Regressão Linear Simples

Trabalho de Estatística III

Kauã Dias

2025-05-17

Contents

1 Lendo os pacotes e a base de dados

```
if (!require("ggplot2")) install.packages(ggplot2)
```

```
## Carregando pacotes exigidos: ggplot2
```

```
if (!require("ggthemes")) install.packages(ggthemes)
```

```
## Carregando pacotes exigidos: ggthemes
```

```
if (!require("cowplot")) install.packages(cowplot)
```

```
## Carregando pacotes exigidos: cowplot
```

```
##
## Attaching package: 'cowplot'

## The following object is masked from 'package:ggthemes':
##
##     theme_map
```

```
library(ggplot2)
library(ggthemes)
library(cowplot)
```

```
dados <- read.csv2("dados/dados_regressao.csv", sep = ",")
```

2 Exploração Inicial dos Dados

```
head(dados)
```

```
##     sexo idade experiencia escolaridade          salario
## 1  Homem    49          33           12 76058.9358791817
## 2  Homem    49          33            9 75116.6201592173
## 3  Homem    44          11            8 52445.0741665169
## 4 Mulher    35          36           19 56891.6064448825
## 5  Homem    43          10           11 47646.5560984471
## 6 Mulher    30          24           11 49467.3043070965
```

A base possui 5 variáveis e 200 linhas, vamos categoriza-las

```
str(dados)
```

```
## 'data.frame':    200 obs. of  5 variables:
##  $ sexo        : chr  "Homem" "Homem" "Homem" "Mulher" ...
##  $ idade       : int  49 49 44 35 43 30 39 59 22 48 ...
##  $ experiencia : int  33 33 11 36 10 24 28 39 18 17 ...
##  $ escolaridade: int  12 9 8 19 11 11 18 17 17 18 ...
##  $ salario     : chr  "76058.9358791817" "75116.6201592173" "52445.0741665169" "56891.6064448825" ..
```

Perceba que temos apenas duas variáveis de formato *string*, porém a variável *salario* se encontra erroneamente definida, dado que intuitivamente, o salário se trata de uma variável numérica, vamos arrumar

```
dados$salario <- as.numeric(dados$salario)
str(dados)
```

```
## 'data.frame':    200 obs. of  5 variables:
##  $ sexo        : chr  "Homem" "Homem" "Homem" "Mulher" ...
##  $ idade       : int  49 49 44 35 43 30 39 59 22 48 ...
##  $ experiencia : int  33 33 11 36 10 24 28 39 18 17 ...
##  $ escolaridade: int  12 9 8 19 11 11 18 17 17 18 ...
##  $ salario     : num  76059 75117 52445 56892 47647 ...
```

Observe que o comando arredondou os valores da variável, vamos seguir e desconsiderar. Vamos descrever algumas estatísticas básicas dos dados

```
summary(dados)
```

```
##      sexo               idade        experiencia     escolaridade
##  Length:200         Min.   :20.00   Min.   : 1.00   Min.   : 8.00
##  Class :character   1st Qu.:31.00   1st Qu.: 9.00   1st Qu.:11.00
##  Mode  :character   Median :41.00   Median :20.00   Median :14.00
```

```
##                     Mean   :40.38   Mean   :20.21   Mean   :14.06
##                     3rd Qu.:49.00   3rd Qu.:30.25   3rd Qu.:17.00
##                     Max.   :60.00   Max.   :40.00   Max.   :20.00
##     salario
##  Min.   :33290
##  1st Qu.:47973
##  Median :56821
##  Mean   :56324
##  3rd Qu.:63312
##  Max.   :88894
```

Observe que o menor salário anual é de 33290 na moeda local, o que da 2774 por mês de acordo com

```
min(dados$salario)/12
```

```
## [1] 2774.193
```

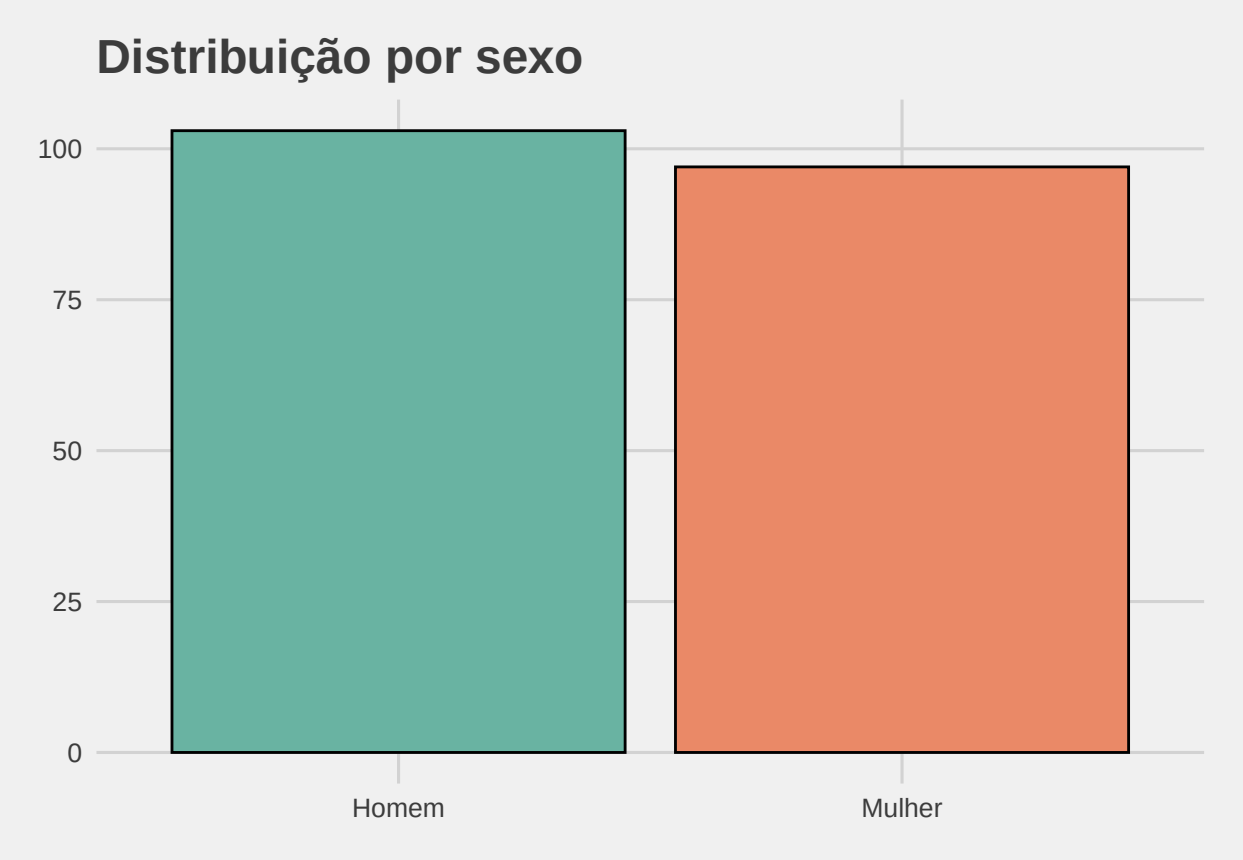
E paralelamente o maior salário anual é de 88894, que por mês é 7407.843, de acordo com

```
max(dados$salario)/12
```

```
## [1] 7407.843
```

