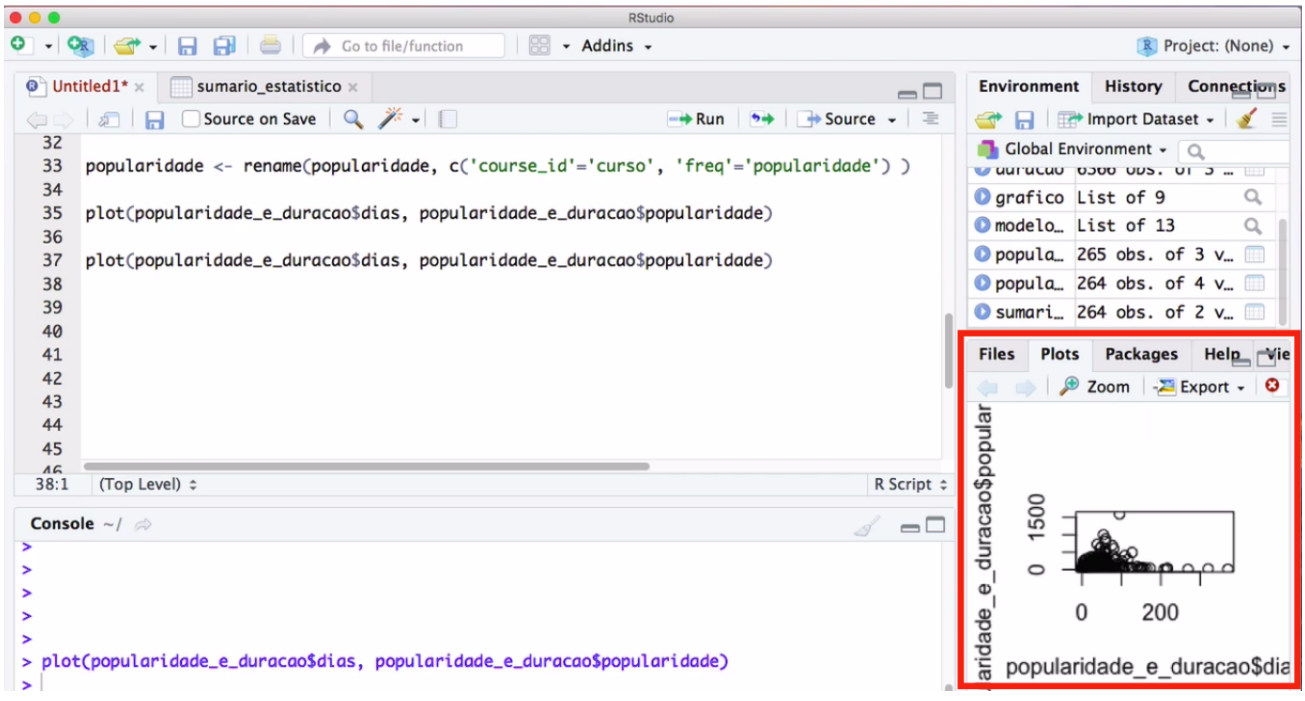
Agora que todas as informações de que precisamos estão em um mesmo objeto, estamos prontos para a análise. Faremos o estudo de correlação para informar à empresa se existe relação entre o número de matrículas de um curso e o tempo médio de duração até sua conclusão. Inicialmente, procuraremos um padrão visual, colocando essas informações em um gráfico para ver se encontramos alguma correlação a "olho nu".

Produziremos um gráfico usando o RStudio, com o comando básico para gráfico `plot` . Entre parênteses, especificaremos o banco de dados (`popularidade_e_duracao`), seguido de cifrão (`$`) e das variáveis que queremos visualizar:

