UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

Trabalho de Estatística III

IVAN CUNHA VIEIRA

VITÓRIA, ES
2024

1 Introdução

O dataset escolhido para o trabalho contém informações acerca dos salários de 375 empregados de uma empresa. Cada linha do dataset representa um empregado diferente, que são descritos pelos seus respectivos salários anuais (USD) e anos de experiência. O objetivo do presente trabalho é o de tentar relacionar tais variáveis.

A princípio, foi feita uma análise descritiva, no intuito de organizar e resumir os importantes aspectos e características de cada uma das variáveis. Após a análise, foi efetuada uma tentativa de um modelo de regressão linear com base nos anos de experiência dos funcionários e seus devidos salários.

2 Análise Descritiva

Para a análise descritiva, apresenta-se abaixo a tabela contendo as medidas resumo relativas às variáveis estudadas (salário anual e anos de experiência), além de gráficos para facilitar a análise.

Observa-se, por meio da tabela das medidas-resumo, uma alta variação nos salários anuais dos funcionários desta empresa, no intervalo de [30000; 250000]. A variável que diz respeito aos anos de experiência apresenta menor variabilidade, no intervalo de [0; 25], o que é justificado pelas maiores limitações da idade humana em comparação às limitações salariais. Pela mesma ótica, percebe-se que a variação dos anos de experiência é bem alta quando considerados o tempo de vida e trabalho das pessoas.

Ao examinar os gráficos de densidade das variáveis, é perceptível a semelhança na tendência de comportamento dos salários e dos anos de experiência, uma vez que, ao mesmo tempo em que há uma baixa densidade de salários próximos aos US$ 250.000,00, também há uma menor densidade de funcionários próximos de 25 anos de experiência, da mesma forma que uma maior densidade de salários próximos aos de US$ 30.000,00 é acompanhada de funcionários próximos a 0 anos de experiência, além de distribuições semelhantes para os demais salários e anos de experiência. Esses comportamentos indicam uma possível correlação entre as variáveis. É possível reparar também que nenhum valor aberrante foi apresentado nos gráficos de boxplot.

Table 1: Medidas-resumo

	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo	D.P.
Experiência	0	4	9	10.0308	15	25	6.557
Salário	30000	55000	95000	100670.2413	140000	250000	48079.583

