SME0823 - Modelos Lineares Generalizados - Lista 2

Danilo Augusto Ganancin Faria – Nº USP: 9609172

20 de dezembro de 2019

Exercício 10

Os dados considerados neste exercício correspondem a uma amostra aleatória do salário anual (entre 110 e 172 mil USD) de 220 executivos (145 homens e 75 mulheres). Identificou-se ainda as seguintes variáves explicativas: sexo (1: masculino; 0: feminino), posição na empresa (varia de 1 a 9), quanto maior o valor mais alta a posição e anos de experiência no cargo ou tempo no cargo (varia de 1,7 a 26,1 anos).

Para iniciar o tratamento dos dados utilizou-se o código abaixo.

```
# Código para ler os dados e renomear as colunas
salary <- read.table("C:\\Users\\DaNiLo\\Documents\\Git\\mlg\\Exercício 10 L2\\salary.dat")
class(salary)
colnames(salary)
names(salary)[names(salary) == "V1"] <- "salario"
names(salary)[names(salary) == "V2"] <- "genero"
names(salary)[names(salary) == "V3"] <- "posicao"
names(salary)[names(salary) == "V4"] <- "experiencia"

# Separando dados por sexo
salary_masc <- salary[salary$genero == "Masculino", ]
salary_fem <- salary[salary$genero == "Feminino", ]
```

Análise de dados preliminar

```
summary(salary)
# Análise descritiva separada por sexo Masculino
summary(salary_masc)
head(salary_masc)
sd(salary_masc$salario)
var(salary_masc$salario)
sd(salary_masc$posicao)
var(salary_masc$posicao)
sd(salary_masc$experiencia)
var(salary_masc$experiencia)

# Feminino
summary(salary_fem)
```

```r
head(salary_fem)
sd(salary_fem$salario)
var(salary_fem$salario)
sd(salary_fem$posicao)
var(salary_fem$posicao)
sd(salary_fem$experiencia)
var(salary_fem$experiencia)
```

Com o intuito de exemplificar a estrutura dos dados, utilizou-se os comandos como segue.

```r
head(salary)
```

```
##   salario    genero posicao experiencia
## 1     148 Masculino       7        16.7
## 2     165 Masculino       7         6.7
## 3     145 Masculino       5        14.8
## 4     139  Feminino       7        13.9
## 5     142  Feminino       6         6.4
## 6     144 Masculino       5         9.1
```

```r
tail(salary)
```

```
##     salario    genero posicao experiencia
## 215     132  Feminino       4         4.6
## 216     146 Masculino       5         8.9
## 217     147 Masculino       5         8.8
## 218     156 Masculino       7        15.1
## 219     132 Masculino       4         4.7
## 220     161 Masculino       7        16.5
```

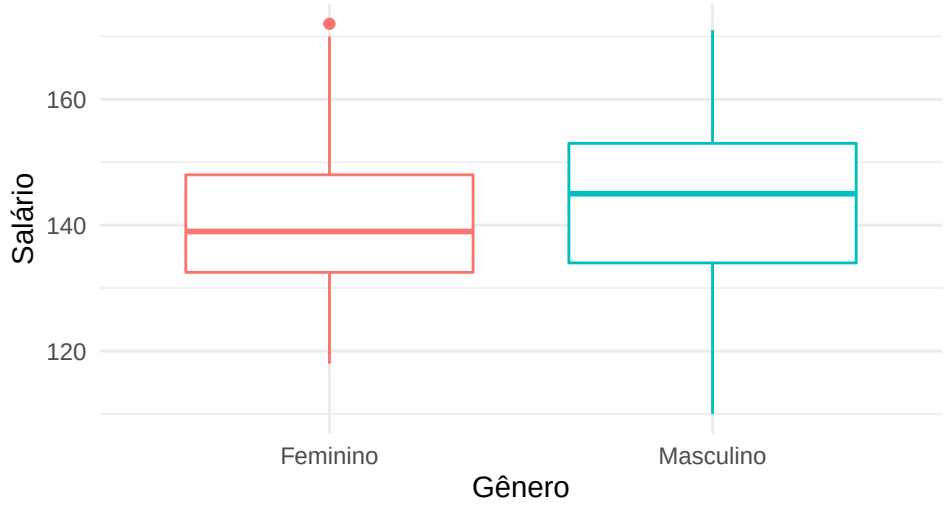

Figura 1: Gráficos de caixa do salário médio dos executivos de ambos os sexos.

