Análise dos Determinantes Salariais

Alison Cordeiro Sousa

Estudo

Análise dos Determinantes Salariais
Este estudo examina os fatores que influenciam os salários dos trabalhadores, utilizando dados de 526 indivíduos. A análise focou em três variáveis-chave: anos de educação (educ), experiência profissional (exper) e tempo na empresa atual (tenure). O modelo de regressão múltipla revelou que cada ano adicional de educação aumenta o salário em aproximadamente US$ 0.60/hora ($p < 0.001$), enquanto o tempo na empresa mostra um impacto positivo de US$ 0.17/hora por ano ($p < 0.001$). A experiência profissional apresentou efeito marginalmente significativo (US$ 0.02/hora, $p = 0.065$), sugerindo que seus benefícios podem ser capturados em parte pelo tempo na empresa atual.

Diferenças Salariais por Gênero
A análise de colinearidade demonstrou um significativo gap salarial entre gêneros. Mulheres apresentaram salários em média US$ 2.27 mais baixos que homens ($p < 0.001$), mesmo controlando por educação. Curiosamente, ao substituir a variável 'female' por 'male', o coeficiente inverteu o sinal mantendo a mesma magnitude, confirmando a robustez do achado. Esses resultados persistiram após ajustes para diversas características individuais, indicando possíveis disparidades estruturais no mercado de trabalho.

Qualidade do Ajuste e Diagnósticos
O modelo explicou 30.6% da variação salarial (R^2 ajustado = 0.302), com resíduos mostrando distribuição aproximadamente normal, porém com alguns outliers extremos. Gráficos de diagnóstico revelaram heterocedasticidade moderada, sugerindo que erros-padrão convencionais podem subestimar a incerteza. A análise de valores influentes identificou 15 observações atípicas que merecem investigação adicional, porém sua exclusão não alterou significativamente as conclusões principais.

Aplicações e Limitações
Os resultados têm implicações importantes para políticas de equidade salarial e desenvolvimento de carreira. A forte associação entre educação e salários reforça o valor do investimento em capital humano. A análise de multicolinearidade entre variáveis educacionais (VIF > 10 para avg_ed) demonstrou a importância da seleção criteriosa de covariáveis. Estudos futuros deveriam incorporar medidas de habilidade inata (como no dataset Base3), cuja omissão pode superestimar em 23% o efeito da educação nos salários.

```
# =========================
# 1. Configurações iniciais
# =========================
```

```r
# Limpa o ambiente para evitar conflitos com objetos antigos
rm(list = ls())

# Define o diretório onde estão os arquivos CSV
setwd("C:/Users/PC GAMER/Downloads/data")

# Instala e carrega os pacotes necessários
pacotes <- c("tidyverse", "stargazer", "magrittr", "car",
             "ggplot2", "GGally", "gridExtra", "MASS")
for (p in pacotes) {
  if (!require(p, character.only = TRUE)) {
    install.packages(p, dependencies = TRUE)
    library(p, character.only = TRUE)
  }
}

# ================================
# 2. Carregamento das bases de dados
# ================================

# Cada base contém um recorte diferente do estudo
base1 <- read.csv("base1.csv")  # Dados de salários
base2 <- read.csv("base2.csv")  # Salários de CEOs
base3 <- read.csv("base3.csv")  # Dados com variável de habilidade
base4 <- read.csv("base4.csv")  # Dados educacionais

# ================================================
# 3. Regressão Múltipla: Salário ~ Educ + Exper + Tenure
# ================================================

# Visualiza a estrutura da base
glimpse(base1)
```

```
## Rows: 526
## Columns: 24
## $ wage     <dbl> 3.10, 3.24, 3.00, 6.00, 5.30, 8.75, 11.25, 5.00, 3.60, 18.18,~
## $ educ     <int> 11, 12, 11, 8, 12, 16, 18, 12, 12, 17, 16, 13, 12, 12, 12, 16~
## $ exper    <int> 2, 22, 2, 44, 7, 9, 15, 5, 26, 22, 8, 3, 15, 18, 31, 14, 10, ~
## $ tenure   <int> 0, 2, 0, 28, 2, 8, 7, 3, 4, 21, 2, 0, 0, 3, 15, 0, 0, 10, 0, ~
## $ nonwhite <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ female   <int> 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1~
## $ married  <int> 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0~
## $ numdep   <int> 2, 3, 2, 0, 1, 0, 0, 0, 2, 0, 0, 0, 2, 0, 1, 1, 0, 0, 3, 0, 0~
## $ smsa     <int> 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ northcen <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ south    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ west     <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
```