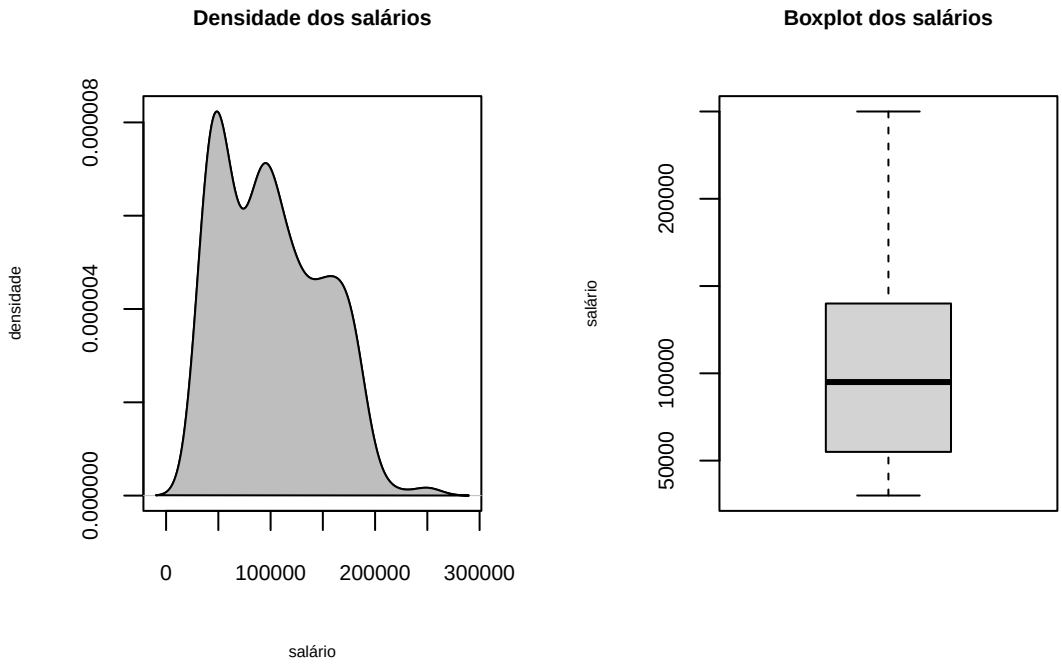

Figure 1: Densidade e boxplot da variável salário

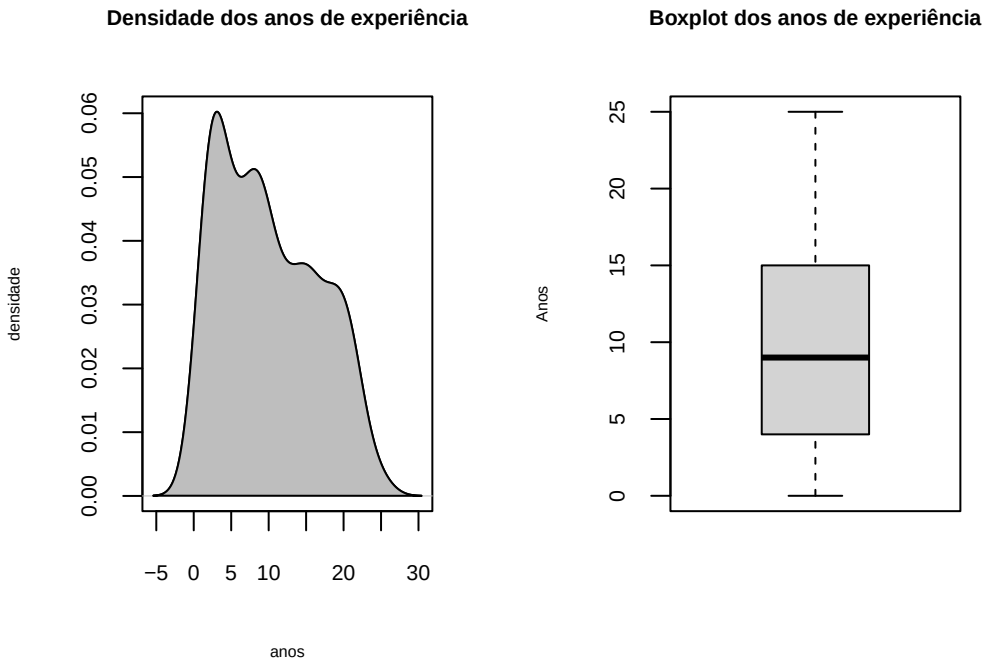

Figure 2: Densidade e boxplot da variável anos de experiência

Adiante, segue-se com o gráfico de dispersão das variáveis. É evidente, através deste gráfico, que há uma relação positiva forte entre os salários e os anos de experiências dos funcionários, fato evidenciado também pela correlação calculada em 0.9309, de forma que maiores anos de experiência influenciam diretamente nos salários dos funcionários. Desse modo, após verificar os comportamentos das variáveis, é possível ajustar um modelo normal linear da forma:

$$salario_i = \alpha + \beta experiencia_i + e_i \ , \ i = 1, \ ..., \ 373.$$

Figure 3: Dispersão das variáveis salário e anos de experiência

Seguindo o ajuste do modelo, chegou-se nos resultados da tabela abaixo. O erro padrão dos resíduos (raíz do quadrado médio dos resíduos) foi calculado como 17580 com 371 graus de liberdade, e seu R^2 foi calculado como 0.8663. Dessa forma, conclui-se que, pelo modelo, cerca de 86% da variação dos salários dos funcionários desta empresa se relaciona linearmente com os anos de experiência, indicando que outras informações não consideradas, como gênero, idade, grau educacional e qual trabalho o funcionário exerce (informações presentes no dataset), podem ser os fatores que resultem na variação dos outros 14% dos dados.