De acordo com a Figura 1, há indícios de que os homens recebam maiores salários em comparação com as mulheres. Nota-se também a presença de um *outlier* em relação ao sexo feminino.

Tabela 1: Medidas resumo da variável Gênero

Gênero	Amostra	Min.	Med.	Máx.	Média	DP	Var.
Feminino	75	118	139	172	140,5	12,5	156,14
Masculino	145	110	145	171	144,1	12,39	153,61

Na Tabela 1 é possível verificar que a média salarial para executivos do sexo masculino é maior em comparação ao sexo feminino, apesar deste fato, o maior salário corresponde à uma executiva do sexo feminino, nota-se também que em ambos os casos temos uma variância salarial consideravelmente alta.

```
# Teste-t para comparação das médias
t.test(salary_masc$salario, salary_fem$salario, alternative = "two.sided", var.equal = FALSE)
```

Com a finalidade de comprovar que as médias salariais entre os sexos são estatisticamente diferentes, foi realizado o Teste-t.

Tabela 2: Teste-t para comparação das médias.

Teste	Estatística	valor-p
t	2,06	0,04

De acordo com a Tabela 2, o valor-p indica que há diferença entre as médias ao nível de 5% de significância. Dessa forma, podemos afirmar que os executivos ganham em média mais do que as executivas.

Tabela 3: Estatísticas descritiva das variáveis Salário, Posição e Experiência

Variável	Min.	Med.	Máx.	Média	DP	Var.
Salário	110	143,5	172	142,9	12,52	156,76
Posição	1	5	9	5,068	1,78	3,19
Experiência	1,7	9,5	26,1	10,48	5,21	27,19

Como pode-se observar na Tabela 3, a variável Salário possui valores entre 110 e 172, com média de 142,9 bem próxima de sua mediana 143,5 e ainda com uma variância razoavelmente grande de 156,76.

Da mesma forma, a variável Experiência está entre 1,7 e 26,1 possui média de 10,48 e mediana de 9,5 sugerindo assim como a variável Salário uma certa simetria pelo fato da média estar razoavelmente próxima da mediana.

Já na Figura 2, os gráficos de caixa nos mostram que em média os homens ocupam cargos mais altos nas empresas e também que eles passam mais tempo ocupando esses cargos quando comparados às mulheres.

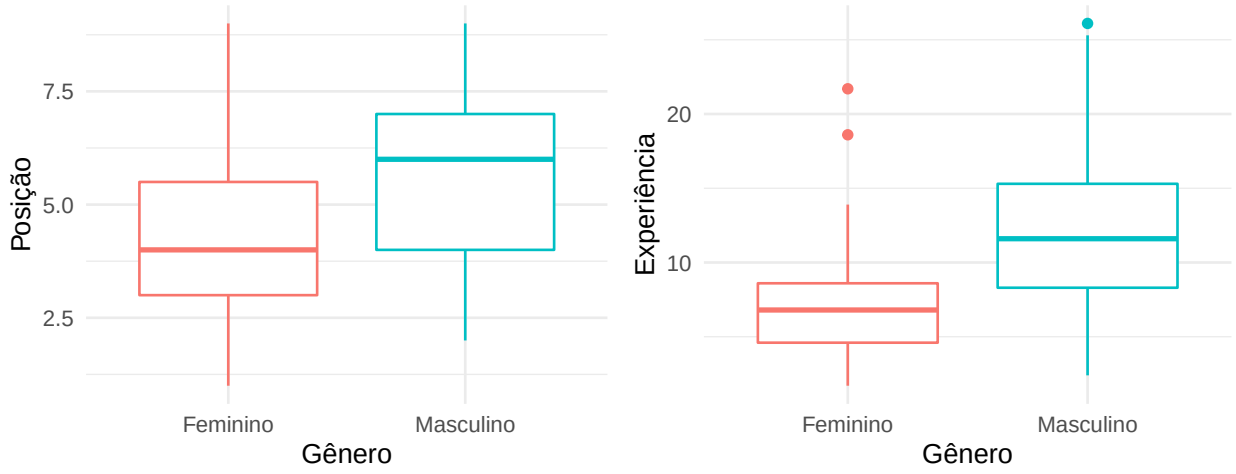

Figura 2: Gráficos de caixa das variáveis Posição e Experiência de ambos os sexos

