Questão 9

Introdução

Para uma amostra aleatória de 220 executivos (Foster, Stine e Waterman, 1998, pgs. 180-188), queremos modelar o salário anual (em mil USD). O salário será modelado pelas variáveis explicativas: sexo, anos de experiência no cargo e posição na empresa. Esperamos, dado as caraterísticas da época, que os homens ganhem mais que as mulheres e que posições mais altas sejam melhor remuneradas.

Para tal análises será usada a metodologia dos modelos lineares homocedásticos, metodologias de verificação da qualidade do ajuste e comparação de modelos apropriado, veja Azevedo (2019). Todas as análises serão feitas com auxílio computacional do R.

Análise descritiva

Como suspeitamos, há um relação positiva entre a posição na empresa e o salário, como observado nos gráficos de dispersão da figura 1 e uma reta parasse que se ajusta bem aos dados. Porém o tempo de experiência não parece explicar bem o salário, nos gráficos da figura 2 observamos que uma variação no tempo de experiência não provoca uma variação no salário.

Pelos boxplots da figura 3 e pela tabela 1, parece que realmente há uma tendência dos homens ganharem mais que as mulheres. Neles observamos que as médias e as medianas do salário dos homens são maiores que os das mulheres.

Considerando as análises descritivas, modelar o salário pelo sexo, tempo de experiência e posição na empresa parece adequado.

Análise inferencial

O primeiro modelo proposto será um modelo de regressão linear homocedástico que considera um coeficiente de cada variável quantitativa (experiência e posição) relativo ao sexo e também um coeficiente relativo ao sexo. As variáveis quantitativas serão centradas no zero, para uma maior interpretabilidade. Em seguida se necessário ele será aprimorado e reduzido.

Os modelos serão ajustados segundo o método dos mínimos quadrados ordinários e a significância de casa parâmetro será testada a partir de uma estatística t, veja Azevedo(2019).

Modelo inicial

Seja Y_{ij} o salário do j-ésimo executivo, para i = 0 (mulher), j = (1,2,. . . ,75) e para i = 1 (homem), j = (1,2,. . . ,145). Temos então o seguinte modelo:

$$Y_{ij} = \alpha_i + \beta_{i0} * (x_{0ij} - \bar{x}_{0i}) + \beta_{i1} * (x_{1ij} - \bar{x}_{1i}) + \varepsilon_{ij}$$

onde x_{0ij} é a posição do j-ésimo executivo dado o sexo, x_{1ij} é a experiência do j-ésimo executivo dado o sexo. E α_i é o efeito do sexo no salário, β_{i0} é o efeito da posição, dado o sexo, no salário e β_{i1} é o efeito da experiência, dado o sexo, no salário. Com a parte aleatória sendo ε_i.

Esse modelo é válido sobre as seguintes suposições: (i) $\varepsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$; (ii) as observações são independentes; (iii) e a variância é constante.

O modelo ajustado resulta no parâmetro da tabela 2, alguns resultado parecem condizentes com o da análise descritiva, outros são um pouco surpreendentes. Como observado a posição na empresa tem uma relação positiva com o salário e é significativa para o modelo. O efeito do sexo também foi significativo e o efeito para mulher é maior, ou seja é esperado que mulheres localizadas na