2.1 Análise do gênero

```
dados |>
  ggplot(aes(x = sexo)) +
    geom_bar(fill = c("#69b3a2","#ea8967"), color = 'black') +
    labs(
      title = "Distribuição por sexo"
    ) +
    theme_fivethirtyeight()
```



Perceba que a base contém um pouco mais de homens do que mulheres.

```
table(dados$sexo)/nrow(dados)*100
```

```
##
##  Homem Mulher
##   51.5   48.5
```

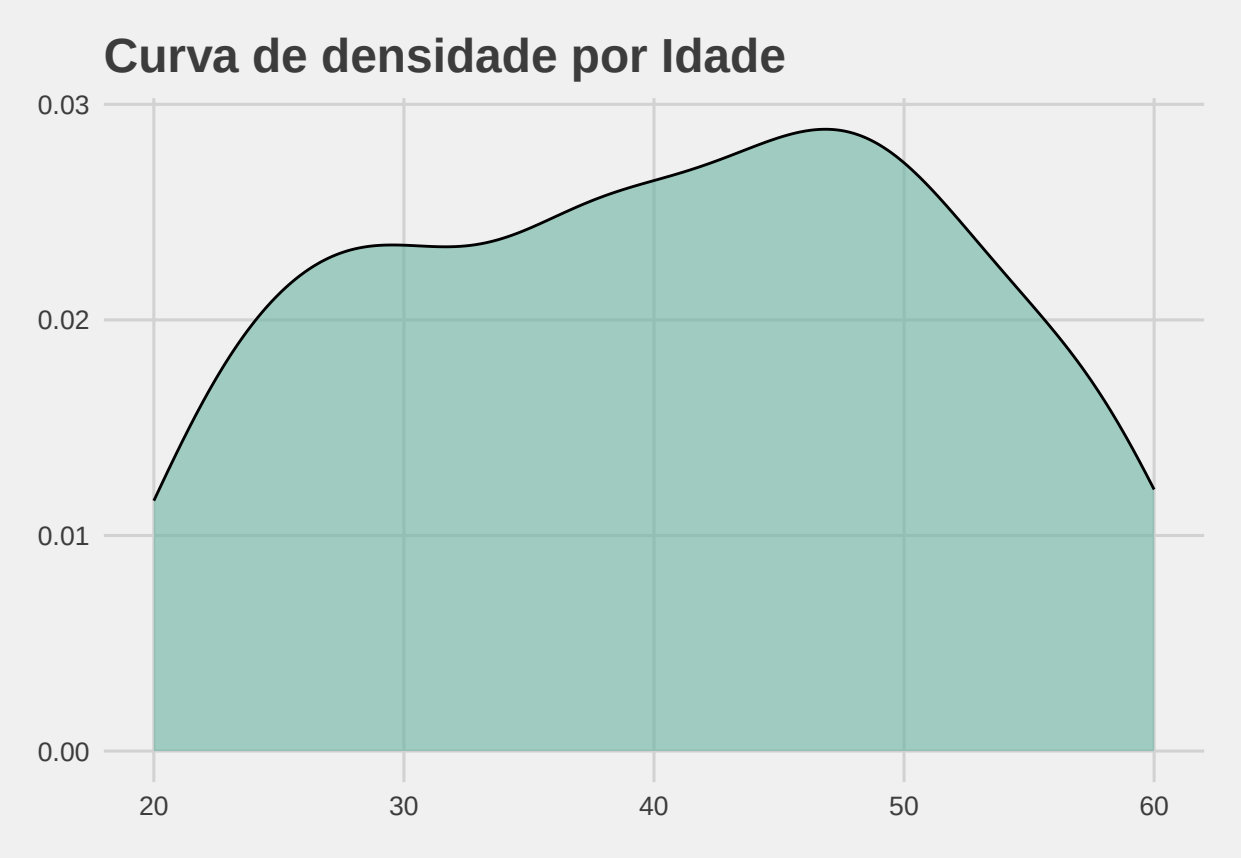
Apenas 3% a mais de homens.

Vamos construir uma função para gerar gráficos para as variáveis numéricas

```
grafico_dens <- function(dado, pos_var, titulo) {
  dado |>
    ggplot(aes(x = dados[[pos_var]])) +
    geom_density(
      aes(y = after_stat(density)),
          fill = "#69b3a2",
          alpha = 0.6
      ) +
    labs(
      title = titulo
    ) +
    theme_fivethirtyeight()
}
```

2.2 Análise da Idade

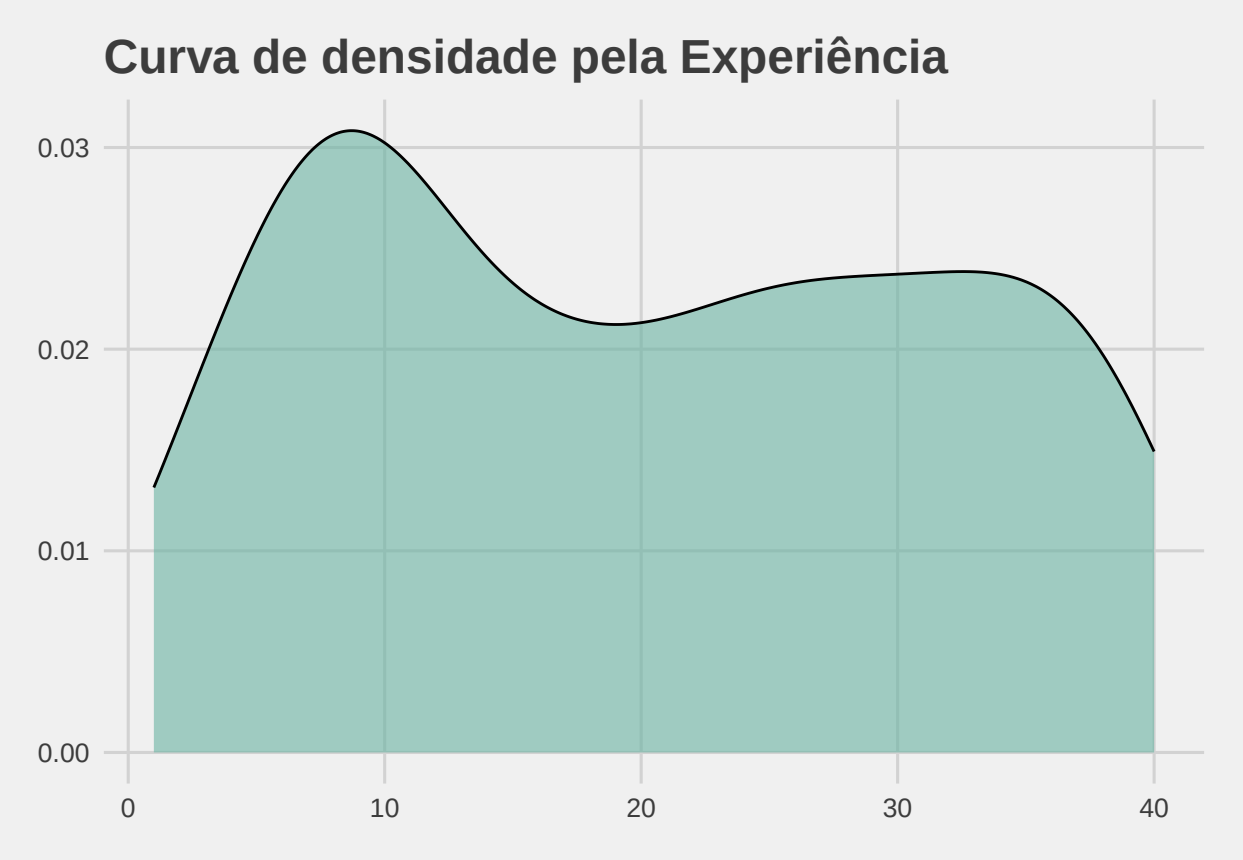
```
grafico_dens(dados, 2, "Curva de densidade por Idade")
```