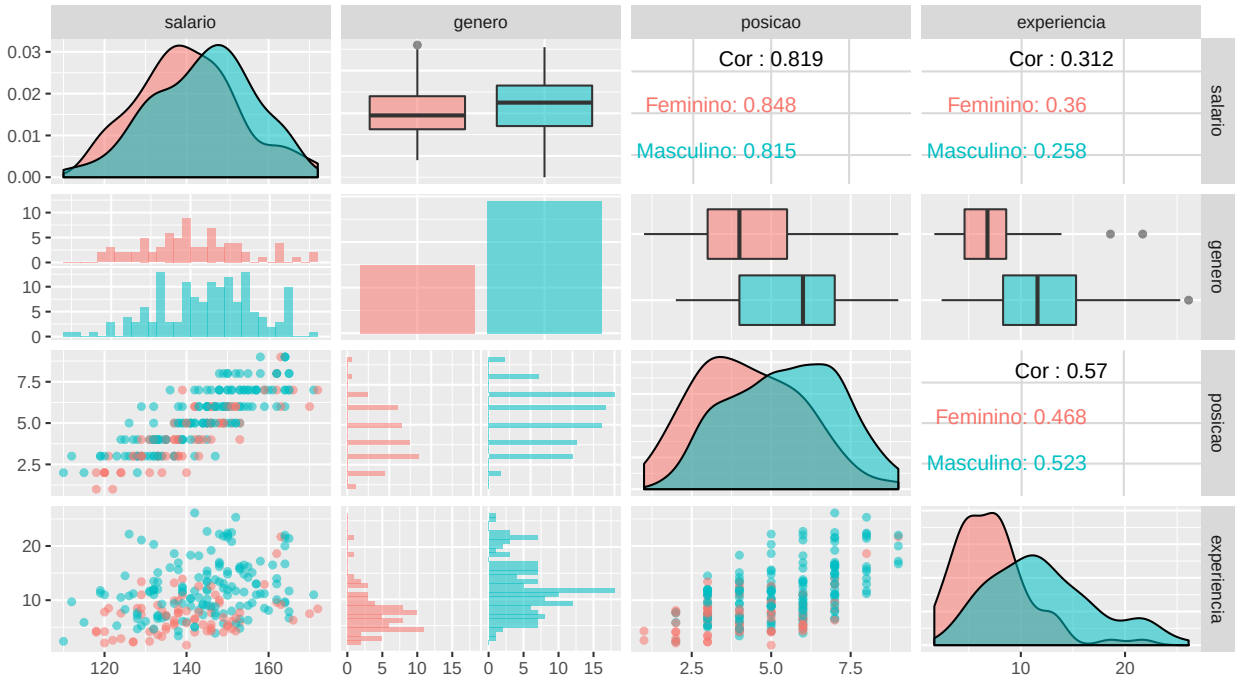

Figura 3: Gráfico de dispersão múltipla das variáveis.

No gráfico de dispersão múltipla da Figura 3, as variáveis Salário e Posição estão fortemente correlacionadas de maneira positiva ($\rho = 0,819$) e as variáveis Posição e Experiência estão moderadamente correlacionadas ($\rho = 0,57$) ou seja, há indícios de que grandes posições e experiência remetem em grandes salários para ambos os sexos, sugerindo inicialmente um modelo linear.

Ajuste do modelo linear

O modelo inicialmente proposto é:

$$y_i = \beta_0 + \beta_1 genero_i + \beta_2 posicao_i + \beta_3 experiencia_i + \varepsilon_i, \tag{1}$$

para $i = 1, \ldots, 220$, em que y_i corresponde ao salário do i-ésimo executivo da amostra e $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

```
# Ajuste do modelo linear Modelo linear completo
modelo <- lm(salario ~ ., data = salary)
summary(modelo)
xtable(stepAIC(modelo))
```

Tabela 4: Estimativas dos parâmetros para o modelo proposto.