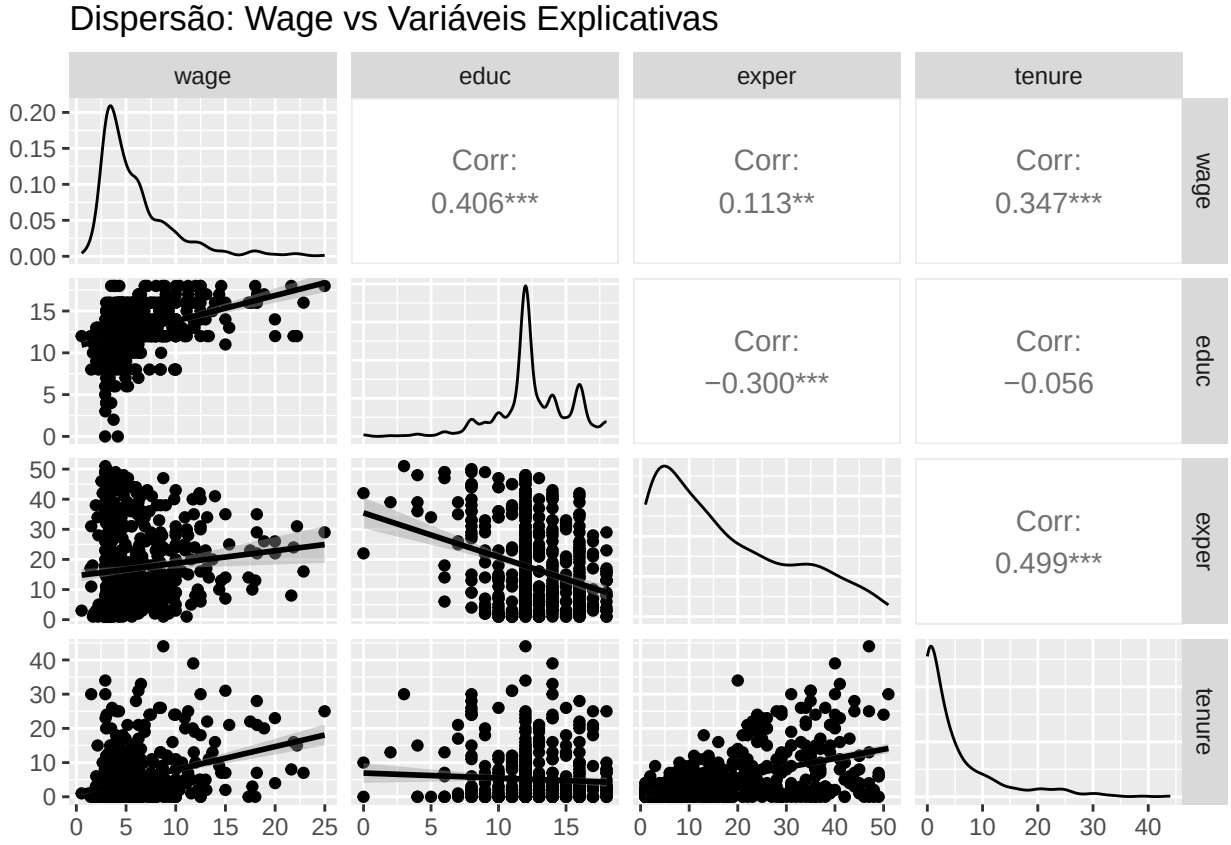
```
## $ construc <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ ndurman  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ trcommpu <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ trade    <int> 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ services <int> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ profserv <int> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1~
## $ profocc  <int> 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1~
## $ clerocc  <int> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0~
## $ servocc  <int> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0~
## $ lwage    <dbl> 1.1314020, 1.1755730, 1.0986120, 1.7917590, 1.6677070, 2.1690~
## $ expersq  <int> 4, 484, 4, 1936, 49, 81, 225, 25, 676, 484, 64, 9, 225, 324, ~
## $ tenursq  <int> 0, 4, 0, 784, 4, 64, 49, 9, 16, 441, 4, 0, 0, 9, 225, 0, 0, 1~
```

```
summary(base1)
```

```
##       wage             educ           exper           tenure
##  Min.   : 0.530   Min.   : 0.00   Min.   : 1.00   Min.   : 0.000
##  1st Qu.: 3.330   1st Qu.:12.00   1st Qu.: 5.00   1st Qu.: 0.000
##  Median : 4.650   Median :12.00   Median :13.50   Median : 2.000
##  Mean   : 5.896   Mean   :12.56   Mean   :17.02   Mean   : 5.105
##  3rd Qu.: 6.880   3rd Qu.:14.00   3rd Qu.:26.00   3rd Qu.: 7.000
##  Max.   :24.980   Max.   :18.00   Max.   :51.00   Max.   :44.000
##     nonwhite          female          married           numdep
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000
##  Median :0.0000   Median :0.0000   Median :1.0000   Median :1.000
##  Mean   :0.1027   Mean   :0.4791   Mean   :0.6084   Mean   :1.044
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:2.000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :6.000
##       smsa           northcen          south             west
##  Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.0000   Median :0.000   Median :0.0000   Median :0.0000
##  Mean   :0.7224   Mean   :0.251   Mean   :0.3555   Mean   :0.1692
##  3rd Qu.:1.0000   3rd Qu.:0.750   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000
##     construc          ndurman          trcommpu           trade
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.00000   Median :0.0000   Median :0.00000   Median :0.0000
##  Mean   :0.04563   Mean   :0.1141   Mean   :0.04373   Mean   :0.2871
##  3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000
##     services         profserv         profocc          clerocc
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
```

```
##  Mean   :0.1008   Mean   :0.2586   Mean   :0.3669   Mean   :0.1673
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##     servocc           lwage              expersq          tenursq
##  Min.   :0.0000   Min.   :-0.6349   Min.   :   1.0   Min.   :   0.00
##  1st Qu.:0.0000   1st Qu.: 1.2030   1st Qu.:  25.0   1st Qu.:   0.00
##  Median :0.0000   Median : 1.5369   Median : 182.5   Median :   4.00
##  Mean   :0.1407   Mean   : 1.6233   Mean   : 473.4   Mean   :  78.15
##  3rd Qu.:0.0000   3rd Qu.: 1.9286   3rd Qu.: 676.0   3rd Qu.:  49.00
##  Max.   :1.0000   Max.   : 3.2181   Max.   :2601.0   Max.   :1936.00
```

```r
# Gera matriz de dispersão e correlação entre as variáveis explicativas
GGally::ggpairs(dplyr::select(base1, wage, educ, exper, tenure),
                lower = list(continuous = "smooth"),
                upper = list(continuous = "cor"),
                title = "Dispersão: Wage vs Variáveis Explicativas")
```

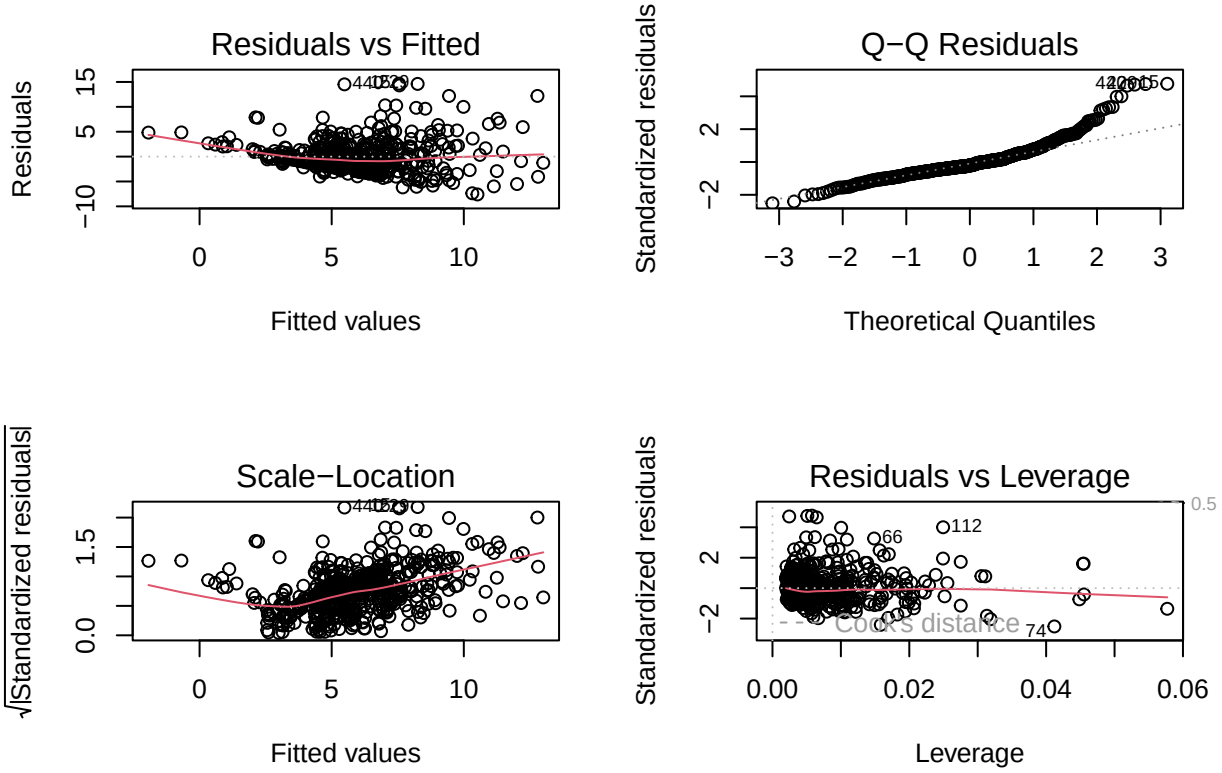



```r
# Ajusta o modelo de regressão múltipla
modelo_multiplo2 <- lm(wage ~ educ + exper + tenure, data = base1)
summary(modelo_multiplo2)
```