Embora a média e mediana da idade se assemelhem, a moda parece se deslocar um pouco mais e isso influencia na pequena assimetria da curva

2.3 Análise dos anos de experiência

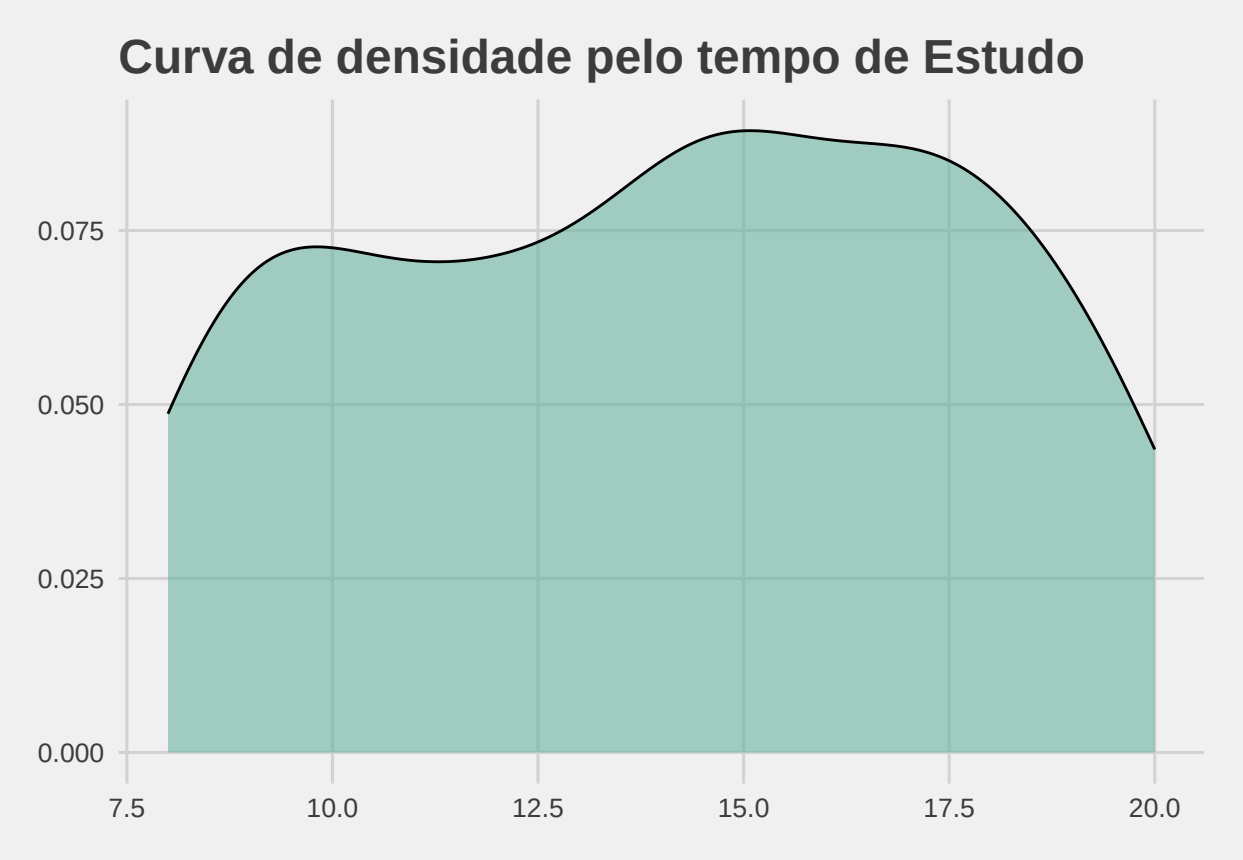
```
grafico_dens(dados, 3, "Curva de densidade pela Experiência")
```



Assim como a curva de densidade de Idade, a média e mediana se assemelham bastante, porém a moda diverge e isso causa uma certa assimetria na curva

2.4 Análise dos anos de escolaridade

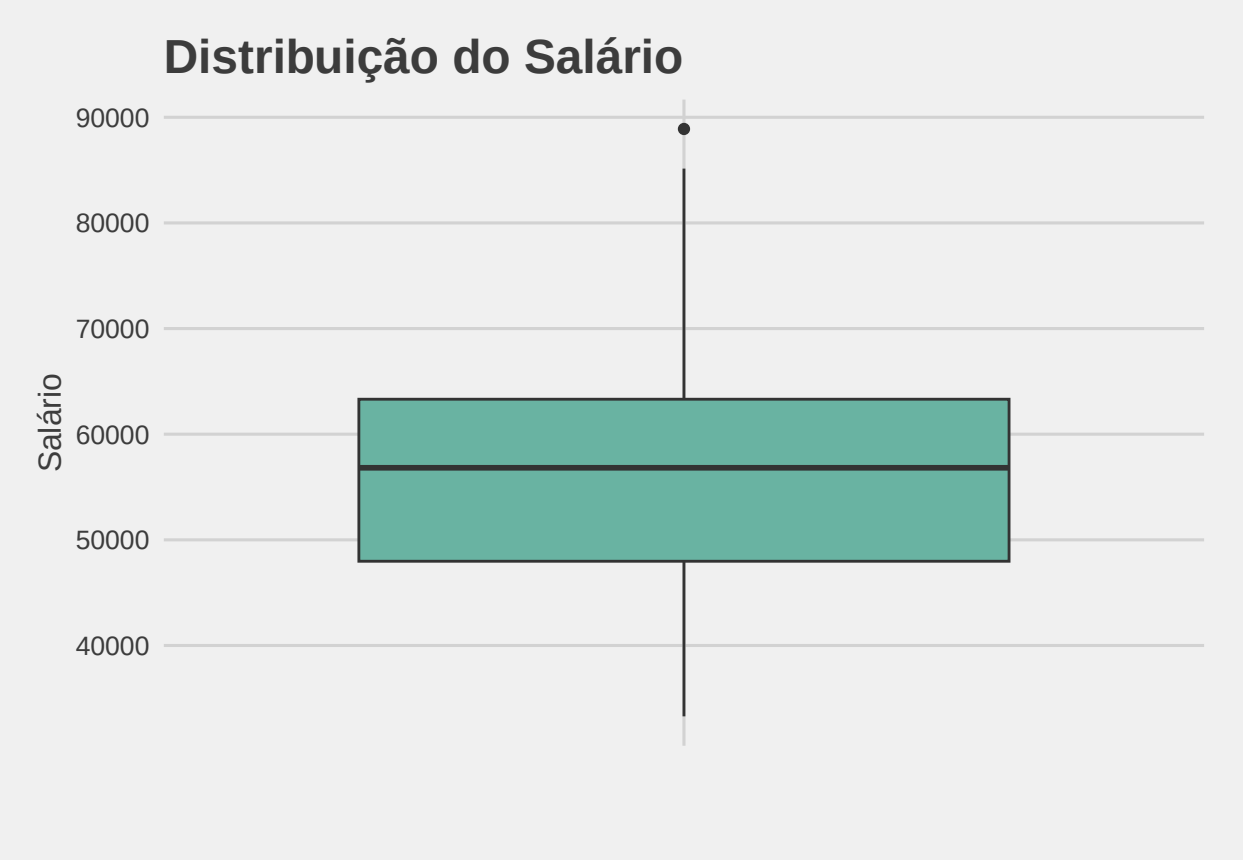
```
grafico_dens(dados, 4, "Curva de densidade pelo tempo de Estudo")
```