Efeito	Estimativa	valor-t	valor-p
Intercepto	115,26	82,25	0,00
GêneroMasculino	-2,20	-2,04	0,04
Posição	6,71	21,46	0,00
Experiência	-0,47	-4,17	0,00
R^2	0,71		
Erro padrão	6,77		
AIC	845,31		

De acordo com as estimativas da Tabela 4, todas as variáveis são significativas ao nível de significância de 5%.

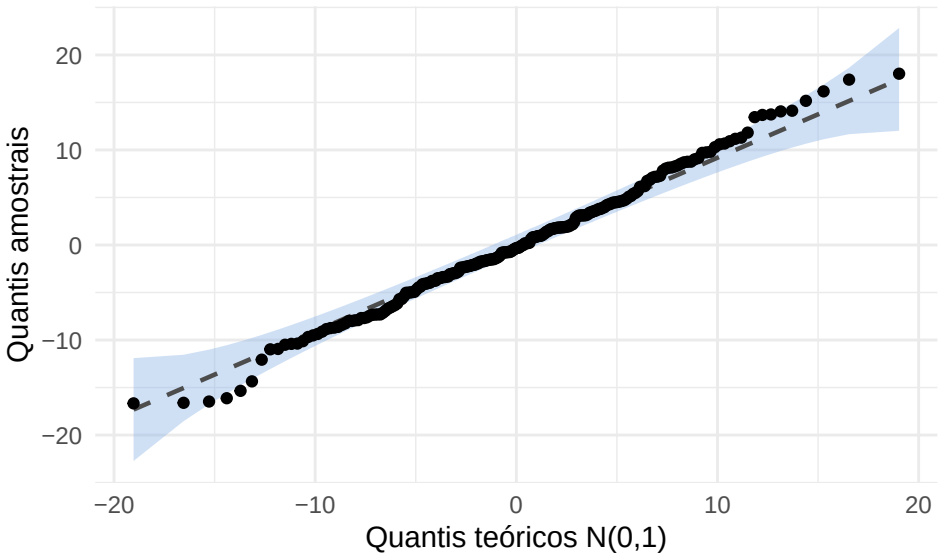

Figura 4: Q-Q plot dos resíduos do modelo ajustado.

```r
# Teste de normalidade de Shapiro-Wilk
shapiro.test(modelo$residuals)
```

Tabela 5: Teste de normalidade de Shapiro-Wilk para os resíduos do modelo ajustado.

Teste	Estatística	valor-p
Shapiro-Wilk	0,994	0,503

O *Q-Q* plot da Figura 4 mostra que grande maioria dos pontos estão próximos à reta apesar de algumas observações nas caudas estarem afastadas. O resultado obtido pelo teste de *Shapiro-Wilk* como mostrado na Tabela 5 afirmou que os resíduos seguem uma distribuição Normal ao nível de significância de 5%.

```r
# Teste de independência de Durbin-Watson
lmtest::dwtest(modelo)
```

Tabela 6: Teste de independência de Durbin-Watson para os resíduos do modelo ajustado.

Teste	Estatística	valor-p
Durbin-Watson	1,879	0,187

Como mostrado na Tabela 6, o teste de *Durbin-Watson* ao nível de 5% de significância atesta que os resíduos são independentes.

```r
# Teste de homocedasticidade de Breusch-Pagan
lmtest::bptest(modelo)
```