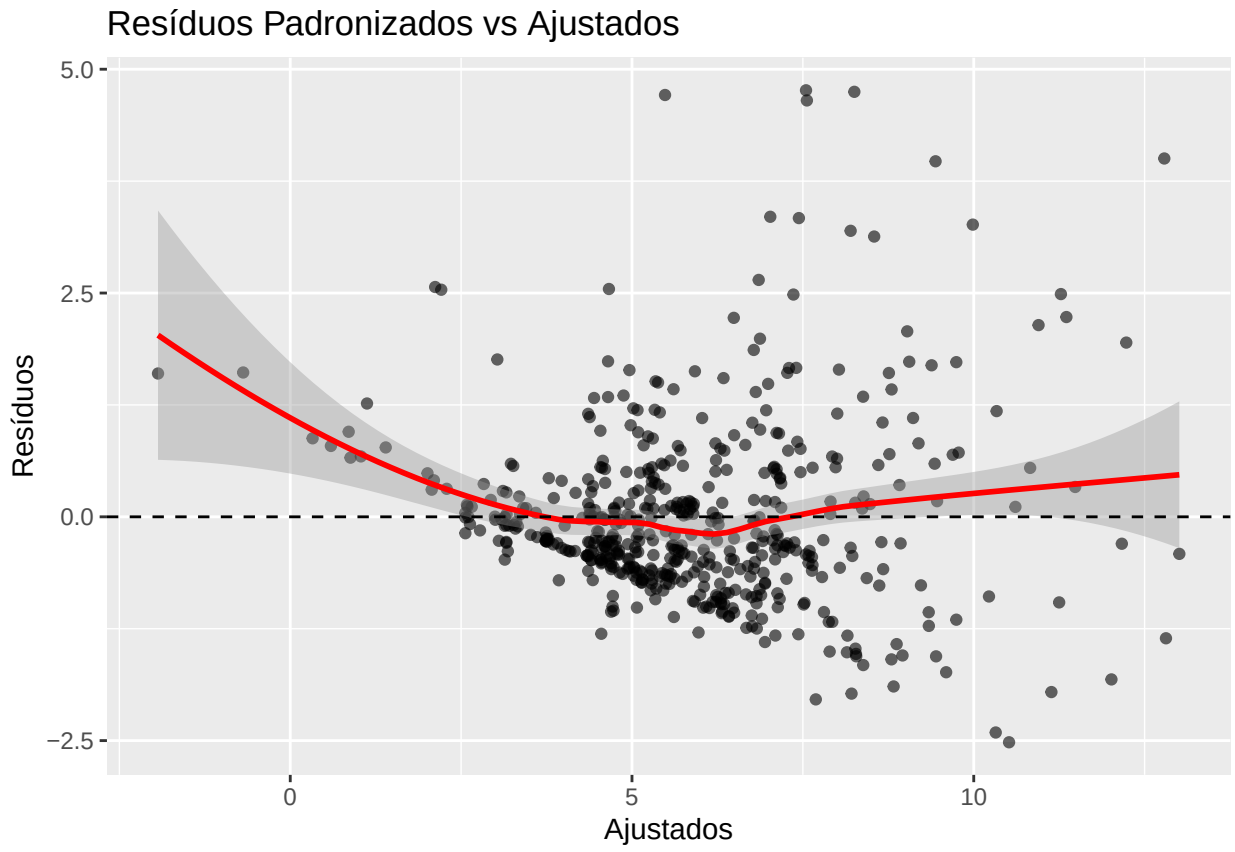
```
##
```

```
## Call:
## lm(formula = wage ~ educ + exper + tenure, data = base1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6068 -1.7747 -0.6279  1.1969 14.6536
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.87273    0.72896  -3.941 9.22e-05 ***
## educ         0.59897    0.05128  11.679  < 2e-16 ***
## exper        0.02234    0.01206   1.853   0.0645 .
## tenure       0.16927    0.02164   7.820 2.93e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.084 on 522 degrees of freedom
## Multiple R-squared:  0.3064, Adjusted R-squared:  0.3024
## F-statistic: 76.87 on 3 and 522 DF,  p-value: < 2.2e-16
```

