04

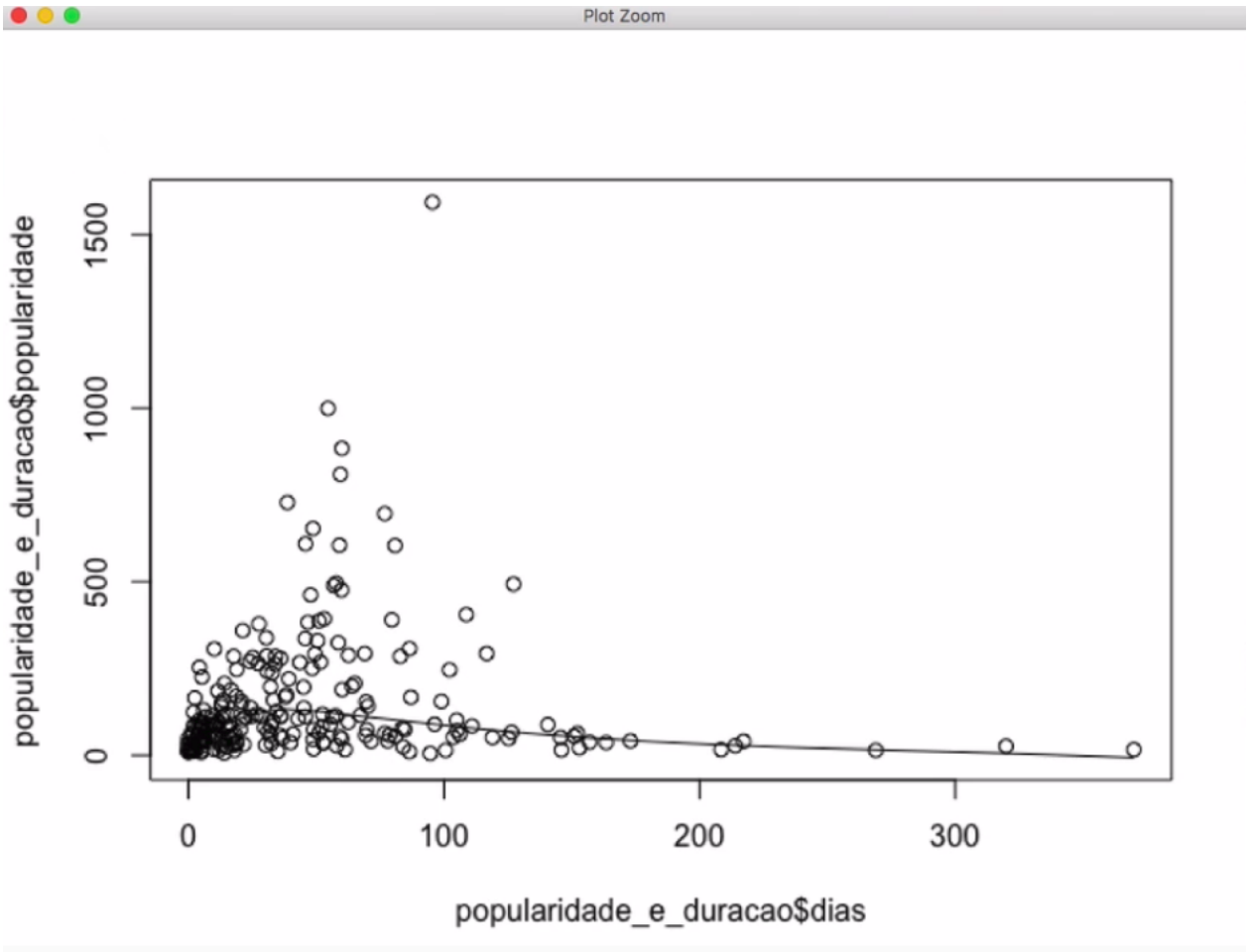
Modelo não linear

Transcrição

Percebemos que o modelo **linear** não é a melhor alternativa, mesmo sendo útil em diversas situações e, por ser simples, é sempre a primeira abordagem em uma análise. No entanto, poderemos aprimorar o que estamos trabalhando, transformando-o em um modelo **não linear**. Para propor um gráfico que ajuste uma curva — e não uma reta — nesse mesmo gráfico, utilizaremos a função "*scatter dot smooth*" (`scatter.smooth`), em português, "dispersão ponto suavização":

```
scatter.smooth(popularidade_e_duracao$dias, popularidade_e_duracao$popularidade)
```

A função foi desenvolvida como no gráfico anterior, em que utilizamos somente as variáveis que queremos analisar. No caso, `dias` e `popularidade` , do banco `popularidade_e_duracao` . Ao executarmos esse comando, é criado um gráfico na janela inferior direita. Clicaremos no botão "Zoom" e expandiremos a janela para melhor visualização, e obteremos a seguinte imagem:

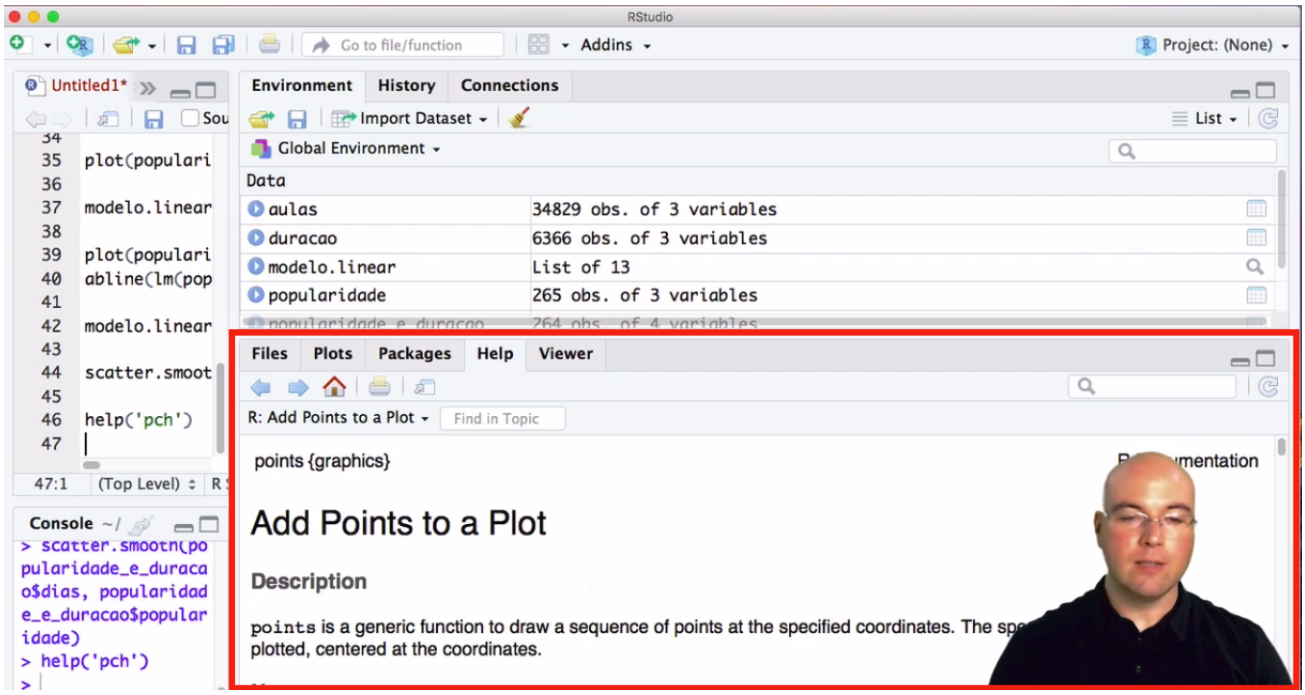


Bacana! Notem que o gráfico é igual ao anterior, com a mesma dispersão de pontos, porém com um modelo não linear. Anteriormente, ajustamos a reta que saía aproximadamente de `138` do eixo vertical e crescia eternamente no gráfico. Agora, temos um modelo não linear ajustado - Reparem que há uma concentração de pontos no início do gráfico, e a partir de `100` do eixo horizontal, ela se dilui. Além disso, a curva é suave, sem quebras ou triangulação. Ela não cresce ou decresce abruptamente em ponto nenhum, como no modelo anterior.