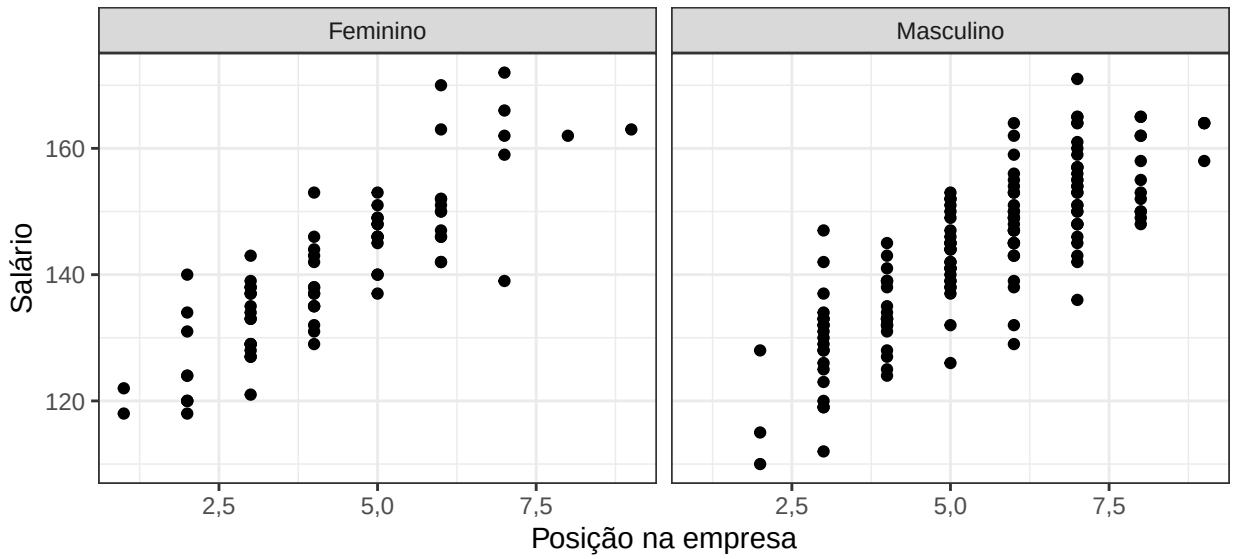

Figura 1: Gráfico de dispersão entre o salário e posição separado pelo sexo

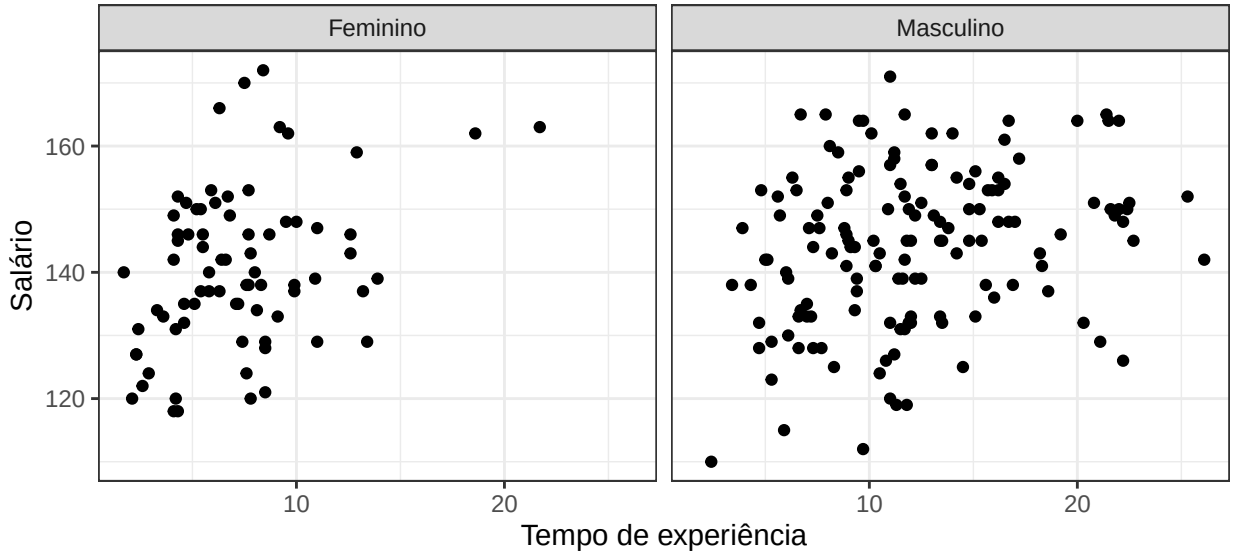

Figura 2: Gráfico de dispersão entre o salário e experiência separado pelo sexo

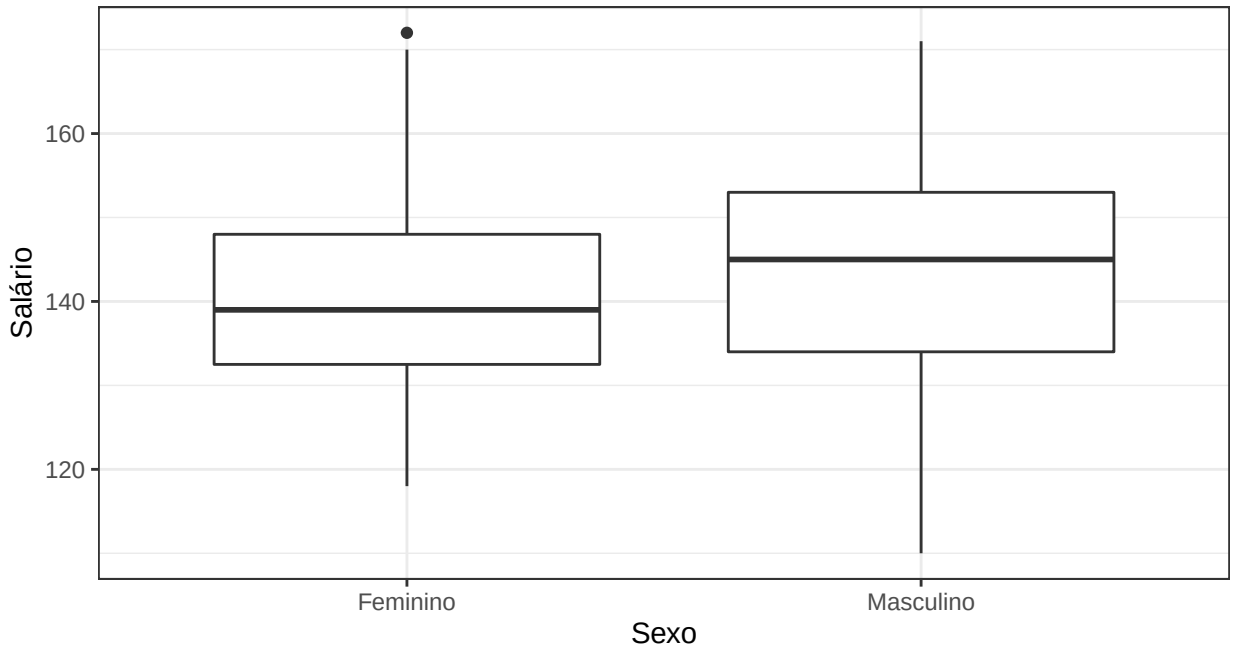

Figura 3: Boxplot do salário pelo sexo

Tabela 1: Estatística do salário para homens e mulheres

Sexo	Média	Desvio Padrão	Coeficiente de Variação
Feminino	140,4667	12,49577	8,90%
Masculino	144,1103	12,39406	8,60%

média feminina(posição e experiência) ganhem mais que homens localizados na média masculina. Por fim os anos de experiência parecem não afetar o salário para as mulheres, pois seu parâmetro foi não significativo e para os os parece afetar negativamente.

Diagnóstico

Observando os gráficos de resíduos na figura 4 não percebemos nenhuma evidência contra a independência dos dados ou contra a homoscedasticidade. Porém através do histograma dos resíduos e o gráfico quantil-quatil nota-se um pequena assimetria e também caudas mais pesadas, o que sugere que a suposição que os erros sigam uma normal não esteja correta.

Em relação a pontos influentes e alavancas, podemos notar na figura 5 que vários pontos são candidatos. Sendo eles as observações: 4,191,163, 30 e 148. Porém ao retirá-los individualmente e ajustarmos o modelo novamente as conclusões não mudam e as estimativas praticamente não se alteram.