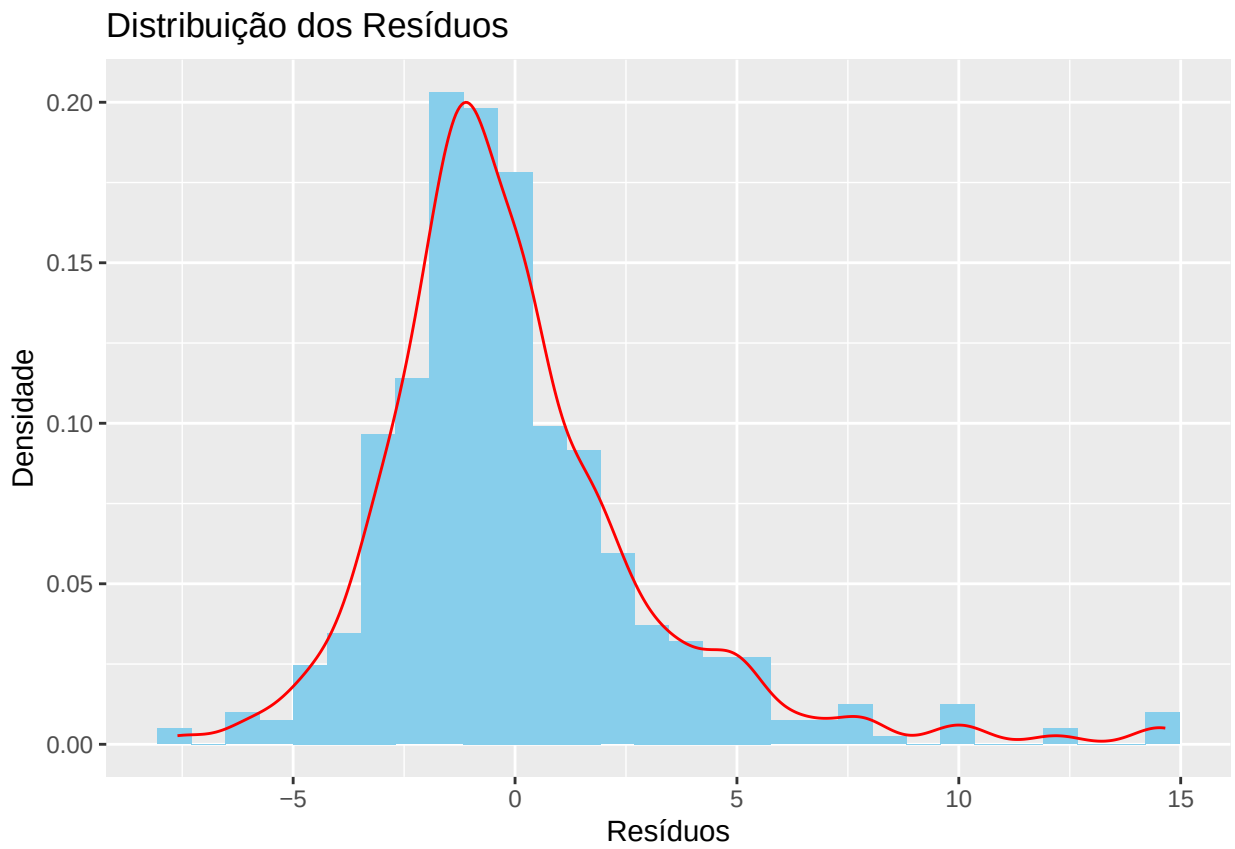
```r
# Diagnóstico gráfico do modelo
par(mfrow = c(2,2))
plot(modelo_multiplo2)
```



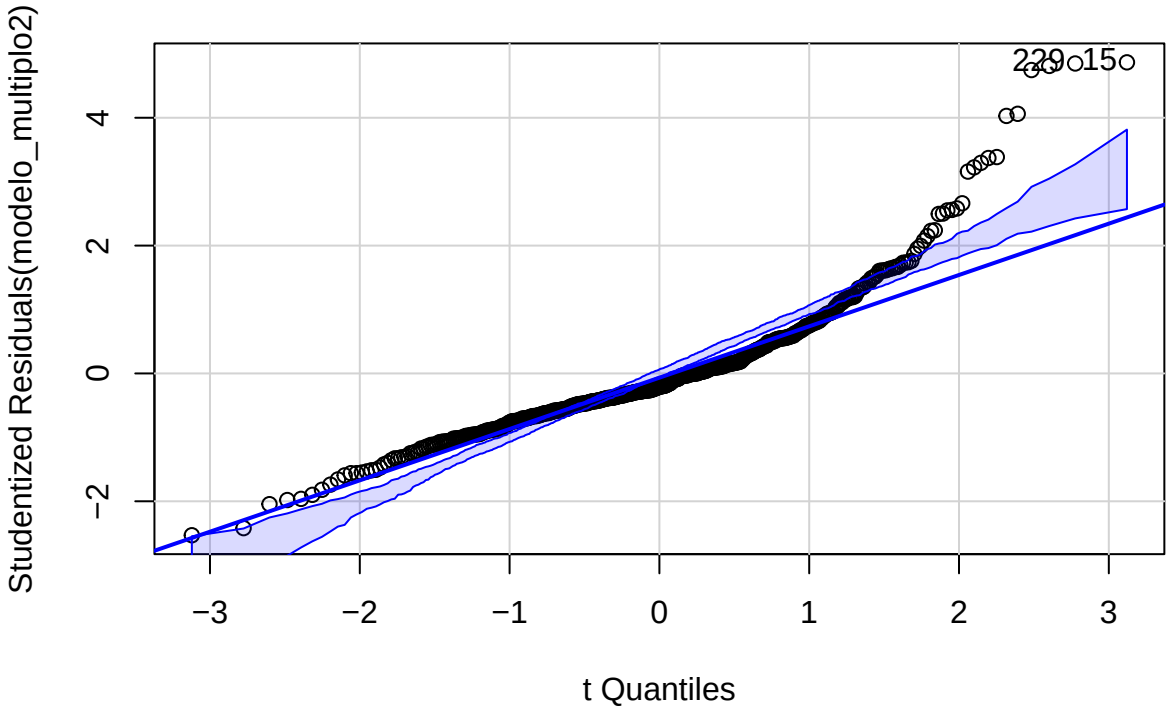
```r
par(mfrow = c(1,1))

# Calcula resíduos, valores ajustados e resíduos padronizados
base1 <- base1 %>%
  mutate(wagehat = fitted(modelo_multiplo2),
         uhat = residuals(modelo_multiplo2),
         std_resid = rstandard(modelo_multiplo2))

# Gráfico de resíduos padronizados vs valores ajustados
ggplot(base1, aes(x = wagehat, y = std_resid)) +
  geom_point(alpha=0.6) + geom_smooth(method="loess", col="red") +
  geom_hline(yintercept=0, linetype="dashed") +
  labs(title="Resíduos Padronizados vs Ajustados", x="Ajustados", y="Resíduos")
```




```r
# Histograma + Curva de densidade dos resíduos
ggplot(base1, aes(x = uhat)) +
  geom_histogram(aes(y=..density..), bins=30, fill="skyblue") +
  geom_density(col="red") +
  labs(title="Distribuição dos Resíduos", x="Resíduos", y="Densidade")
```




```
# QQ-plot para verificar normalidade dos resíduos
car::qqPlot(modelo_multiplo2)
```