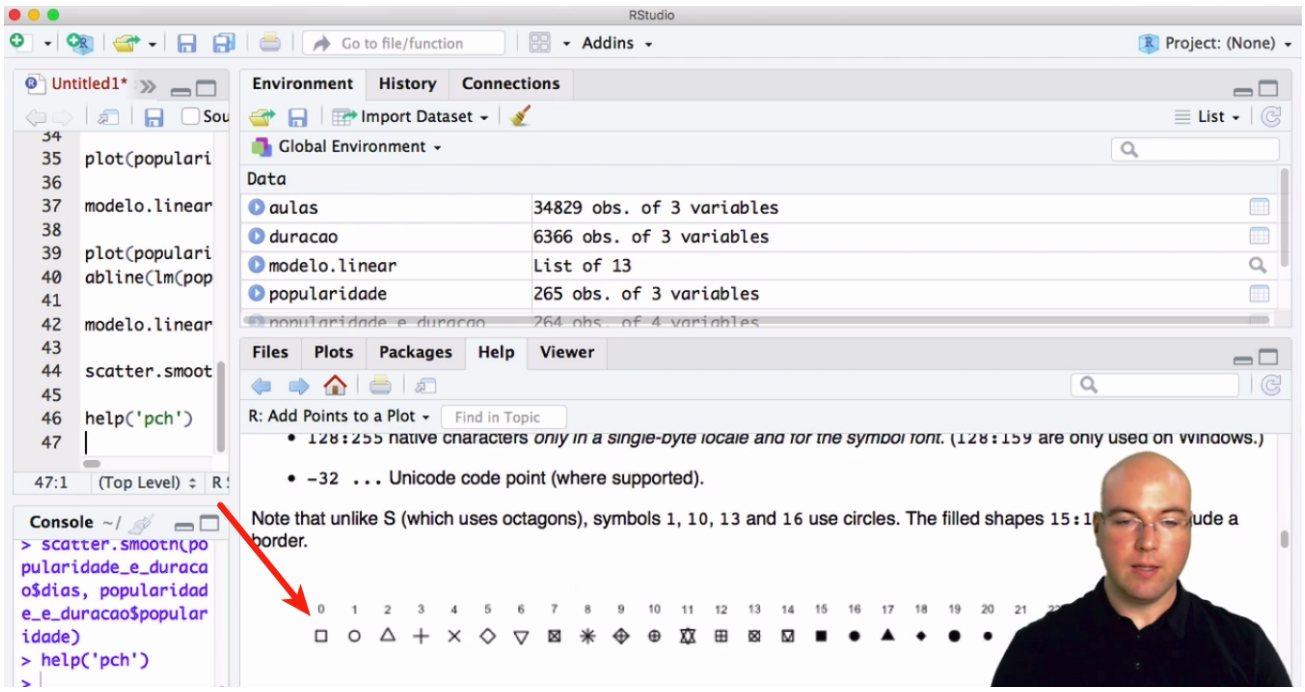
Essa é a ideia de "suavização" da curva. O problema é que há uma concentração de pontos próxima a `0` e o símbolo utilizado — um círculo oco — prejudica a visualização. Tentaremos melhorá-lo, alterando-o. Para isso, adicionaremos um parâmetro chamado `pch`, e escolheremos um novo símbolo no R Script, por meio da função `help` com o `pch` especificado entre parênteses:

```
help('pch')
```