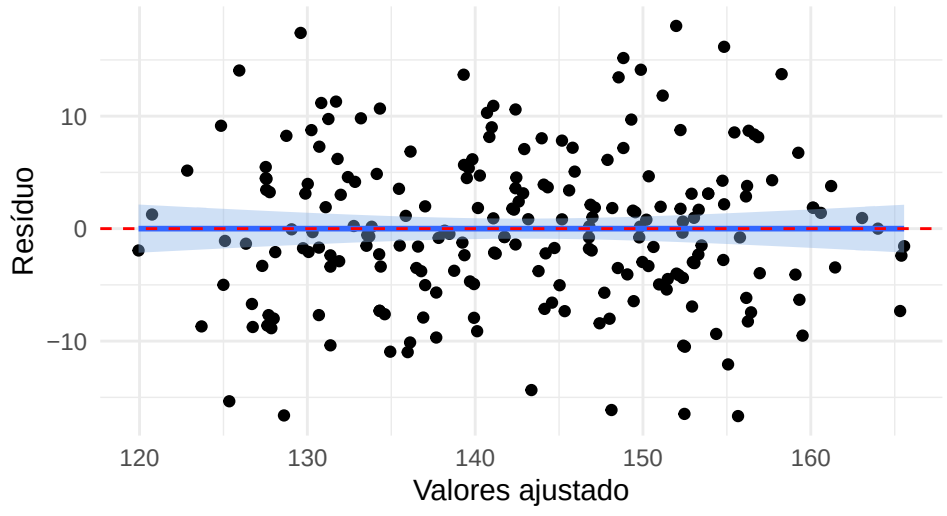

Figura 5: Gráfico dos Resíduos versus Valores ajustados.

Tabela 7: Teste de homocedasticidade de Breush-Pagan para a variância dos resíduos do modelo ajustado.

Teste	Estatística	valor-p
Breusch-Pagan	0,106	0,991

A Figura 5 não apresenta nenhuma tendência e está bem espalhada, temos indícios de independência e de que as variâncias podem ser constantes. O que é comprovado com o teste de *Breusch-Pagan* conforme resultado pode ser conferido na Tabela 7, ao nível de 5% de significância.

Após ter feito a comprovação de todas as suposições acerca do modelo de regressão linear múltipla pode-se então prosseguir para a seleção de modelos.

Seleção do modelo

A variável Gênero possui dois níveis (masculino e feminino). Pelo fato desses dois níveis não possuir a mesma variação em relação às outras variáveis, utiliza-se a interação entre os fatores para avaliar tal situação.

Por exemplo, a presença de interação entre os fatores Gênero e Experiência significa que a diferença entre os salários médios de executivos e executivas não é a mesma à medida que varia o tempo de experiência.

```
# Interação entre os fatores Mondelo completo + genero*experiencia
modelo1 <- lm(salario ~ . + genero * experiencia, data = salary)
summary(aov(modelo1))
stepAIC(modelo1)

# Modelo completo + genero*posicao
modelo2 <- lm(salario ~ . + genero * posicao, data = salary)
summary(aov(modelo2))
stepAIC(modelo2)

# Modelo completo + experiencia*posicao
modelo3 <- lm(salario ~ . + experiencia * posicao, data = salary)
summary(aov(modelo3))
stepAIC(modelo3)
summary(modelo3)
```

Tabela 8: Estimativas das interações entre os fatores.

Interação	R^2	AIC	valor-F	valor-p
Gênero*Experiência	0,714	845,66	1,615	0,205
Gênero*Posição	0,712	847,31	0,001	0,974
Experiência*Posição	0,722	839,67	7,594	0,006

De acordo com a Tabela 8, será incluída no modelo somente a interação *experiencia*posicao*.

Dessa forma, o modelo final selecionado é dado por:

$$y_i = \beta_0 + \beta_1 genero_i + \beta_2 posicao_i + \beta_3 experiencia_i + \beta_4 experiencia * posicao + \varepsilon_i, \tag{2}$$

para $i = 1, \ldots, 220$, em que y_i corresponde ao salário do i-ésimo executivo da amostra e $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

Tabela 9: Estimativas dos parâmetros para o modelo final.

Efeito	Estimativa	valor-t	valor-p
Intercepto	108,042	36,48	0,000
GêneroMasculino	-2,811	-2,59	0,010
Posição	8,096	13,73	0,000
Experiência	0,336	1,07	0,285
Posicao*Experiência	-0,135	-2,76	0,006
R^2	0,722		
Erro padrão	6,667		

Na Tabela 9 encontram-se as estimativas dos parâmetros para o modelo selecionado. Note que, o modelo final apresenta melhor coeficiente de determinação pois consegue explicar mais do que 72% dos valores observados em relação à variável resposta.

Portanto, o modelo ajustado é dado por:

$$\hat{y}(\mathbf{x}) = 108,042 - 2,811genero + 8,096posicao + 0,336experiencia - 0,135experiencia * posicao, \quad (3)$$

em que $\mathbf{x} = (genero, posicao, experiencia)^T$.