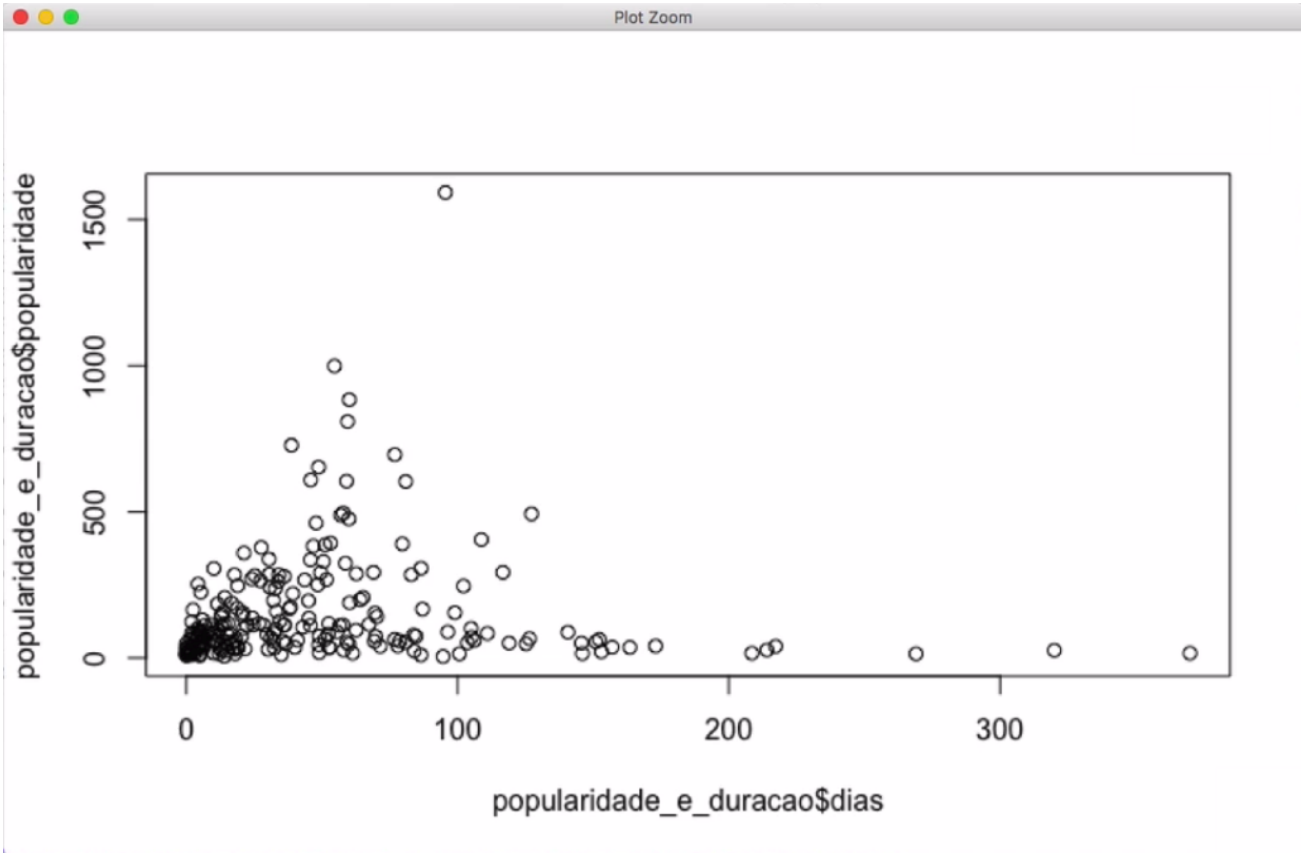
- `dias` , correspondente ao número de dias que cada curso leva para ser concluído em média;
- `popularidade` , referente ao número de matrículas de cada curso da amostra.

```
plot(popularidade_e_duracao$dias, popularidade_e_duracao$popularidade)
```