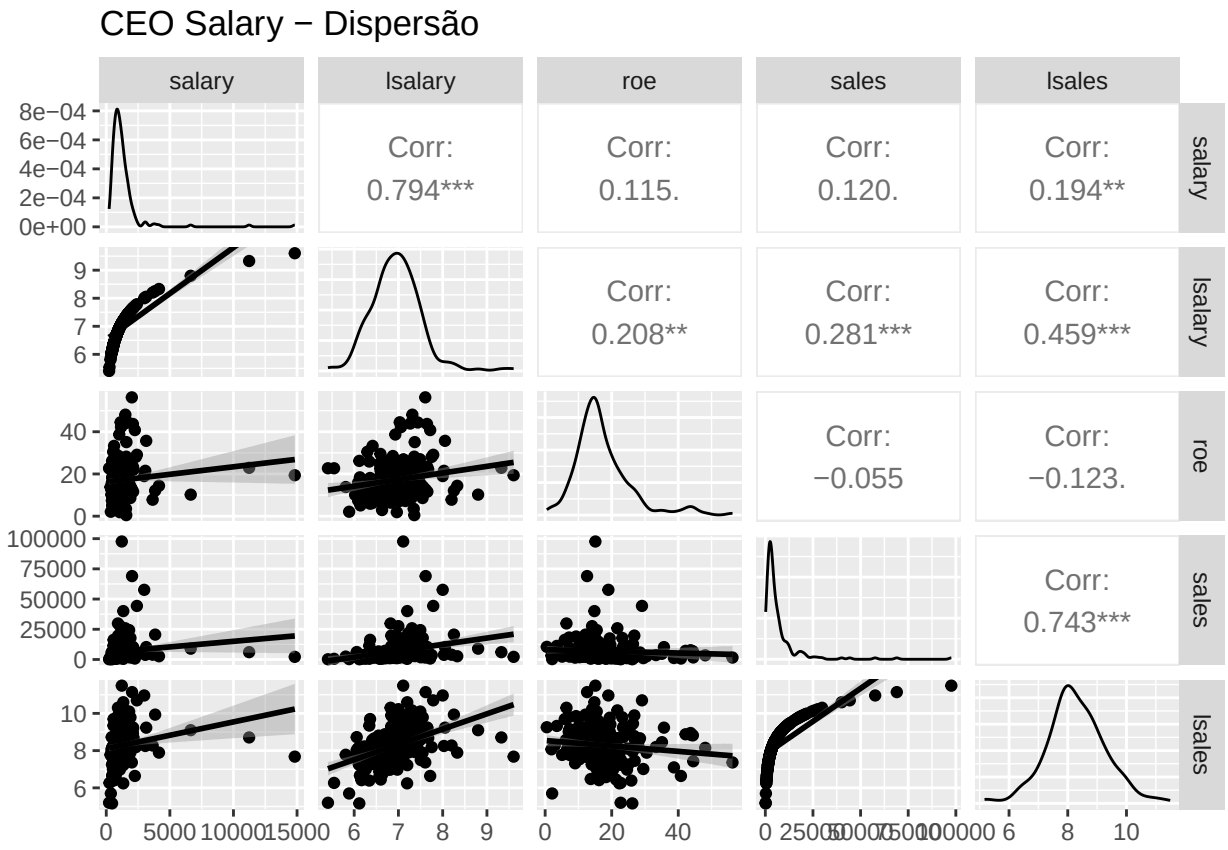



```
## [1]  15 229
```

```
# ======================================
# 4. Estudo Salarial de CEOs (base2)
# ======================================

glimpse(base2)
```

```
## Rows: 209
## Columns: 12
## $ salary    <int> 1095, 1001, 1122, 578, 1368, 1145, 1078, 1094, 1237, 833, 567~
## $ pcsalary <int> 20, 32, 9, -9, 7, 5, 10, 7, 16, 5, 7, -3, -9, 9, 49, 4, 12, 9~
## $ sales     <dbl> 27595.0, 9958.0, 6125.9, 16246.0, 21783.2, 6021.4, 2266.7, 29~
## $ roe       <dbl> 14.1, 10.9, 23.5, 5.9, 13.8, 20.0, 16.4, 16.3, 10.5, 26.3, 25~
## $ pcroe     <dbl> 106.4, -30.6, -16.3, -25.7, -3.0, 1.0, -5.9, -1.6, -70.2, -23~
## $ ros       <int> 191, 13, 14, -21, 56, 55, 62, 44, 37, 37, 109, -10, 41, 44, 6~
## $ indus     <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ finance   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ consprod <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ utility   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ lsalary   <dbl> 6.998509, 6.908755, 7.022868, 6.359574, 7.221105, 7.043160, 6~
## $ lsales    <dbl> 10.225389, 9.206132, 8.720281, 9.695602, 9.988894, 8.703075, ~
```

```r
# Matriz de dispersão entre as variáveis principais
GGally::ggpairs(dplyr::select(base2, salary, lsalary, roe, sales, lsales),
                lower = list(continuous = "smooth"),
                upper = list(continuous = "cor"),
                title = "CEO Salary - Dispersão")
```