Interpretação do modelo ajustado:

- Para cada aumento de uma unidade da variável Gênero, o salário médio dos executivos diminui em 2,811 unidades;
- Para cada aumento de uma unidade da variável Posição, o salário médio dos executivos aumenta em 8,096 unidades;
- Para cada aumento de uma unidade da variável Experiência, o salário médio dos executivos aumenta em 0,336 unidades;
- Para cada aumento de uma unidade da variável combinada Experiência*Posição, o salário médio dos executivos diminui em 0,135 unidades.

Diagnóstico do modelo ajustado

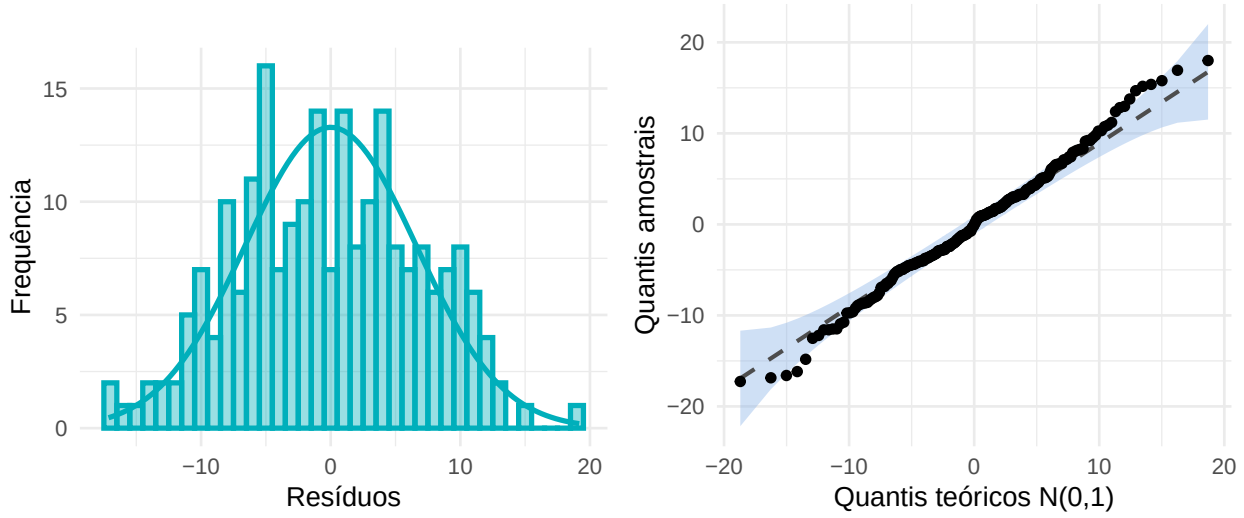

Figura 6: Histograma e Q-Q plot dos resíduos do modelo final ajustado

```
# Teste de normalidade de Shapiro-Wilk
shapiro.test(modelo3$residuals)

# Teste de independência de Durbin-Watson
lmtest::dwtest(modelo3)

# Teste de homocedasticidade de Breusch-Pagan
lmtest::bptest(modelo3)
```

Tabela 10: Testes de normalidade, independência e de homocedasticidade para os resíduos do modelo ajustado.

Teste	Estatística	valor-p
Shapiro-Wilk	0,994	0,482
Durbin-Watson	1,825	0,099
Breusch-Pagan	0,806	0,938

De acordo com a Figura 6 e o gráfico do resíduo *versus* valor ajustado da Figura 7, e ainda com as informações da Tabela 10, ao nível de 5% de significância, o modelo na fórmula (2) atende à todas as suposições (normalidade, indepêndencia e variância constante) de um modelo de regressão linear múltipla com erros normais.