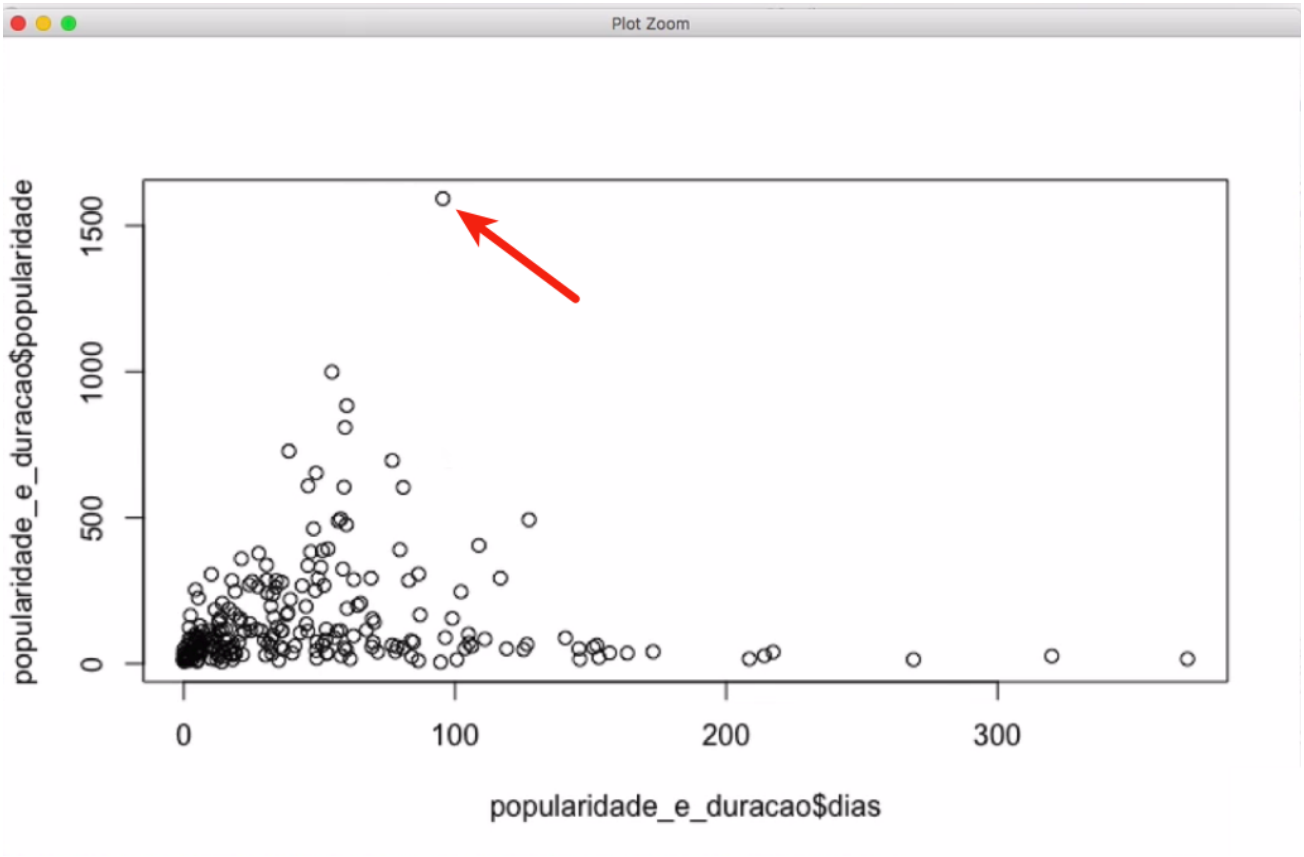
Ao executarmos a linha desse comando, na janela inferior direita visualizaremos o gráfico produzido por RStudio.



Porém, a visualização é ruim para análise, então clicaremos no botão de "Zoom", exportando-a para uma janela que podemos ampliar, obtendo a seguinte imagem:

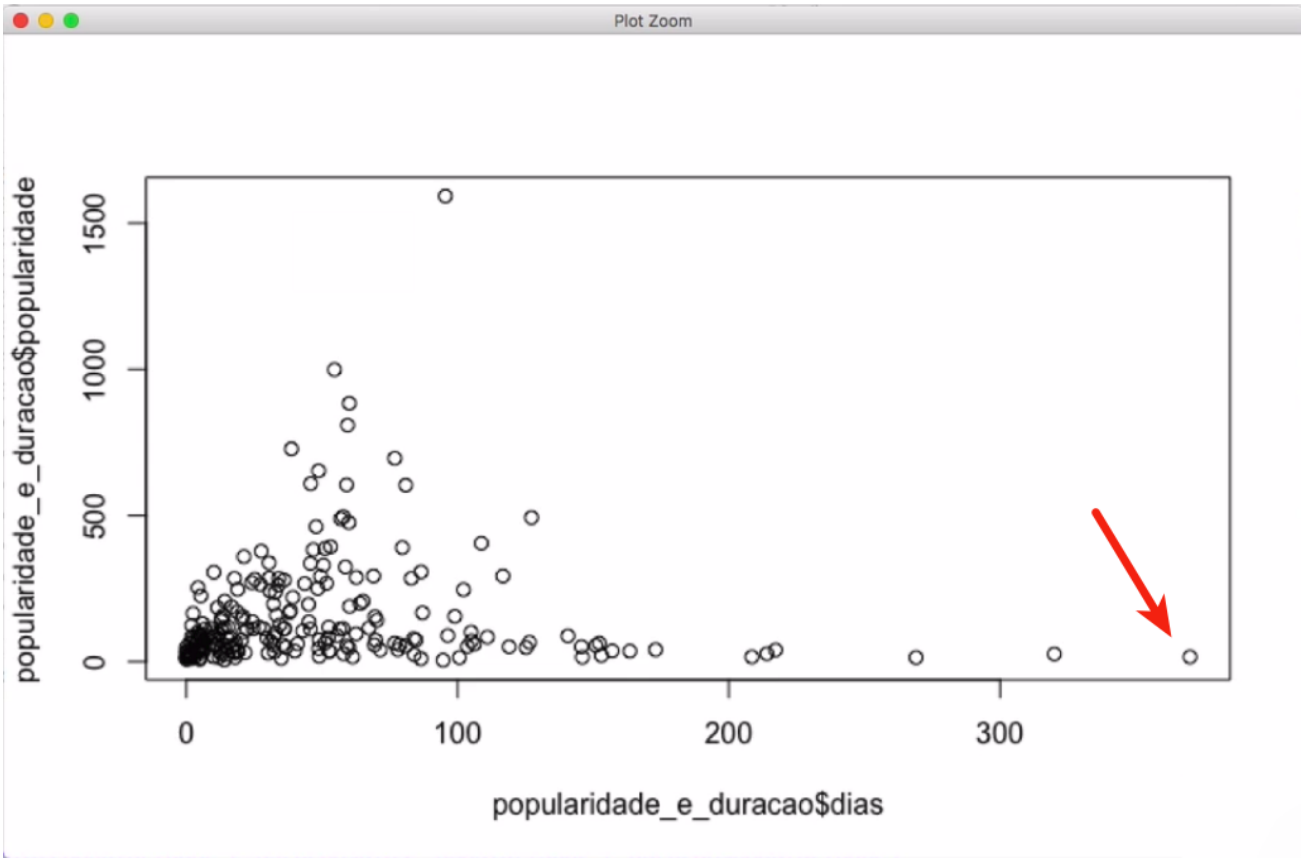


Agora, o tamanho do gráfico está apropriado para análise. No eixo horizontal, a medida vai de `0` para além de `300` dias. Como vimos anteriormente, há cursos que levam menos de `1` dia para serem completados, que se concentram próximo a `0`, e cursos que levam mais de `200` dias até a conclusão, à direita do valor. Há um ponto isolado, acima de `1500` do eixo vertical, e próximo a `100` do eixo horizontal.



A posição do ponto indica um curso com mais de `1500` alunos matriculados — supondo uma matrícula por aluno, pois não tem como sabermos se o curso foi feito repetidas vezes —, que levaram uma média de `100` dias para concluí-lo.

Partimos de `100` no eixo horizontal e fomos até o ponto localizado pouco acima de `1500` no eixo vertical, o que aponta o número de matrículas. Pra ficar claro, analisaremos os pontos discrepantes, como o que está no canto inferior direito.



Ele está localizado à direita do valor `300`, quase em `400`, no eixo horizontal, indicando um curso que os alunos levam, em média, mais de um ano para concluir. Se olharmos o número de matrículas, constataremos que não se trata de um curso popular, e tem poucos alunos matriculados.

São informações que poderemos levar à empresa, para que ela analise do ponto de vista qualitativo. Devemos ficar alertas para não deixar que os pontos discrepantes influenciem a análise que estamos fazendo, pois eles não nos fornecem uma visão apropriada do que é esperado que aconteça. Não é normal um aluno demorar mais de um ano para concluir um curso.

Até por isso, temos pouquíssimos casos nos extremos, como os casos em que os alunos levam mais de `200` dias para a conclusão do curso. A maioria leva menos de `100` dias para concluir um curso. E como vimos anteriormente, a média de dias, arrendondada, era de `48` dias, e a mediana era `8`. Isto significa que 50% das matrículas são concluídas em menos de `8` dias.