Também apresenta uma assimetria, mesmo que a média ≈ mediana ≈ moda.

Vamos construir um gráfico boxplot para entender a distribuição desses dados

```
dados |>
  ggplot(aes(x = "", y = salario)) +
  geom_boxplot(fill = "#69b3a2") +
  labs(
      title = "Distribuição do Salário",
    y = "Salário",
    x = ""
  ) +
  theme_fivethirtyeight() +
  theme(
    axis.title = element_text()
  )
```



Em geral os dados se apresentam bem distribuídos, é possível notar que possui um outlier para além do máximo do intervalo interquartil, vamos descobrir em qual linha se encontra

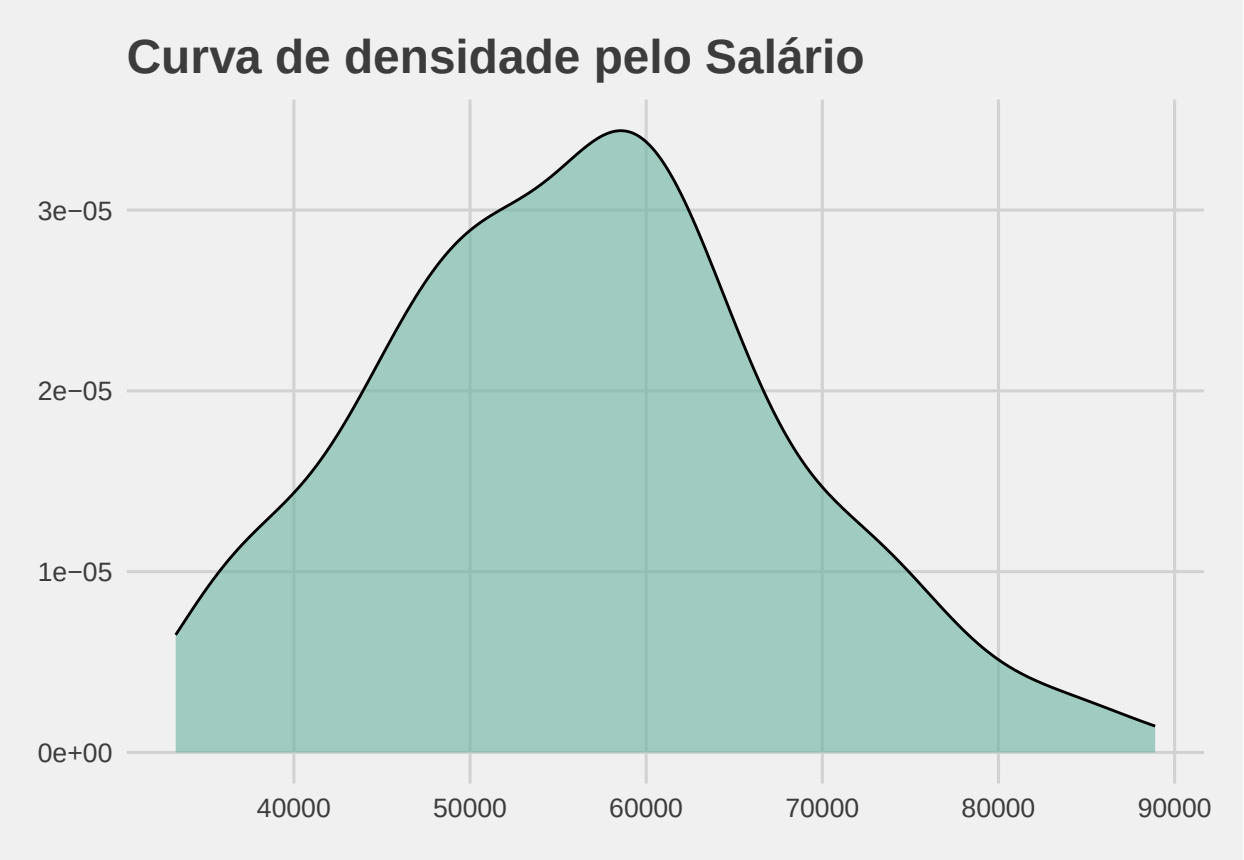
```
dados[dados$salario %in% boxplot.stats(dados$salario)$out, ]
```

```
##       sexo idade experiencia escolaridade  salario
## 138 Mulher    60          38           19 88894.11
```

Observe que a linha é condizente com a realidade dos dados, para além do gênero feminino (que veremos depois uma desvalorização do mercado de trabalho apontada pelo modelo), a escolaridade e a experiência condizem com o salário. Vamos manter o dado já que não se trata de um erro ou coisa do tipo.

2.5 Análise dos salários

```
grafico_dens(dados, 5, "Curva de densidade pelo Salário")
```



A curva de densidade do *Salário* apresenta uma forma mais normal que o resto das variáveis.