Modelo reduzido

Como vimos que para o sexo feminino o efeito do tempo de experiência não era significativa vamos propor um novo modelo, reduzido, sem esse parâmetro. Seja Y_{ij} o salário do j-ésimo executivo, para i = 0 (mulher), j = (1,2,. . . ,75) e para i = 1 (homem), j = (1,2,. . . ,145). Temos então o seguinte modelo:

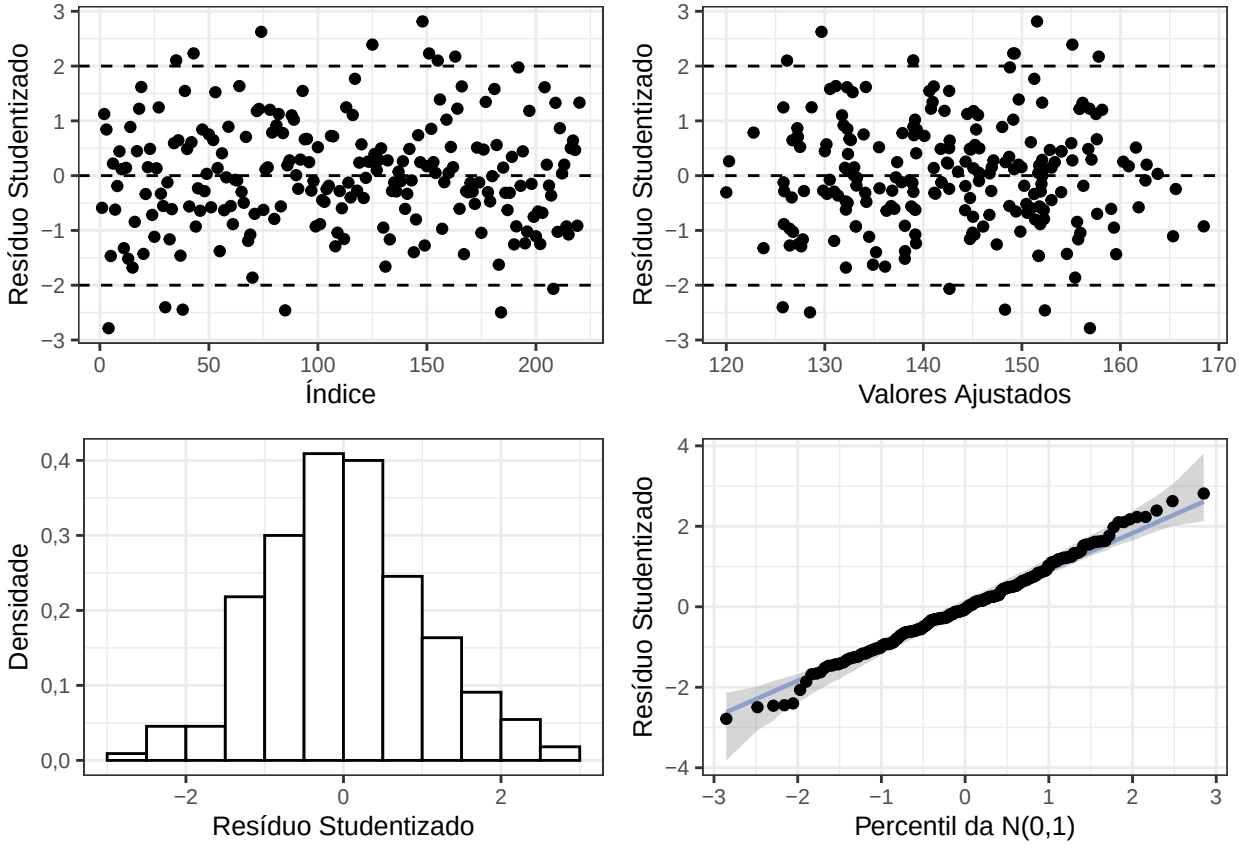

Figura 4: Diagnóstico do modelo inicial

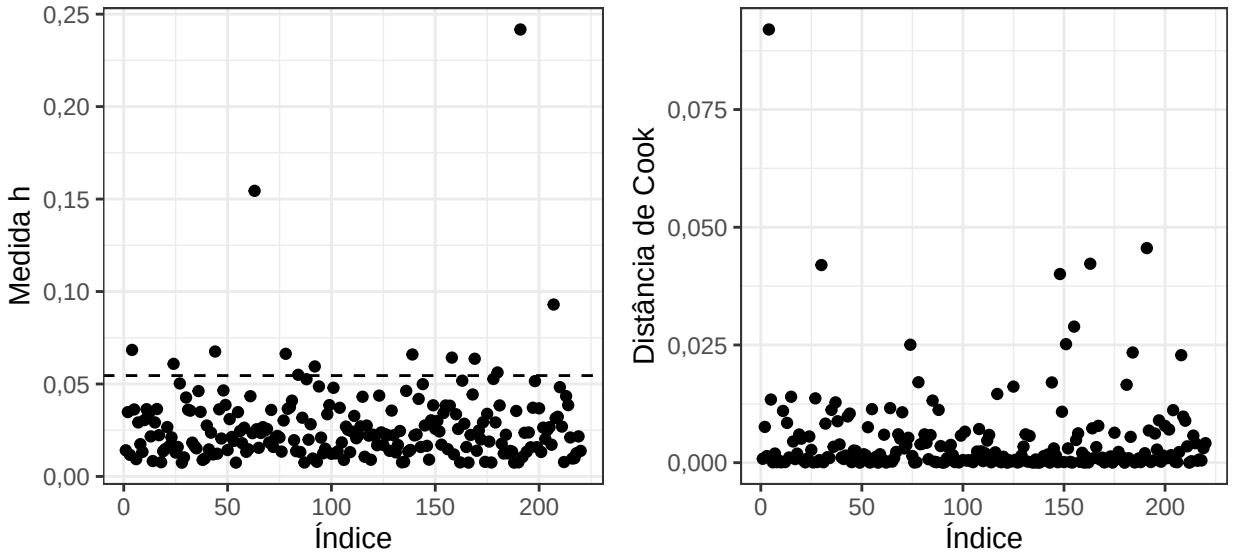

Figura 5: Análise de pontos influentes e alavancas para o modelo inicial

Tabela 2: Estimativa para os parâmetros do modelo completo.

Termo	Estimativa	Erro Padrão	IC (95%)	Estatística t	p-valor
α_0	145,079	1,049	[143,022 ; 147,136]	138,262	< 0.001
α_1	142,207	0,593	[141,044 ; 143,37]	239,713	< 0.001
β_{00}	6,411	0,525	[5,382 ; 7,441]	12,209	< 0.001
β_{10}	6,839	0,390	[6,074 ; 7,603]	17,533	< 0.001
β_{01}	-0,163	0,245	[-0,643 ; 0,317]	-0,666	0,506
β_{11}	-0,558	0,129	[-0,811 ; -0,305]	-4,326	< 0.001