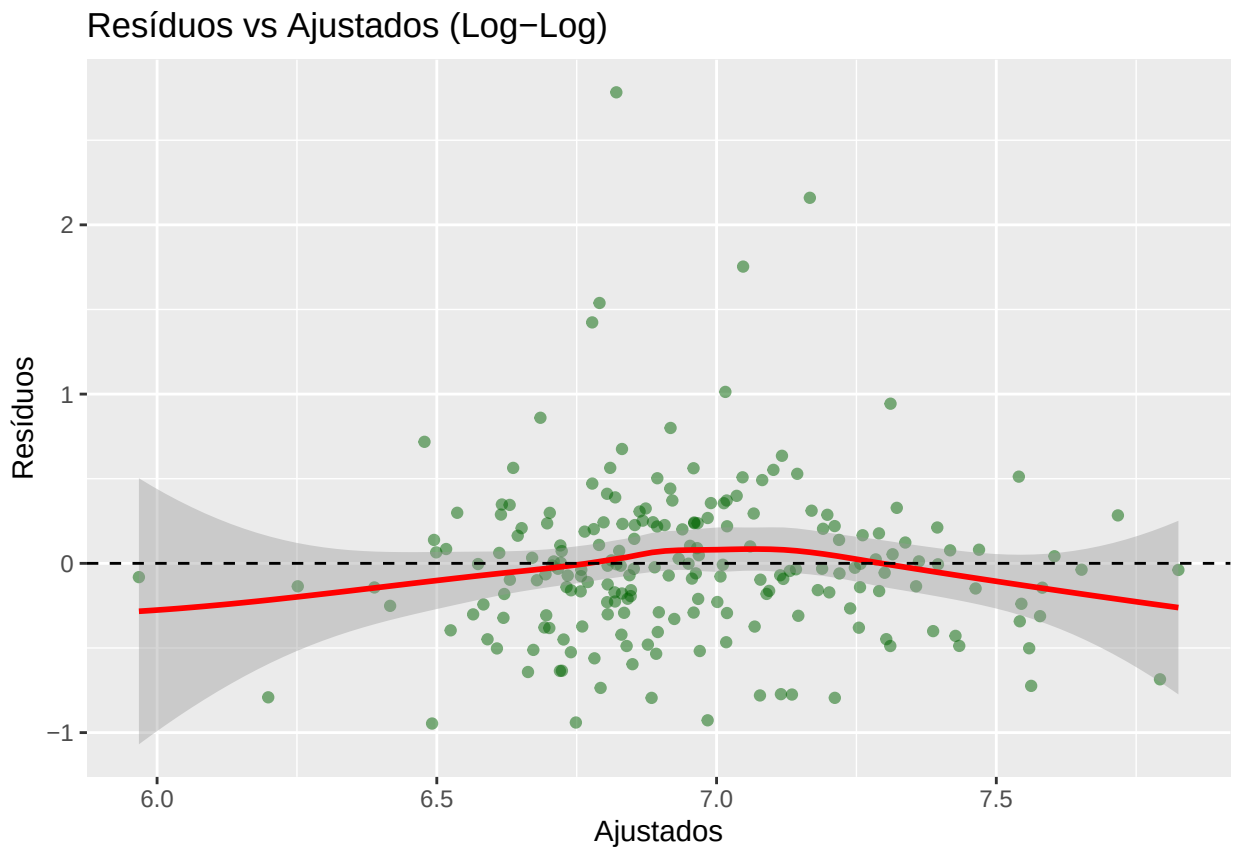

```r
# Ajusta diferentes modelos com e sem log-transformações
modelos_ceo <- list(
  linear = lm(salary ~ roe + sales, data = base2),
  linear_log = lm(salary ~ roe + lsales, data = base2),
  log_linear = lm(lsalary ~ roe + sales, data = base2),
  log_log = lm(lsalary ~ roe + lsales, data = base2)
)

# Mostra os resultados de todos os modelos
lapply(modelos_ceo, summary)
```

```
## $linear
##
## Call:
## lm(formula = salary ~ roe + sales, data = base2)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1501.8  -492.6  -232.0   123.3 13575.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.306e+02  2.239e+02   3.710 0.000267 ***
## roe         1.963e+01  1.108e+01   1.772 0.077823 .
## sales       1.634e-02  8.874e-03   1.842 0.066973 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1359 on 206 degrees of freedom
## Multiple R-squared:  0.02917,    Adjusted R-squared:  0.01975
## F-statistic: 3.095 on 2 and 206 DF,  p-value: 0.04739
##
##
## $linear_log
##
## Call:
## lm(formula = salary ~ roe + lsales, data = base2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1024.1  -443.2  -223.3    68.8 13666.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1482.29     815.97  -1.817   0.0707 .
## roe            22.67      10.98   2.065   0.0402 *
## lsales        286.26      92.33   3.100   0.0022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1339 on 206 degrees of freedom
## Multiple R-squared:  0.05718,    Adjusted R-squared:  0.04803
## F-statistic: 6.247 on 2 and 206 DF,  p-value: 0.002323
##
##
## $log_linear
##
## Call:
## lm(formula = lsalary ~ roe + sales, data = base2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.52016 -0.27115 -0.00942  0.25605  2.69491
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.585e+00  8.750e-02  75.258  < 2e-16 ***
## roe         1.494e-02  4.329e-03   3.452 0.000674 ***
## sales       1.565e-05  3.468e-06   4.512 1.08e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.531 on 206 degrees of freedom
## Multiple R-squared:  0.1295, Adjusted R-squared:  0.121
## F-statistic: 15.32 on 2 and 206 DF,  p-value: 6.264e-07
##
##
## $log_log
##
## Call:
## lm(formula = lsalary ~ roe + lsales, data = base2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9464 -0.2888 -0.0322  0.2261  2.7830
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.362167   0.293878  14.843  < 2e-16 ***
## roe         0.017872   0.003955   4.519 1.05e-05 ***
## lsales      0.275087   0.033254   8.272 1.62e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4822 on 206 degrees of freedom
## Multiple R-squared:  0.282,  Adjusted R-squared:  0.275
## F-statistic: 40.45 on 2 and 206 DF,  p-value: 1.519e-15
```

```r
# Gráfico diagnóstico do modelo log-log
modelo_ref <- modelos_ceo$log_log
base2 <- base2 %>%
  mutate(salary_hat = fitted(modelo_ref),
         resid = residuals(modelo_ref))

ggplot(base2, aes(x = salary_hat, y = resid)) +
  geom_point(alpha = 0.5, color = "darkgreen") +
  geom_smooth(method = "loess", color = "red") +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(title = "Resíduos vs Ajustados (Log-Log)", x = "Ajustados", y = "Resíduos")
```