Ao executarmos o comando, na janela inferior direita visualizamos a documentação de ajuda do comando que solicitamos. Poderemos ajustar o tamanho das janelas para melhor visualização:



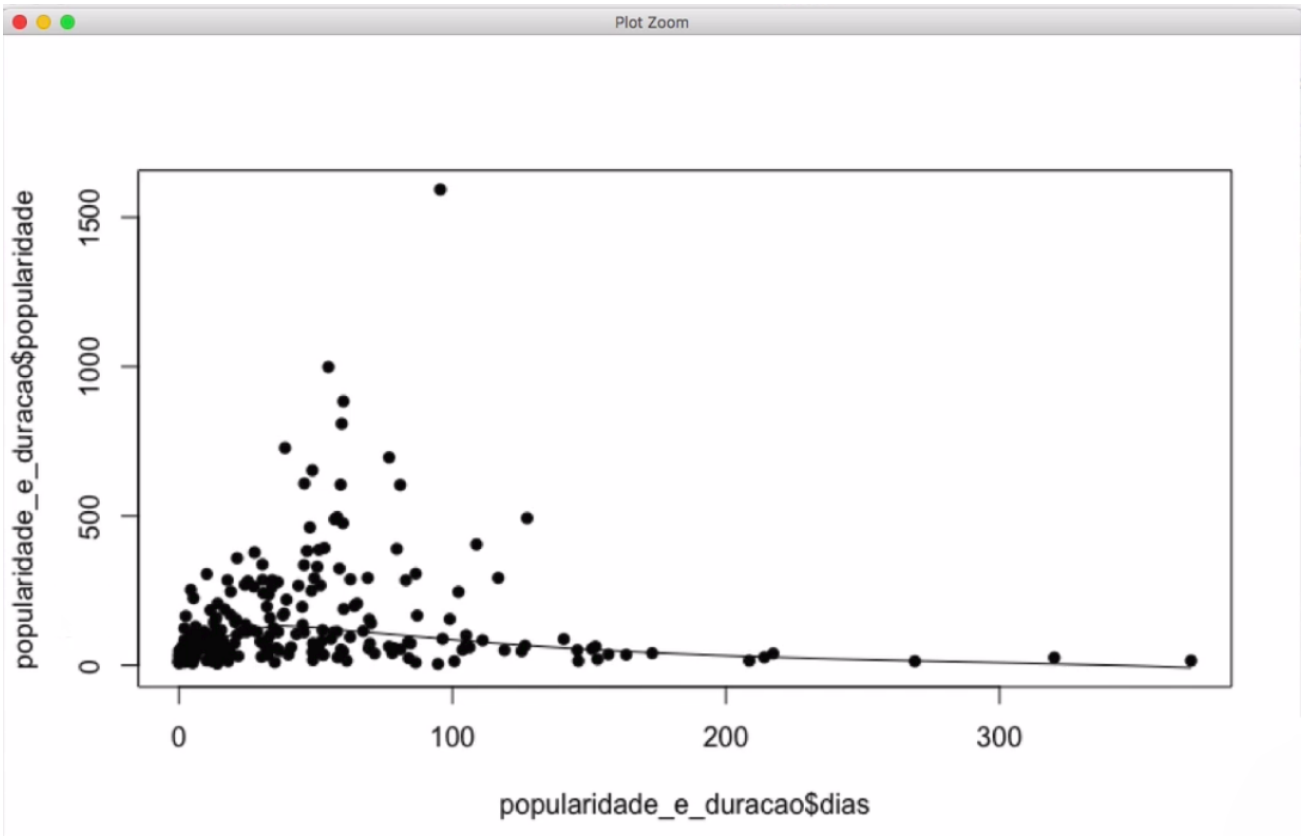
Nela, há uma descrição do pacote "*Add Points to a Plot*" (em português, "Adicionar Pontos em um Gráfico"), mas o que nos interessa são as opções de pontos disponíveis. Para cada número, há um símbolo diferente, e se colocarmos `pch = 0`, o desenho será um quadrado. O desenho padrão ou *default* é o círculo oco, que corresponde ao número `1`. O `2` é um triângulo, e assim segue.