Agora, buscaremos uma medida um pouco mais quantitativa da correlação. A empresa pode nos indagar: "vocês nos deram essa dispersão, mas se um curso leva em média `50` dias para ser concluído, quantas matrículas podemos esperar para ele?". Vejam que em torno de `50`, no eixo horizontal, há uma dispersão grande no número de matrículas. Em uma análise visual, não obteremos o número de matrículas esperadas para um curso de maneira imediata.

Para isso, complementaremos a análise gráfica visual com modelos preditivos, estatísticos ou matemáticos, estimando um modelo matemático para prever qual seria a popularidade de um curso, com base em sua duração média. Fecharemos a janela de visualização do gráfico e criaremos um modelo, ao qual daremos o nome de `modelo.linear`, considerando que ele criará uma linha a partir dos pontos da dispersão, ajustando-os a uma reta que melhor prediz os pontos.

Não há garantias de que será uma boa predição ou previsão. Utilizaremos esses termos de maneira intercambiável, a reta que seria "a melhor das retas", por assim dizer. Criaremos o `modelo.linear` — e o importante é informar algo sobre a variável ou objeto no nome, então se preferirem podem dar outro nome — atribuindo (`<-`) a ele a função `lm()` que vem de "*linear model*".

Entre parênteses, o parâmetro da função será o que chamamos de "fórmula", cujo primeiro componente será o que queremos prever ou explicar, no caso, a `popularidade` . Então, passaremos a variável correspondente após especificarmos o banco em que está (`popularidade_e_duracao`):

```
modelo.linear <- lm(popularidade_e_duracao$popularidade~popularidade_e_duracao$dias)
```

Na sintaxe, utilizamos o acento til (**~**), lido no código como "modelado". No caso, a primeira variável (`popularidade_e_duracao$popularidade`) é **modelada** por outra (`popularidade_e_duracao$dias`). O acento til informa ao programa que uma variável será modelada por outra, e após acrescentarmos o til, inseriremos a variável `popularidade_e_duracao$dias` , que chamamos de **explicativa** ou **variável independente**, que apontará naquele ponto o quanto achamos que a outra variável irá medir.

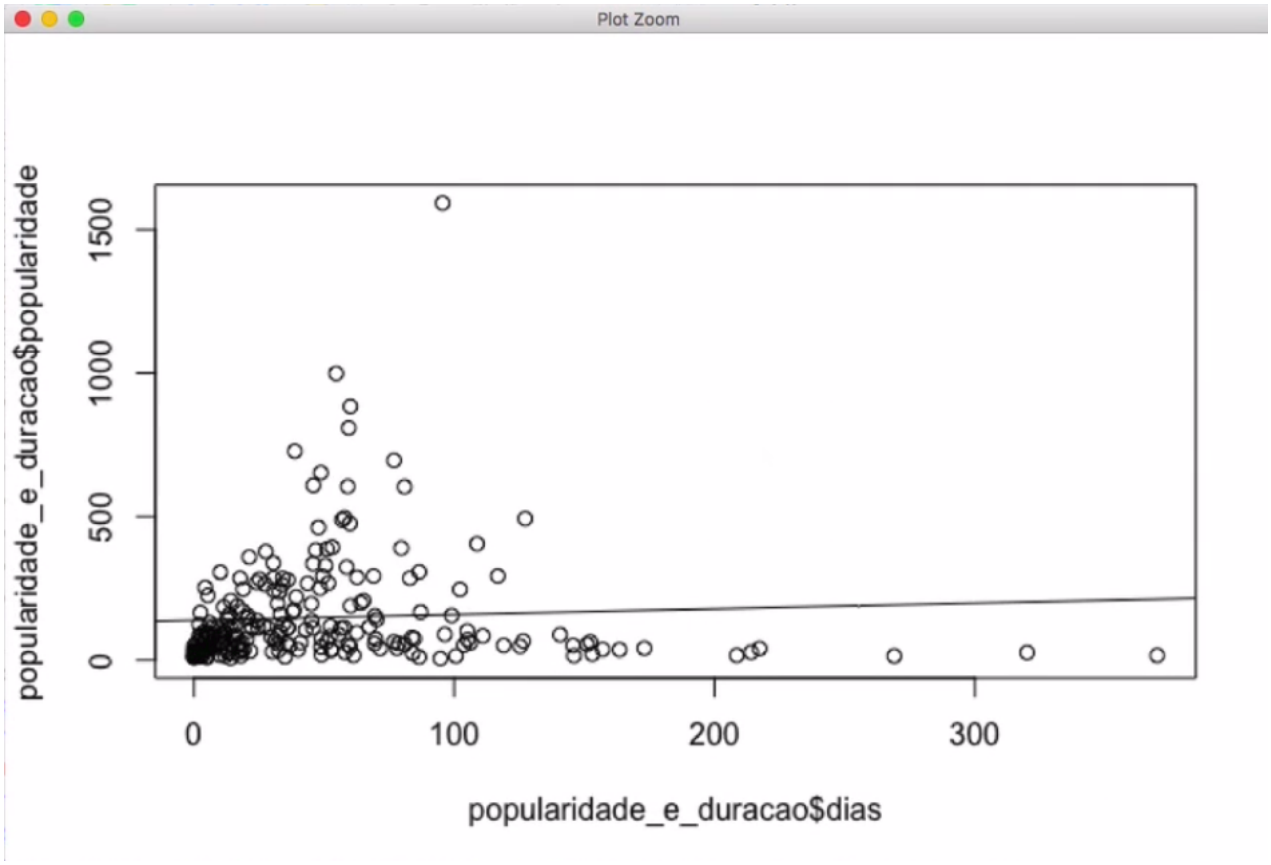
Ao ser executado o comando, é criado um modelo linear e nada mais é informado no Console, indicando que não houve erro de sintaxe. O programa não informa se a lógica estiver errada ou se o modelo não for o mais apropriado, e nós é que temos que descobrir isto. Visualizaremos o modelo linear no gráfico, posicionando o cursor em:

```
plot(popularidade_e_duracao$dias, popularidade_e_duracao$popularidade)
```