Tabela 3: Estimativa para os parâmetros do modelo reduzido

Termo	Estimativa	Erro Padrão	IC (95%)	Estatística t	p-valor
α_0	145,474	0,864	[143,78 ; 147,168]	168,320	< 0.001
α_1	142,207	0,592	[141,046 ; 143,368]	240,024	< 0.001
β_{00}	6,248	0,463	[5,339 ; 7,156]	13,481	< 0.001
β_{10}	6,839	0,390	[6,075 ; 7,602]	17,556	< 0.001
β_{11}	-0,558	0,129	[-0,81 ; -0,305]	-4,332	< 0.001

$$Y_{ij} = \alpha_i + \beta_{i0} * (x_{0ij} - \bar{x}_{0i}) + \beta_{11} * (x_{1ij} - \bar{x}_{1i}) + \varepsilon_{ij}$$

onde x_{0ij} é a posição do j-ésimo executivo dado o sexo, x_{1ij} é a experiência do j-ésimo executivo dado o sexo. E α_i é o efeito do sexo no salário, β_{i0} é o efeito da posição, dado o sexo, no salário e β_{11} é o efeito da experiência para o sexo masculino no salário. Com a parte aleatória sendo ε_i.

Esse modelo é válido sobre as seguintes suposições: (i) $\varepsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$; (ii) as observações são independentes; (iii) e a variância é constante.