Pode-se escolher qualquer um deles; selecionaremos o número `16`, correspondente a um círculo sólido, preenchido com a cor preta. De volta à linha de `scatter.smooth`, acrescentaremos `pch=16` ao final do código:

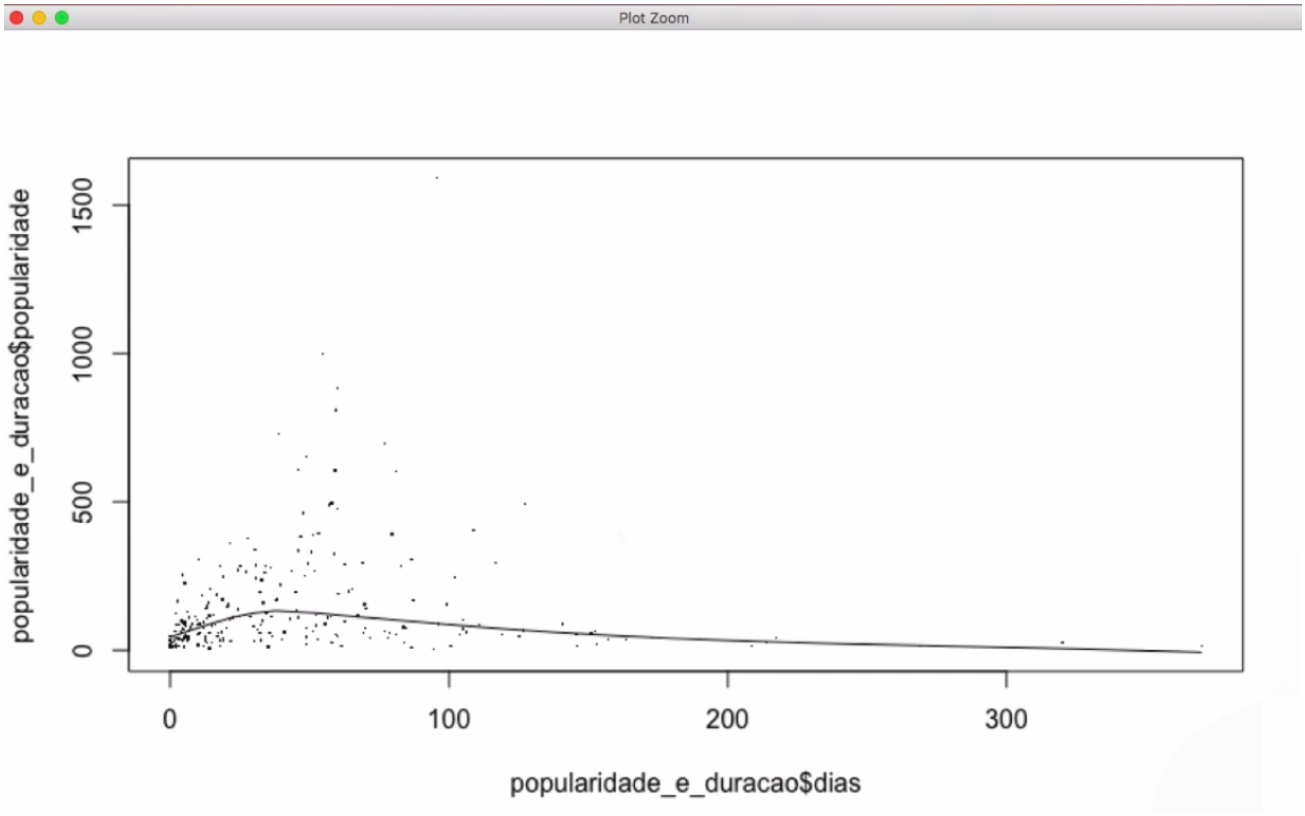
```
scatter.smooth(popularidade_e_duracao$dias, popularidade_e_duracao$popularidade, pch=16)
```