E clicando em "Run", após o qual colocaremos a linha que acabamos de criar, acima do gráfico. Adiante, exportaremos para ver o resultado, e utilizaremos o comando `abline` e, entre parênteses, colocaremos a modelagem:

```
abline(lm(popularidade_e_duracao$popularidade~popularidade_e_duracao$dias))
```

Assim, estaremos solicitando ao programa uma linha com o modelo que criamos. Ao executá-la, na janela inferior direita teremos uma linha no gráfico. Clicaremos em "Zoom" e expandiremos para visualizá-la melhor:



É o mesmo gráfico de dispersão dos pontos, agora com uma reta ajustada, que fornece uma "previsão" do número de matrículas que um curso deve ter ou tem na amostra, de acordo com a média dos dias que os alunos levam para concluí-lo. Com a reta, poderemos passar para a empresa uma primeira estimativa. Lembrando que ela perguntou qual o número de matrículas esperado para um curso que os alunos levam, em média, `50` dias para concluir.

Não sabíamos qual era o número aproximado de matrículas, devido a dispersão dos pontos. Com a reta ajustada poderemos raciocinar de maneira mais precisa. Então, analisaremos o gráfico, na região próxima a `50` no eixo horizontal (referente a `dias`), subindo até chegar na reta. Ao atingi-la, verificaremos que sua medida no eixo vertical está um pouco abaixo de `250` . Assim, poderemos responder a questão da empresa da seguinte forma:

> "Para um curso que leva, em média, `50` dias para ser concluído, pode-se esperar aproximadamente `250` matrículas."

O problema do modelo linear é que os pontos não são ajustados em nenhuma região do gráfico. A reta passa por cima de uma concentração de pontos no início do gráfico, e há uma quantidade relevante de pontos na região entre `50` e `100` dias, com cerca de `300` matrículas; vamos estimar assim, por onde a reta não passa.

Notem também que, no final do gráfico, a linha passa por cima de alguns pontos que representam cursos pouco populares, com poucas matrículas, e que levam muitos dias até a conclusão. Analisando visualmente, perceberemos que a dispersão é bastante "não linear". É difícil fazer o ajuste da dispersão por uma reta. Há uma concentração no início, à esquerda do gráfico, em seguida há uma aparente ascensão, uma concentração e popularidade maiores dos cursos, a medida que o número de dias para serem concluídos aumenta.

Depois há uma queda, formando uma "montanha" que o modelo linear não capta. Essa é uma informação importante para levarmos à empresa, que ajusta muito bem os valores médios. Então, muito provavelmente, na média e na mediana ela faz um bom trabalho, mas a real dinâmica da correlação não é captada nos extremos, na não linearidade, que veremos adiante.