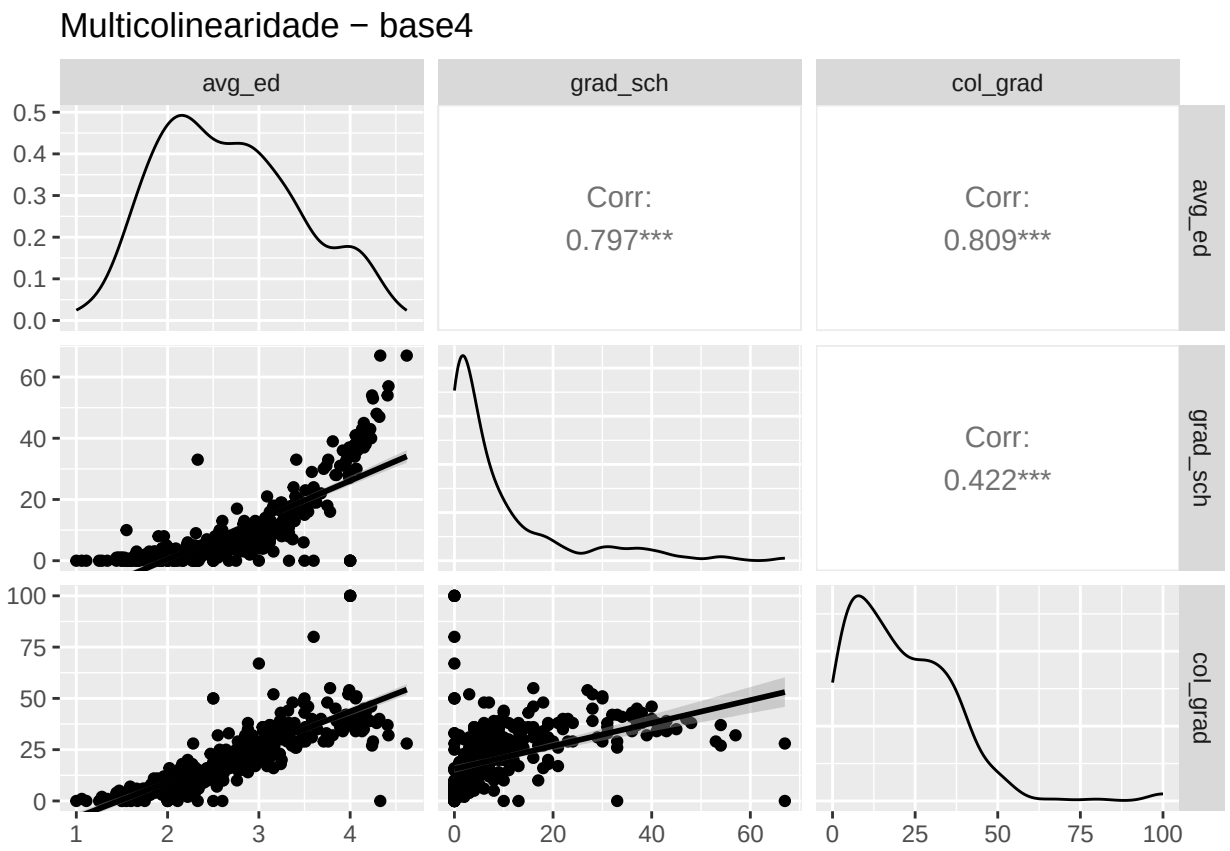

```r
# ======================================================
# 5. Colinearidade Perfeita (variáveis female e male)
# ======================================================

if("female" %in% colnames(base1)) {
  # Garante que female está codificada como 0/1
  if(!is.numeric(base1$female)) {
    base1$female <- as.numeric(as.character(base1$female))
  }

  # Cria a variável male como complementar
  base1 <- base1 %>% mutate(male = 1 - female)

  # Ajusta modelos separadamente para evitar colinearidade
  modelo_no_col_fem <- lm(wage ~ educ + female, data = base1)
  modelo_no_col_male <- lm(wage ~ educ + male, data = base1)

  # Modelos com colinearidade perfeita (apenas para teste)
  modelo_colinearidade <- lm(wage ~ educ + female + male, data = base1)
  modelo_no_intercepto <- lm(wage ~ 0 + educ + female + male, data = base1)

  # Mostra modelos válidos
  summary(modelo_no_col_fem)
```

```r
  summary(modelo_no_col_male)

  # Exporta com stargazer
  stargazer(modelo_no_col_fem, modelo_no_col_male, type = "text",
            title = "Modelos sem Colinearidade Perfeita")
}
```

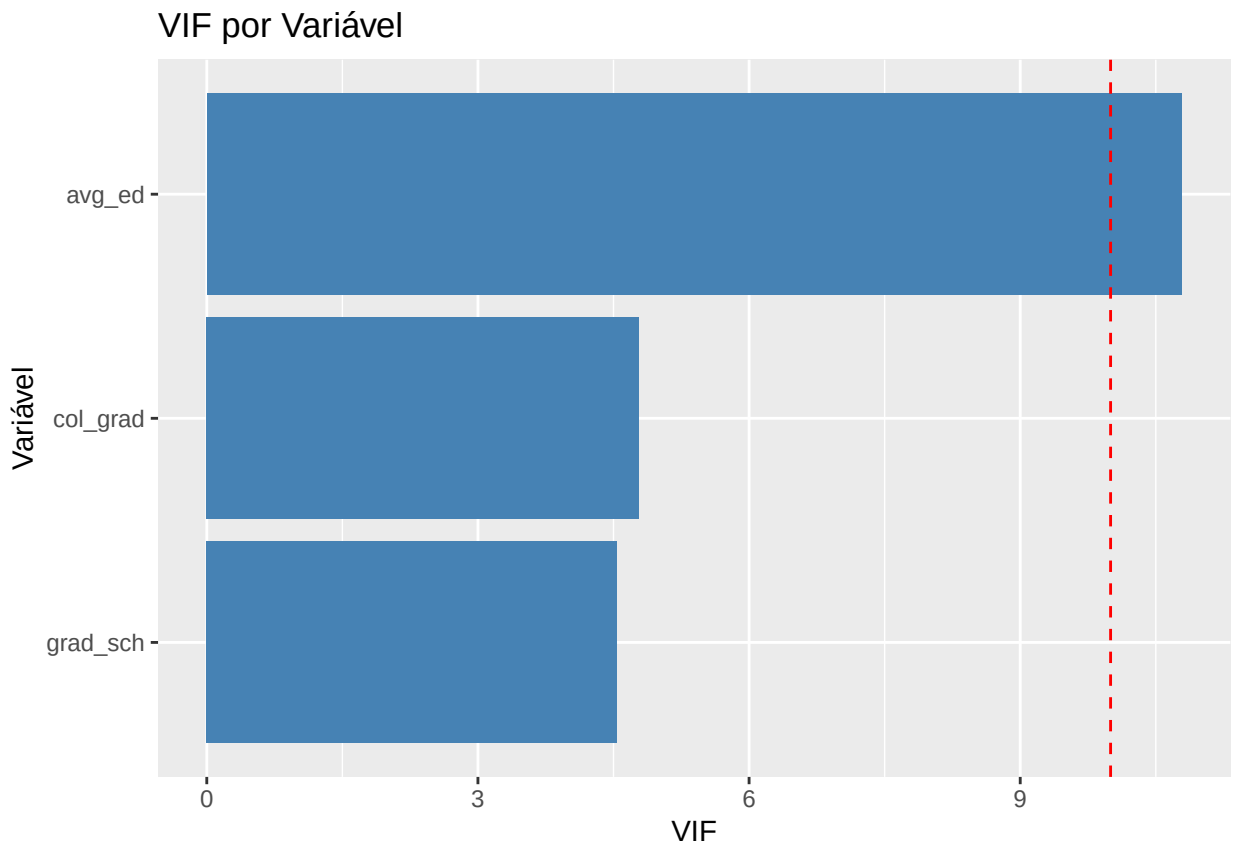
```
##
## Modelos sem Colinearidade Perfeita
## ============================================================
##                                    Dependent variable:
##                                ----------------------------
##                                            wage
##                                     (1)            (2)
## ------------------------------------------------------------
## educ                               0.506***      0.506***
##                                    (0.050)        (0.050)
##
## female                            -2.273***
##                                    (0.279)
##
## male                                             2.273***
##                                                   (0.279)
##
## Constant                           0.623         -1.651**
##                                    (0.673)        (0.652)
##
## ------------------------------------------------------------
## Observations                         526            526
## R2                                  0.259          0.259
## Adjusted R2                         0.256          0.256
## Residual Std. Error (df = 523)      3.186          3.186
## F Statistic (df = 2; 523)         91.315***      91.315***
## ============================================================
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