Executaremos o comando, clicaremos no botão "Zoom" da janela inferior direita e expandiremos a visualização do gráfico, para obter a seguinte imagem:



Notem que o símbolo utilizado para cada ponto mudou; melhorou um pouco. Conseguimos observar melhor a curva na região inicial do gráfico, porém isto ainda não é o ideal. Na documentação de ajuda, não encontramos alternativas de símbolos que proporcionem uma boa visualização da curva. Poderemos brincar com o parâmetro e colocar outros símbolos, como um ponto (`.`). Como é um caractere, e não um número, devemos colocá-lo entre aspas simples (`'`).

A ideia é que cada informação, cada ponto, ocupe o menor espaço possível para evitar a sobreposição de valores. Queremos um símbolo pequeno, e por isso escolhemos o ponto (`.`). Executaremos o comando, aplicaremos "Zoom" e ampliaremos a janela:



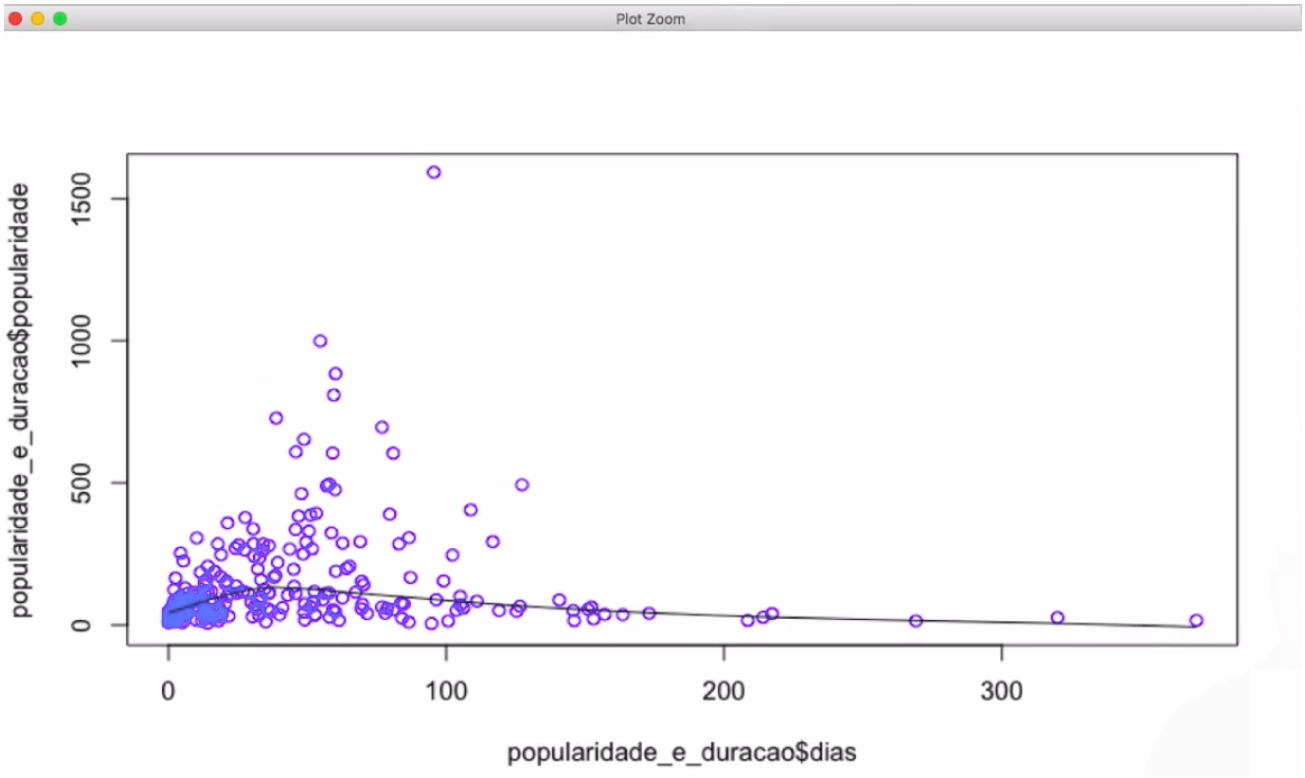
Desta vez conseguimos ver nitidamente a curva suave, seu início e o comportamento do meio até o final do caminho. Observem que o programa utilizou o ponto (`.`), reduzindo a sobreposição de símbolos e permitindo a visualização da concentração inicial e a dispersão na região intermediária.