O novo modelo ajustado resulta no parâmetro da tabela 3, as estimativas e as colusões se mantém similares ao do modelo inicial e agora todos os parâmetros são significativos.

Observando os gráficos de resíduos na figura 6 chegamos as mesmas conclusões que para o modelo inicial. Que as suposições de homoscedasticidade e indecência são aceitáveis e a de normalidade não, que no caso do modelo reduzido parece até pior.

Também similar ao modelo inicial temos alguns candidatos a pontos influentes ou alavanca, observados na figura 7, mas que também ao serem retirados não alteram muito as estimativas ou a conclusão.

Comparações

Através das medidas de comparação, vistas na tabela 4, percebemos que para a maioria das medidas o modelo reduzido possuí um valor mais baixo, então em teoria seria um modelo melhor.

Conclusão

Mesmo com uma suposição quebrada, a que os dados seguem uma normal, os modelos se ajustaram relativamente bem. Essa quebra da suposição já era esperada pois salário é uma variável positiva e foi confirmada na análise de resíduos. Entretanto, devido as restrições, optamos pelo o modelo que apresentou o melhor ajuste, que foi o modelo reduzido. A partir dele concluímos que, homens na média masculina de experiência e de posição recebem menos que mulheres na média feminina de experiência e posição.

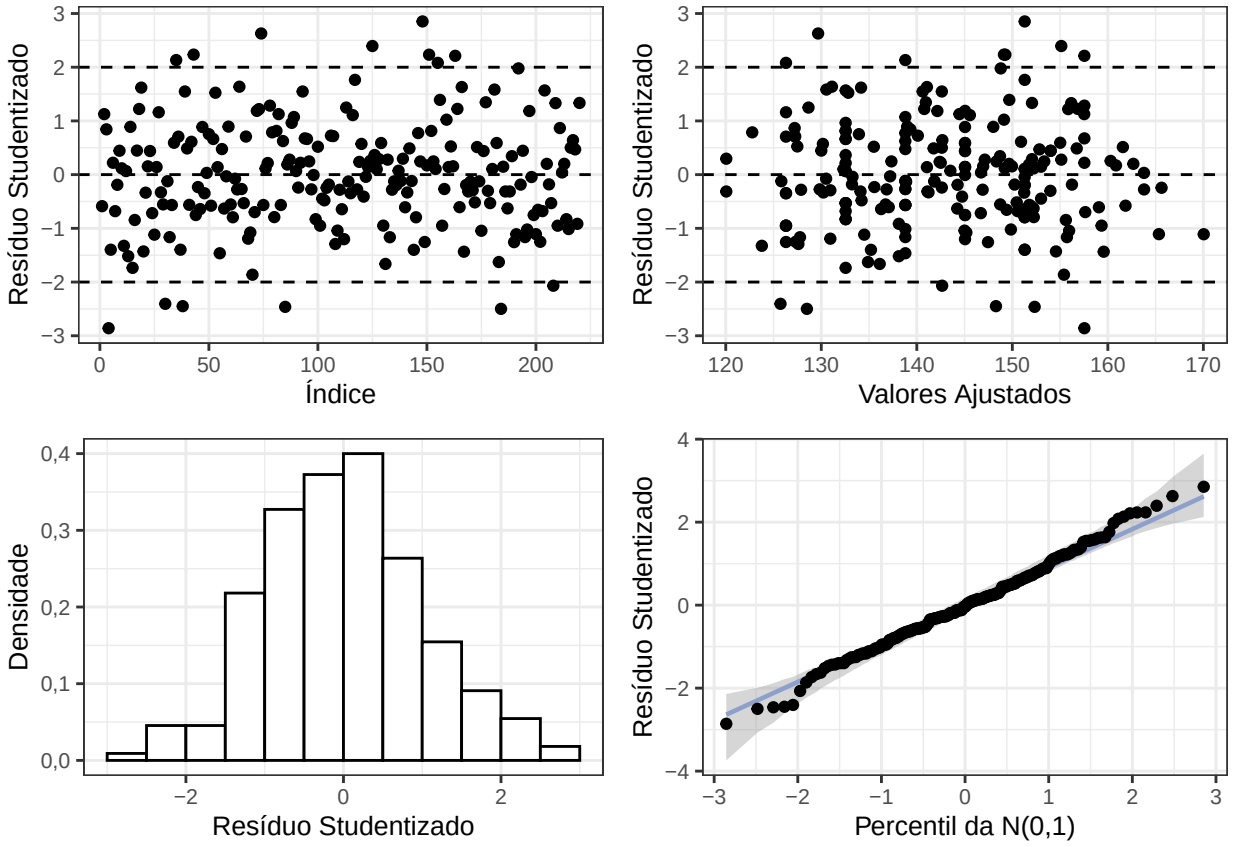

Figura 6: Diagnóstico do modelo reduzido

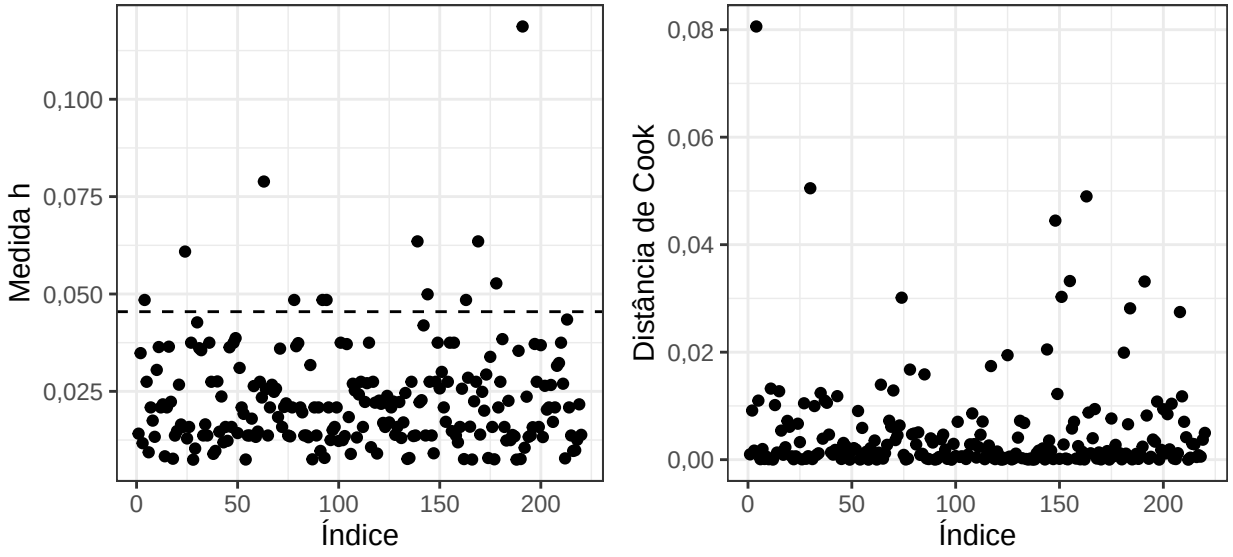

Figura 7: Análise de pontos influentes e alavancas para o modelo reduzido

Tabela 4: Medidas de comparação do modelo.

	Completo	Reduzido
AIC	1473,559	1472,015
BIC	1497,315	1492,377
AICc	1473,954	1472,295
SABIC	1472,907	1471,138
HQCIC	1479,782	1476,867
-2log.lik	1459,559	1460,015

Que tanto para mulheres ou para homens um aumento na posição implica em um aumento no valor esperado do salário e para os homens o tempo de experiência influência na média o salário negativamente.

Referências

Azevedo, C. L. N (2019). Notas de aula sobre Análise de regressão, http://www.ime.unicamp.br/~cnaber/Material_ME613_1S_2019.htm

Paula, G. A. (2013). Modelos de regressão com apoio computacional, versão pré-eliminar, https://www.ime.usp.br/~giapaula/texto_2013.pdf