Como se observa no gráfico de envelope a seguir, há diversos pontos fora dos limites do envelope. Essa observação, juntamente ao R^2 abaixo de 90%, traz indícios de que o modelo utilizado para modelar estes dados não é, provavelmente, o mais adequado, apesar de conseguir explicar parte do comportamento das variáveis.

Table 2: Ajuste do modelo normal linear

	Estimativas	Erro Padrão	t-valor	p-valor
α	32199.4985	1665.4921	19.3333	<2e-16
β	6826.0289	139.0342	49.096	<2e-16

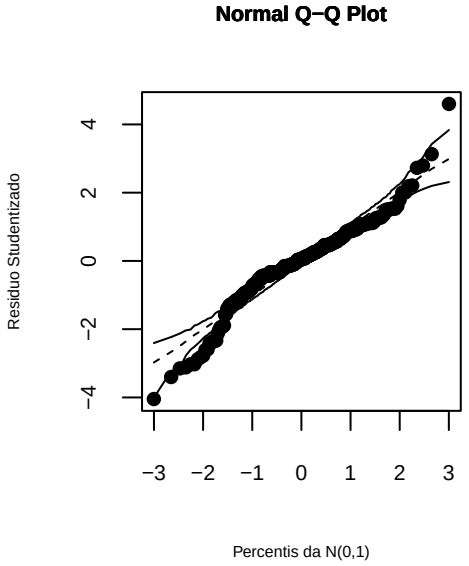

Figure 4: Envelope do ajuste normal

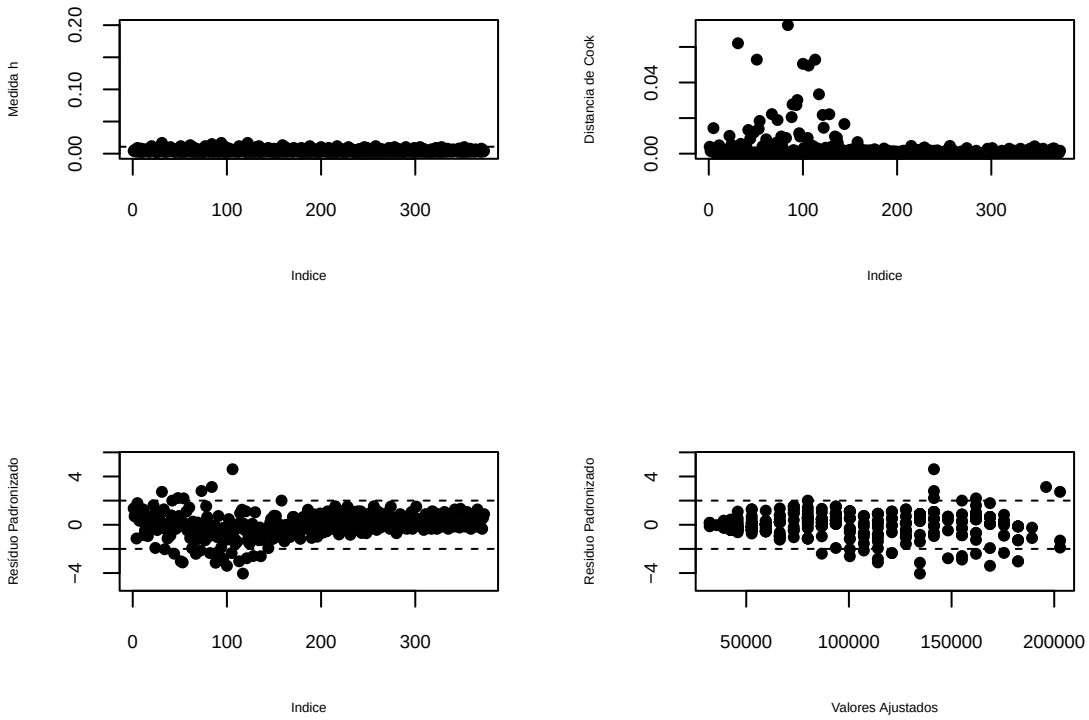

Figure 5: Diagnóstico do ajuste normal

Corroborando com o que foi analisado através do gráfico de envelope, os gráficos de diagnóstico também apresentam informações que indicam que o modelo utilizado não é o mais adequado para os dados. Apesar de não apresentar pontos de alavancagem e apresentando pontos de influência pouco relevantes (pela distância de Cook, pontos muito longe de 1), o modelo apresenta diversas observações que demonstram haver heterocedasticidade no conjunto dos dados, o que vai contra a ideia do modelo normal linear. No caso dessas