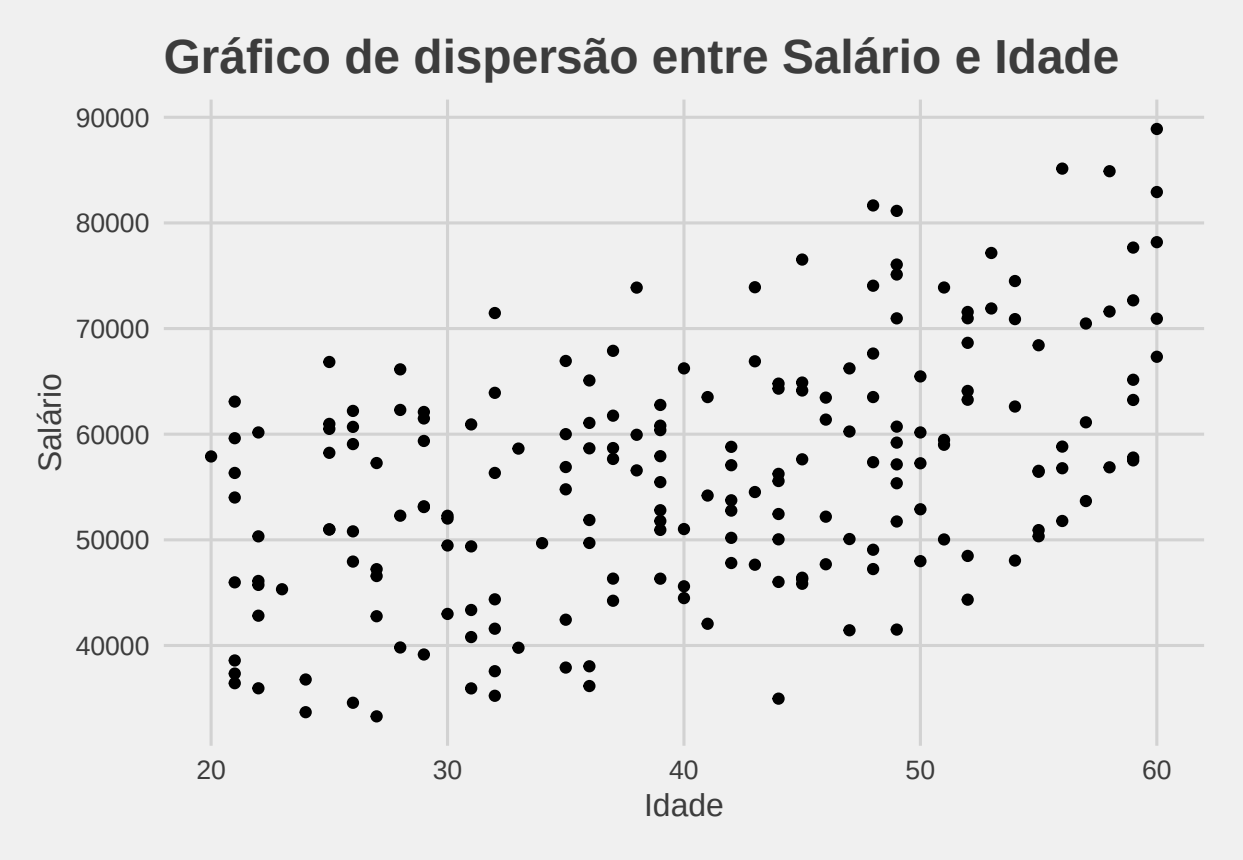
3 Relação linear entre as variáveis dos Dados

Vamos criar uma função para criar gráficos de dispersão do salario em função das demais variáveis numéricas

```
grafico_disp <- function(dado, pos_var, titulo, nome_var) {
  dado |>
    ggplot(aes(x = dado[[pos_var]], y = salario)) +
    geom_point() +
    labs(
      title = titulo,
      y = "Salário",
      x = nome_var
    ) +
    theme_fivethirtyeight() +
    theme(
      axis.title = element_text()
    )
}
```

3.1 Relação linear entre Salário e Idade

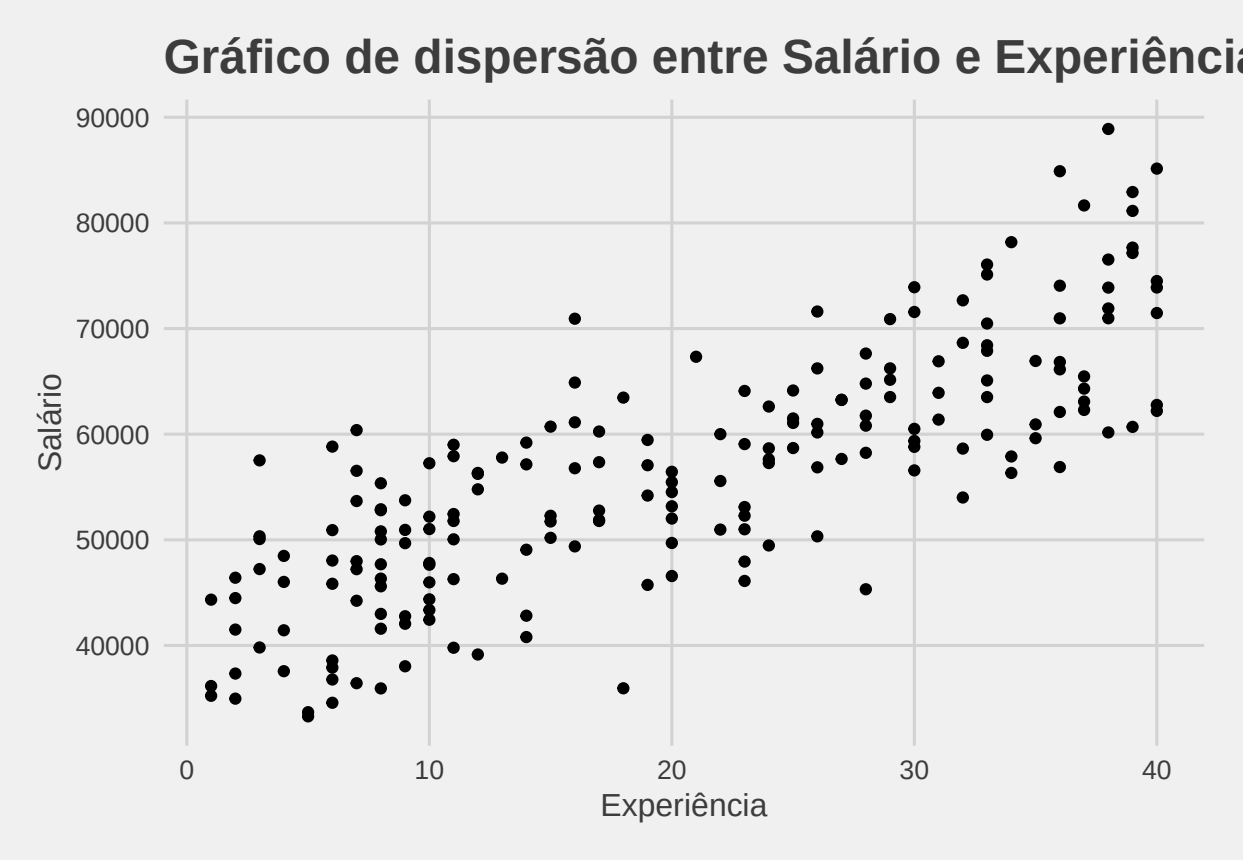
```
grafico_disp(dados, 2, "Gráfico de dispersão entre Salário e Idade", "Idade")
```



O gráfico não apresenta uma relação linear muito forte, porém é notório que o *Salário* tende a aumentar de acordo com a idade, mas com uma variabilidade considerável.