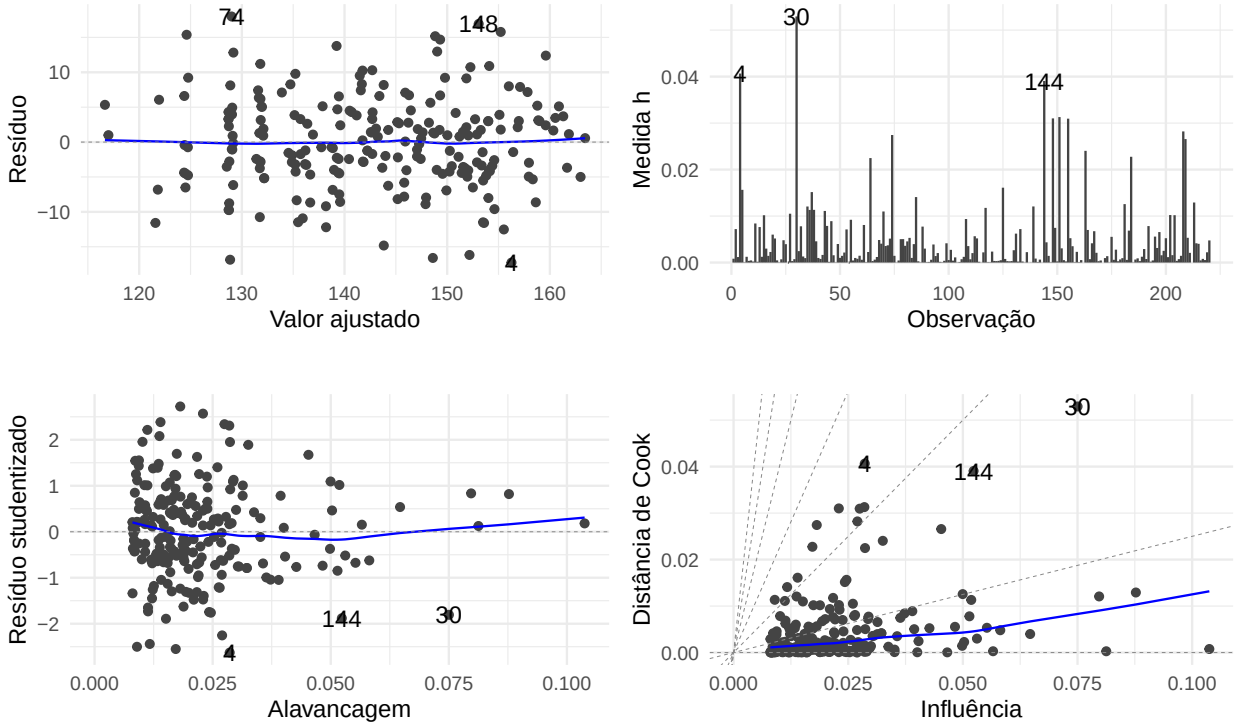

Figura 7: Diagnóstico do modelo ajustado.

Com o propósito de identificar as observações influentes, utilizou-se os gráficos da Figura 7. Foi possível identificar que apenas duas observações causam variações desproporcionais em algumas estimativas, porém não interferem em mudanças inferenciais, são as observações de números 4 e 30.

Identificou-se que a observação de número 30 corresponde a um executivo com salário anual de USD 110 mil, posição 2 com 2,4 anos de experiência. E também que a observação de número 4 é de uma executiva com salário anual de USD 139 mil, posição 7 e 13,9 anos de experiência.

Partindo do modelo ajustado dado pela fórmula em (3), tem-se o modelo ajustado para cada sexo.

O modelo ajustado para o grupo de executivas é dado por:

$$\hat{y}(\mathbf{x}) = 108,042 + 8,096 posicao + (0,336 - 0,135 posicao) * experiencia \tag{4}$$

em que $\mathbf{x} = (genero, posicao, experiencia)^T$.

O modelo ajustado para o grupo de executivos é dado por:

$$\hat{y}(\mathbf{x}) = 105,231 + 8,096 posicao + (0,336 - 0,135 posicao) * experiencia \tag{5}$$

em que $\mathbf{x} = (genero, posicao, experiencia)^T$.

Qual seria o salário previsto para executivos com 7 anos de experiência e posição 3?

- Executivo: USD 129,036 mil
- Executiva: USD 131,847 mil.

Conclusões

Quando desprezadas as variáveis Posição e Experiência os salários anuais dos executivos são em média significativamente maiores do que das executivas. No entanto, quando essas variáveis são consideradas pelo modelo ocorre o contrário, para uma mesma posição e mesma experiência as executivas ganham em média mais do que os executivos.

A interação entre Experiência e Posição mostra que a tendência entre o salário médio e o tempo no cargo não é a mesma para todas as posições. Em geral essa interação indica que não vale a pena do ponto de vista salarial ficar muito tempo no mesmo cargo.