observações, se torna difícil enumerar cada uma, pois são muitas, e remover muitos dados para conseguir encaixar a modelagem não seria produtivo. Portanto, para o caso do problema da heterocedasticidade, seguiu-se com a transformação de Box-Cox.

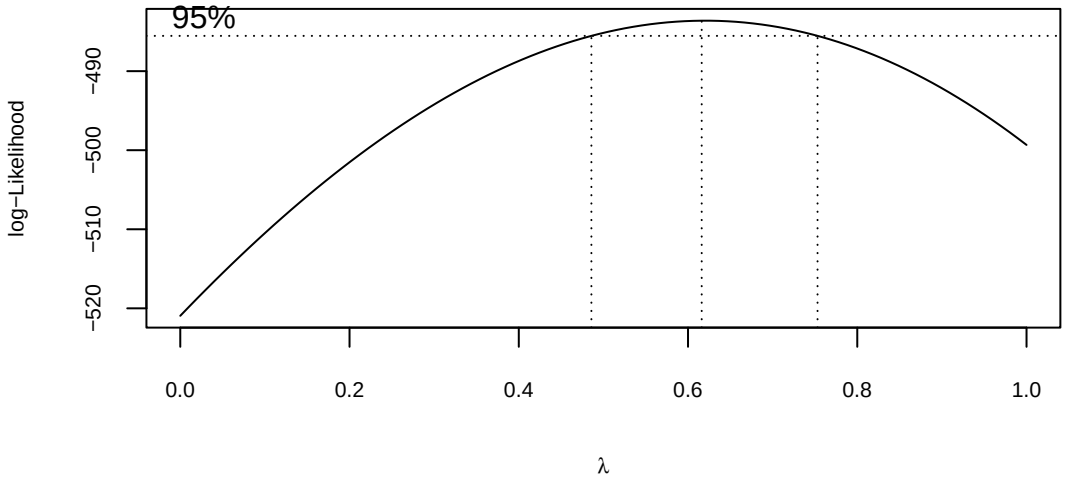

Figure 6: Transformação de boxcox

Foi possível perceber, pelo gráfico apresentado, que o valor de $\lambda = 0.5$ se encontra entre os possíveis valores de lambda. Este é um valor interessante para usar, pois, na transformação de Box-Cox, ele apresenta uma transformação mais fácil de interpretar, de tal forma que $Y' = \frac{\sqrt{Y} - 1}{0.5}$, o que também facilita seu uso.

Prosseguindo com a transformação, percebe-se, através dos gráficos de envelope e dos gráficos de diagnóstico do novo modelo, que somente a transformação não foi suficiente para resolver os problemas de não normalidade e de heterocedasticidade presentes na modelagem anterior, apesar dos dados apresentarem melhores resultados referentes à normalidade, mas ainda com pontos fora do envelope. Essas observações demonstram novamente que a modelagem escolhida foi inadequada para o conjunto de dados.