3.2 Relação linear entre Salário e Experiência

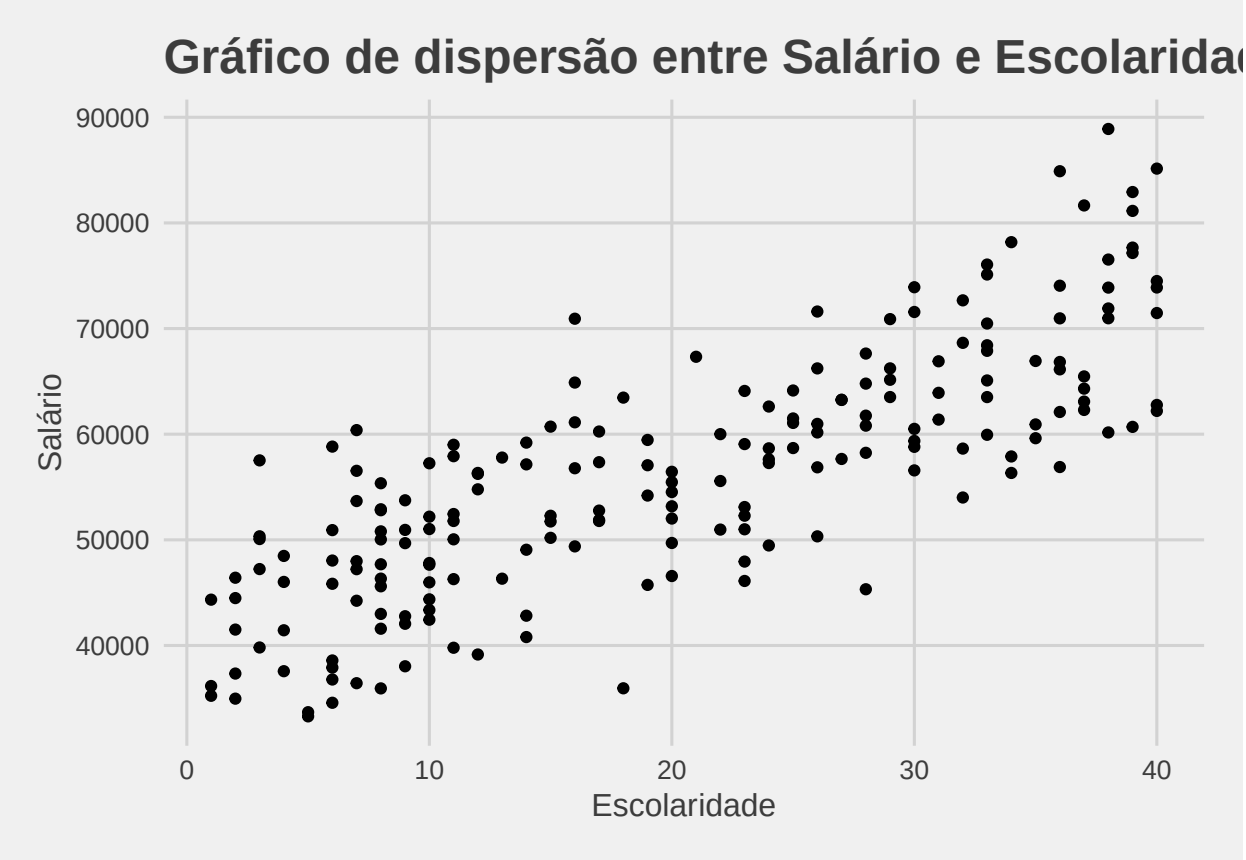
```
grafico_disp(dados, 3, "Gráfico de dispersão entre Salário e Experiência", "Experiência")
```



Assim como a variável *Idade*, o *Salário* tende a aumentar de acordo com a experiência, porém com uma menor variabilidade.

3.3 Relação linear entre Salário e Escolaridade

```
grafico_disp(dados, 3, "Gráfico de dispersão entre Salário e Escolaridade", "Escolaridade")
```



Com uma menor variabilidade, o gráfico sugere também uma tendência de aumentar o *Salário* proporcinalmente com a *Escolaridade*.

4 Ajuste do Modelo de Regressão

A variável escolhida para a construção do modelo será a *Experiência*, pois como vimos, possui uma forte relação linear (o *Salário* tende a aumentar de acordo com a *Experiência*), tem menor variabilidade que as outras variáveis e parece ser uma variável com um poder explicativo maior, pois a experiência tende a ter um impacto mais direto no salário em determinadas áreas.

```
# Construindo o  modelo
modelo <- lm(salario ~ experiencia, dados)

# Avaliando o modelo
summary(modelo)
```

```
##
## Call:
## lm(formula = salario ~ experiencia, data = dados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18662.6  -5061.5   -367.4   4381.4  18763.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  40640.6      973.2   41.76   <2e-16 ***
## experiencia    776.0       41.6   18.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6933 on 198 degrees of freedom
## Multiple R-squared:  0.6374, Adjusted R-squared:  0.6356
## F-statistic: 348.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

Os coeficientes fornecem a seguinte equação

$$\hat{Y} = \theta_1 + \theta_2 X,$$

onde θ_1, θ_2 são o intercepto (quando a *Experiência é zero*) e o coeficiente angular (a inclinação da reta na equação dada), respectivamentes. No contexto dado a equação tem forma

$$\hat{salário} = 40640.6 + 776 \times experiencia$$

.

5 Avaliação do Modelo

Vamos utilizar o modelo para prever os dados

```
dados$salario_predit <- predict(modelo, dados)
```

5.1 Calculando o *EQM, MAE, MAPE*

```
eqm <- mean((dados$salario - dados$salario_predit)^2)
mae <- mean(abs(dados$salario - dados$salario_predit))
mape <- mean(abs(dados$salario - dados$salario_predit)/dados$salario)*100
print(list(
  "EQM" = eqm,
  "MAE" = mae,
  "MAPE" = mape)
  )
```