À direita do gráfico não tínhamos muitos problemas. Conseguíamos ver os símbolos e a curva, mas a dificuldade estava à esquerda, onde há sobreposição dos valores. Agora, conseguimos visualizar melhor. O problema é que o símbolo ficou pequeno demais, e uma região grande do gráfico ficou em branco, dificultando a análise de alguns pontos.

A solução é tentarmos outro valor e acrescentar cores, que podem ser alteradas também. Colocaremos, por exemplo:

```
scatter.smooth(popularidade_e_duracao$dias, popularidade_e_duracao$popularidade, pch=21,
               col='blue')
```

Assim, especificamos que usaremos o símbolo `21` , referente ao círculo oco, na cor azul (`blue`). Ao executarmos, aplicarmos o "Zoom" e expandirmos a janela, teremos a seguinte imagem:



Agora, o círculo oco está com a borda azul, o que nos parece uma solução melhor. Conseguimos informar onde está a distribuição dos pontos, e a sobreposição (*overlapping*) de valores não está cobrindo a visualização da curva. No início do gráfico, onde há a concentração de pontos, conseguiremos ver o começo da curva e acompanhamos seu comportamento integralmente. Há uma confirmação da intuição inicial de que o modelo linear não era o mais apropriado para essa análise.

Visualmente, identificamos um padrão complexo, não linear, em comparação com antes, quando notávamos uma concentração no início diferente. A partir de agora, o modelo não linear com a curva "suavizada" explica o que acontece no gráfico. Por exemplo, de `0` ao trecho entre `40` e `50` dias, a popularidade dos cursos cresce, e então começa a decrescer até o final do espectro, ao fim do qual a curva se ajusta muito melhor aos pontos do que à reta anterior, que passava por cima deles e crescia sem parar, ficando distante dos símbolos.

Ou seja, esse modelo de curva é muito melhor, e traz uma informação muito relevante para levarmos à empresa. Ela confirma as estatísticas que encontramos anteriormente. Lembram que a média de duração era em torno de `48` dias? Conferimos que há uma concentração nesse ponto do eixo horizontal, e que a popularidade dos cursos é maior nessa região do gráfico.

Há casos discrepantes (*outliers*), como o curso que dura em torno de `100` dias, sendo também o mais popular da amostra, com mais de `1500` matrículas. Há outros cursos populares, com o número de matrículas entre `500` e `1000`, que demoram de `50` a `100` dias até a conclusão. A maior concentração de cursos leva menos de `1` dia ou pouquíssimos dias — entre `0` e `20` — até sua conclusão.

No entanto, o grande *insight* são os cursos que levam de `48` a `50` dias até a conclusão. Poderíamos passar a informação para a empresa da seguinte forma:

> "Os cursos com duração média entre `48` e `50` dias são os mais populares. Eles possuem um número maior de matrículas".

A partir disto, a empresa poderá realizar uma análise qualitativa de forma direcionada, economizando custos, tempo de funcionários e esforço, pois trata-se de uma informação bastante relevante.