Ao testar outras mudanças na modelagem, foi encontrada uma modelagem que abrangia outra covariável, sendo esta o tipo de trabalho exercido pelo funcionário, tal que o modelo encontrou $R^2 = 0.966$. Dessa maneira, é possível interpretar que é importante, para esse conjunto de dados, modelagens que envolvam múltiplas covariáveis, como a seguinte:

$$salario_i = \alpha + \beta_1 experiencia_i + \beta_2 trabalho_i + e_i \ , \ i = 1, \ ..., \ 373.$$

```
## [1] 0.9294028
```

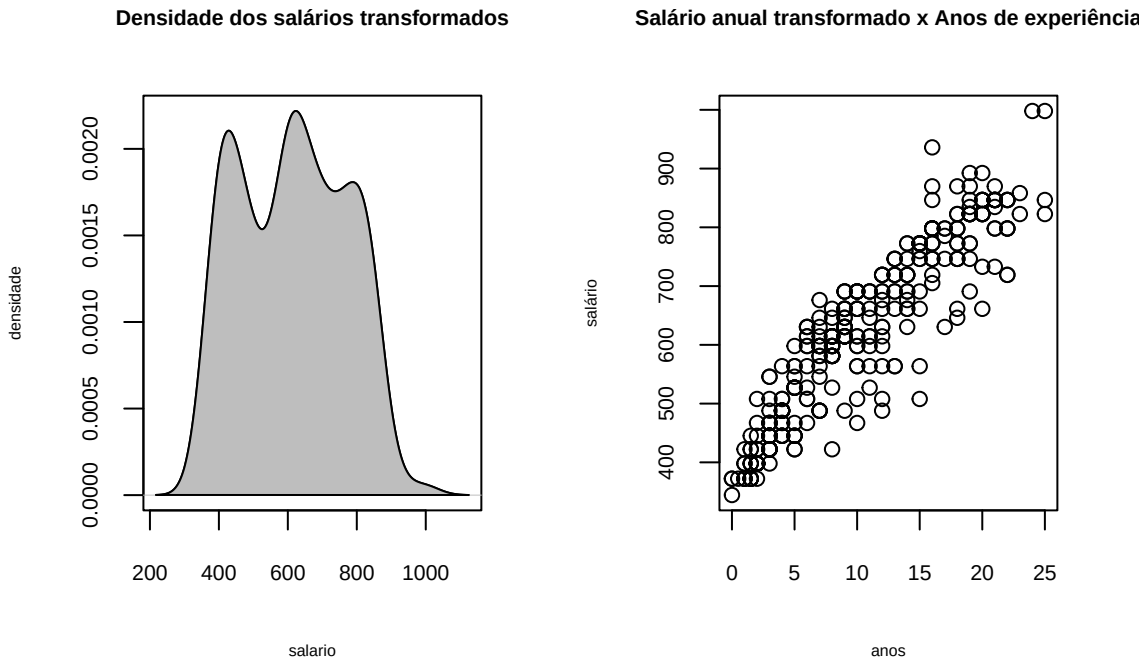

Figure 7: Densidade e dispersão da variável transformada

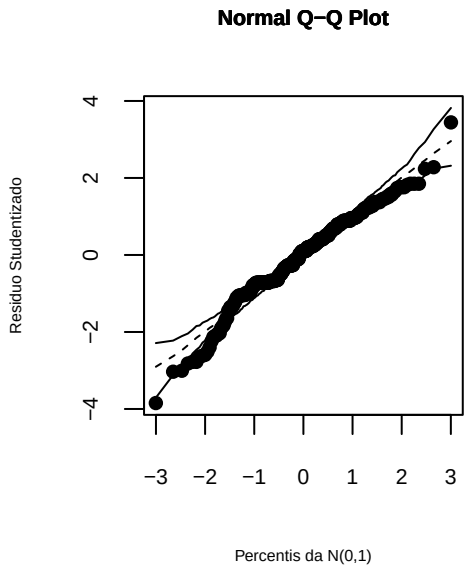

Figure 8: Envelope do ajuste normal da variável transformada

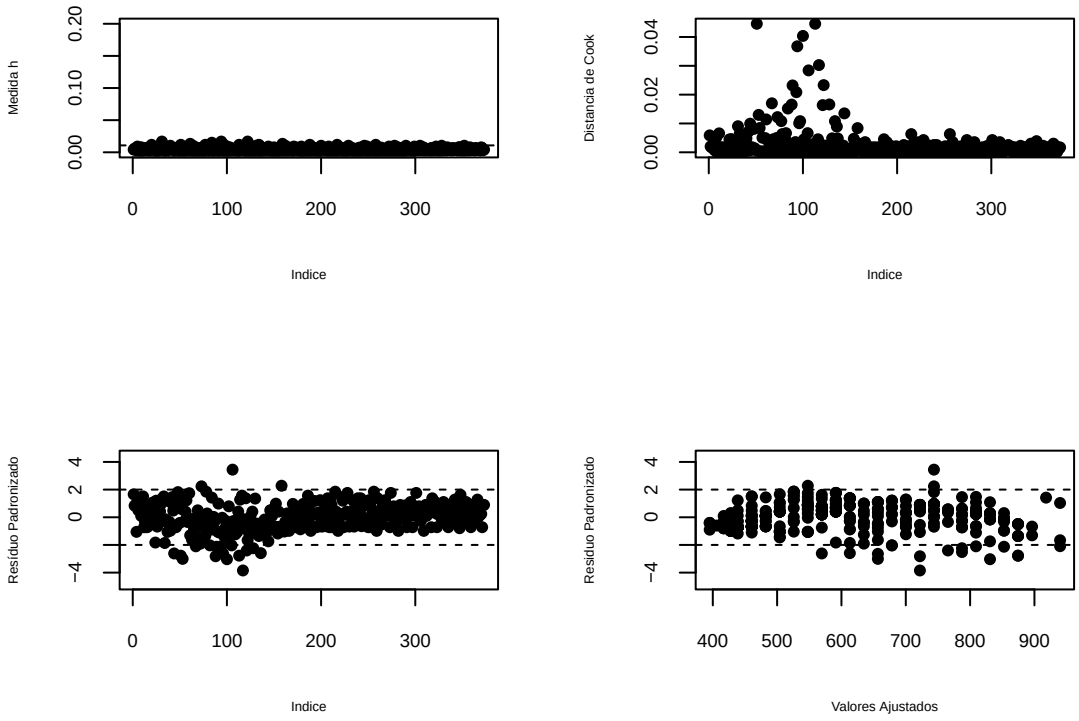

Figure 9: Diagnóstico do ajuste normal da variável transformada

3 Conclusões

A partir do estudo da correlação entre salários anuais de funcionários de uma empresa e seus respectivos anos de experiência, foi possível observar que, conforme mais anos de experiência o funcionário tiver, o salário desse funcionário também será maior, apresentando assim uma forte correlação positiva. Além disso, pela variação dos anos de experiência ser numericamente baixa, percebeu-se que mesmo alguns valores próximos causavam mudanças significativas na variável resposta. A dispersão dos dados também apresentou um formato de funil, o que trouxe indícios de uma possível heterocedasticidade no momento da modelagem, que se provou verdade posteriormente.

Os dados apresentarem diversas observações atípicas na modelagem, principalmente nos gráficos dos resíduos padronizados. Dessa forma, não foi possível determinar todas as observaçoes atípicas, uma vez que seria improdutivo enumerar tantas observações, provando que o modelo escolhido para o estudo desses dados foi inadequado. O gráfico de envelope também trouxe as mesmas conclusões sobre a modelagem. Ainda assim, por conta dos fortes indícios de heterocedasticidade, utilizou-se do método de Box-Cox para procurar uma transformação que tornasse o modelo homocedástico, o que se provou ineficiente quando analisou-se novamente os gráficos de envelope e de diagnóstico. Percebeu-se, ao fim, que, além da modelagem não ter sido a mais adequada, é provável que o uso de apenas uma covariável no modelo tenha sido insuficiente, se fazendo necessário estudar modelos que acomodem mais covariáveis do banco de dados.