```
## $EQM
## [1] 47584647
##
## $MAE
## [1] 5599.72
##
## $MAPE
## [1] 10.43424
```

Como o Erro Quadrático Médio (EQM) é uma métrica mais apropriada para comparação entre diferentes modelos e tem interpretação limitada por estar em uma escala quadrática, focamos nossa análise nas métricas MAE (Erro Absoluto Médio) e MAPE (Erro Percentual Absoluto Médio), que possuem interpretação mais direta.

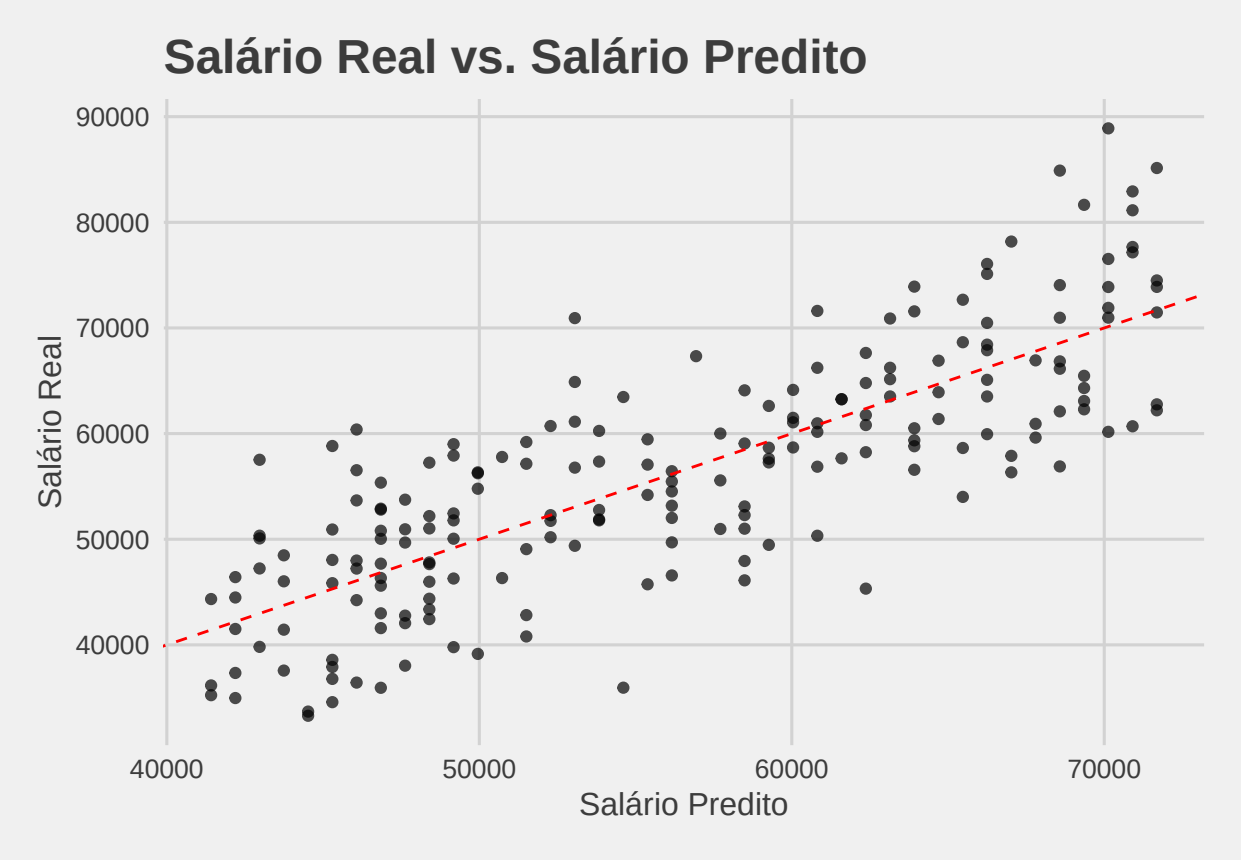
MAPE $\approx$ 10,43%: Isso significa que, em média, o modelo erra cerca de 10% no valor previsto dos salários. Para muitos contextos de negócios ou estudos sociais, esse é um nível de erro considerado baixo e aceitável, o que indica boa acurácia preditiva.

MAE $\approx$ 5599: O erro médio absoluto nas previsões é de aproximadamente 5.599,00. Dado que os salários na base variam entre 30.000 e 90.000, esse erro representa uma pequena fração do valor total, reforçando a ideia de que o modelo tem um bom desempenho.

Com base nessas métricas, podemos afirmar que o modelo de regressão linear simples utilizando a variável *experiência* como preditora apresenta bom poder preditivo e é adequado para estimar salários dentro do contexto dos dados fornecidos.

Por fim, vamos analisar os valores preditos em função dos valores reais

```
dados |>
  ggplot(aes(x = salario_predit, y = salario)) +
  geom_point(alpha = .7) +
  geom_abline(intercept = 0,
              slope = 1,
              color = "red",
              linetype = "dashed") +
  labs(
    title = "Salário Real vs. Salário Predito",
    x = "Salário Predito",
    y = "Salário Real"
  ) +
  theme_fivethirtyeight() +
   theme(
      axis.title = element_text()
   )
```



O gráfico sugere que os dados estão predominantemente próximos da linha de referência $y = x$, o que indica boa fidelidade do modelo, já que os valores preditos se aproximam dos valores reais. A dispersão é relativamente simétrica em torno da linha, reforçando a consistência das previsões. No entanto, observa-se que nos extremos da distribuição (salários muito altos ou muito baixos), há maior discrepância entre os valores preditos e reais, sugerindo que o modelo apresenta maior variabilidade ou imprecisão em casos

extremo.

6 Comparação por Gênero

Vamos dividir a base de dados por *sexo*:

```
dados_h <- dados[dados$sexo == "Homem",]
dados_m <- dados[dados$sexo == "Mulher",]
```

6.1 Criação do modelo por gênero

```
modelo_h <- lm(salario ~ experiencia, dados_h)
modelo_m <- lm(salario ~ experiencia, dados_m)

summary(modelo_h); summary(modelo_m)
```

```
##
## Call:
## lm(formula = salario ~ experiencia, data = dados_h)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18989.9  -4654.7     88.4   4399.2  13590.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 41728.86    1342.87   31.07   <2e-16 ***
## experiencia   733.77      58.39   12.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6779 on 101 degrees of freedom
## Multiple R-squared:  0.6099, Adjusted R-squared:  0.6061
## F-statistic: 157.9 on 1 and 101 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = salario ~ experiencia, data = dados_m)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12059.8  -5245.0   -477.4   4391.9  18317.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39544.51    1418.17   27.88   <2e-16 ***
## experiencia   816.86      59.56   13.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7120 on 95 degrees of freedom
## Multiple R-squared:  0.6644, Adjusted R-squared:  0.6609
## F-statistic: 188.1 on 1 and 95 DF,  p-value: < 2.2e-16
```

O modelo construído para *Homem* fornece a equação

$$\widehat{salário} = 41728.86 + 733.77 \times experiencia,$$

enquanto que para *Mulher*

$$\widehat{salário} = 39544.51 + 816.86 \times experiencia.$$

A interpretação desses resultados mostra que o intercepto θ_1 é maior para os *Homens*, o que indica que, para zero anos de experiência, o modelo prevê um salário inicial maior para os *Homens* em comparação às *Mulheres*, demonstrando assim, uma valorização maior ao trabalho masculino no início que ao feminino.

Por outro lado, o coeficiente angular θ_2 é maior no modelo feminino, o que sugere que o salário das *Mulheres* cresce mais rapidamente conforme a experiência aumenta. Isso pode ser interpretado como uma valorização proporcional maior da experiência feminina.

6.2 Validação do modelo por gênero

Vamos utilizar o modelo para prever os dados

```
dados_h$salario_predit <- predict(modelo_h, dados_h)
dados_m$salario_predit <- predict(modelo_m, dados_m)
```

6.3 Calculando o *EQM, MAE, MAPE*:

```
data.frame(
  "Gênero" = c("Masculino", "Feminino"),
  "EQM" = c(
    mean((dados_h$salario - dados_h$salario_predit)^2),
    mean((dados_m$salario - dados_m$salario_predit)^2)
  ),
  "MAE" = c(
    mean(abs(dados_h$salario - dados_h$salario_predit)),
    mean(abs(dados_m$salario - dados_m$salario_predit))
  ),
  "MAPE" = c(
    mean(abs(dados_h$salario - dados_h$salario_predit)/dados_h$salario)*100,
    mean(abs(dados_m$salario - dados_m$salario_predit)/dados_m$salario)*100
  )
)
```

```
##      Gênero      EQM      MAE     MAPE
## 1 Masculino 45056668 5451.538 10.10944
## 2  Feminino 49644071 5715.297 10.64559
```

O modelo ajustado para o grupo Masculino apresentou valores ligeiramente menores em todas as métricas de erro (*EQM, MAE e MAPE*), indicando que ele realiza previsões um pouco mais precisas do que o modelo construído para o grupo Feminino.

Apesar dessa diferença, os valores de erro entre os dois grupos são muito próximos, o que sugere que o desempenho dos modelos é semelhante em termos de capacidade preditiva. Dessa forma, não há evidências de que o modelo para um dos gêneros seja significativamente superior ao outro.

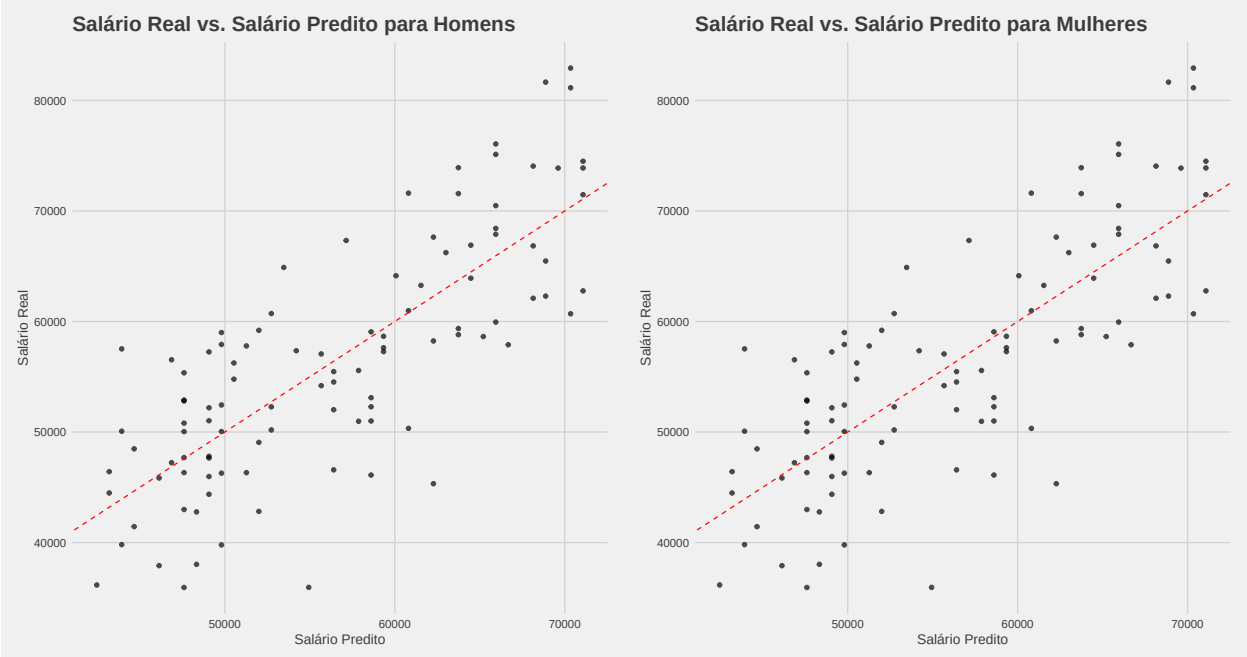
Por fim, vamos analisar os valores preditos em função dos valores reais para os modelos por gênero

```
g1 <- dados_h |>
  ggplot(aes(x = salario_predit, y = salario)) +
  geom_point(alpha = .7) +
```

```
  geom_abline(intercept = 0,
              slope = 1,
              color = "red",
              linetype = "dashed") +
  labs(
    title = "Salário Real vs. Salário Predito para Homens",
    x = "Salário Predito",
    y = "Salário Real"
  ) +
  theme_fivethirtyeight() +
   theme(
      axis.title = element_text()
   )

g2 <- dados_h |>
  ggplot(aes(x = salario_predit, y = salario)) +
  geom_point(alpha = .7) +
  geom_abline(intercept = 0,
              slope = 1,
              color = "red",
              linetype = "dashed") +
  labs(
    title = "Salário Real vs. Salário Predito para Mulheres",
    x = "Salário Predito",
    y = "Salário Real"
  ) +
  theme_fivethirtyeight() +
   theme(
      axis.title = element_text()
   )

plot_grid(g1, g2)
```



O modelo apresenta um bom desempenho preditivo para ambos os gêneros, com resultados visivelmente semelhantes. Entretanto, observa-se que o modelo para o grupo masculino apresenta desempenho ligeiramente superior, com menor dispersão dos pontos em torno da linha ideal $y = x$ e métricas de erro um pouco menores. Para o grupo feminino, há uma maior variabilidade nas predições, especialmente em faixas salariais mais altas, o que pode indicar uma menor precisão nessas regiões. Uma possível explicação para essa diferença está na distribuição das observações: a base de dados é levemente desbalanceada, contendo aproximadamente 51,5% de observações masculinas e 48,5% femininas. Esse leve desequilíbrio pode impactar a qualidade do ajuste do modelo, favorecendo o grupo mais representado.

7 Interpretação e Conclusões

Com base nos resultados obtidos, é possível concluir que o modelo de regressão linear simples apresenta um bom desempenho preditivo, tanto no conjunto geral de dados quanto nas divisões por gênero. As métricas de erro (*EQM, MAE e MAPE*) são bastante semelhantes entre os grupos, com o *MAPE* em torno de 10% para ambos, o que indica que o modelo comete, em média, um erro percentual relativamente baixo nas previsões salariais.

A análise dos coeficientes mostra que, segundo o modelo, o mercado tende a atribuir um salário inicial maior aos homens (intercepto mais alto), enquanto o ganho associado à experiência (coeficiente angular) é mais elevado para as mulheres. Isso pode sugerir uma valorização crescente da experiência feminina no mercado de trabalho, embora partindo de um ponto inicial inferior.

Por fim, vale destacar que o modelo pode ser aprimorado com a inclusão de mais observações e, principalmente, com a adição de outras variáveis explicativas que possam captar melhor a complexidade que influencia os salários.