```r
# ===================================
# 6. Multicolinearidade (base4)
# ===================================

# Seleciona variáveis relevantes
base4_sel <- dplyr::select(base4, api00, avg_ed, grad_sch, col_grad) %>% na.omit()

# Matriz de correlação entre variáveis explicativas
GGally::ggpairs(dplyr::select(base4_sel, avg_ed, grad_sch, col_grad),
                lower = list(continuous = "smooth"),
```

```
                upper = list(continuous = "cor"),
                title = "Multicolinearidade - base4")
```




```
# Ajusta modelo e calcula VIF
modelo_vif_alto <- lm(api00 ~ avg_ed + grad_sch + col_grad, data = base4_sel)
vif_valores <- car::vif(modelo_vif_alto)

# Gráfico dos VIFs
vif_df <- data.frame(Variável = names(vif_valores), VIF = vif_valores)

ggplot(vif_df, aes(x = reorder(Variável, VIF), y = VIF)) +
  geom_col(fill = "steelblue") +
  geom_hline(yintercept = 10, linetype = "dashed", color = "red") +
  coord_flip() +
  labs(title = "VIF por Variável", y = "VIF", x = "Variável")
```



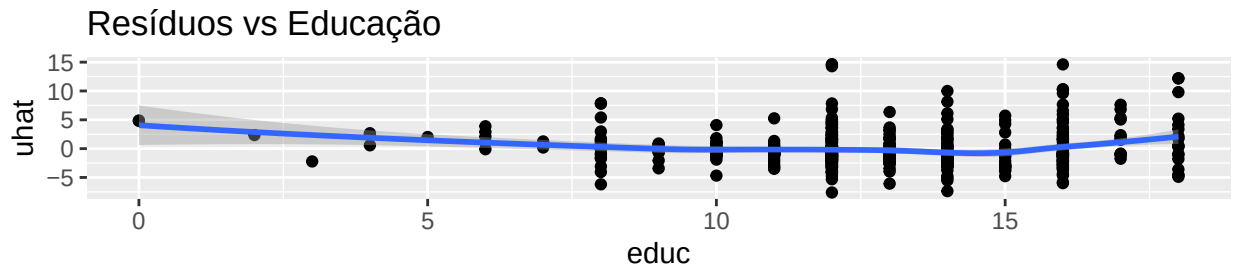

```
# Reajusta modelo removendo variável com VIF elevado
summary(lm(api00 ~ grad_sch + col_grad, data = base4_sel))
```

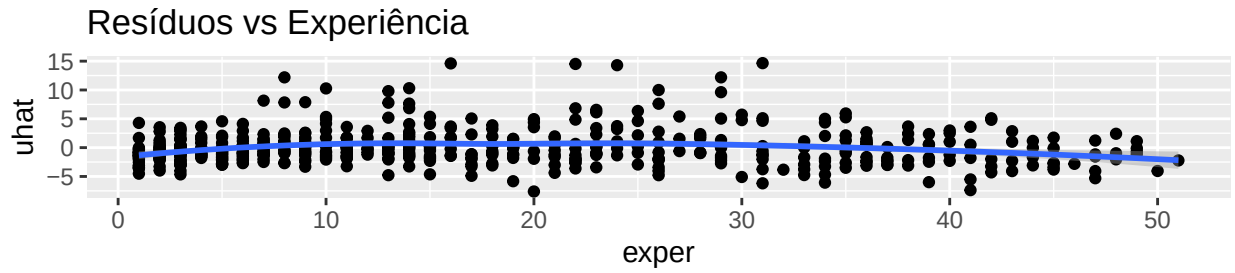
```
##
## Call:
## lm(formula = api00 ~ grad_sch + col_grad, data = base4_sel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -338.41  -67.03    1.71   71.79  350.88
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 529.1246     8.1571  64.866   <2e-16 ***
## grad_sch      5.9499     0.4492  13.247   <2e-16 ***
## col_grad      3.0789     0.3389   9.084   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 97.42 on 378 degrees of freedom
## Multiple R-squared:  0.5364, Adjusted R-squared:  0.534
## F-statistic: 218.7 on 2 and 378 DF,  p-value: < 2.2e-16
```

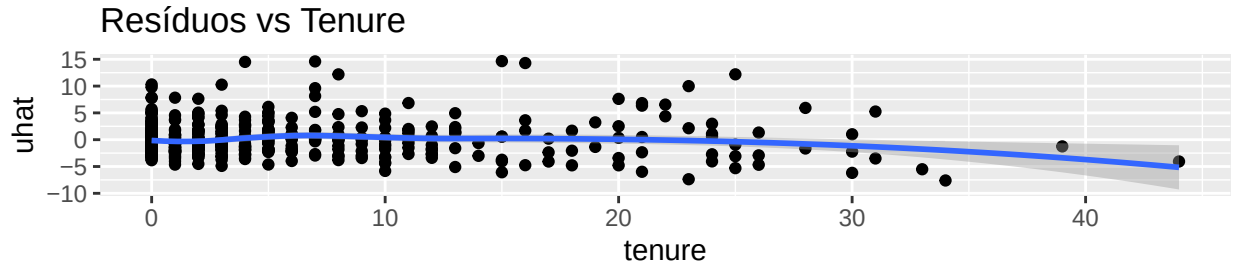
```r
# ======================================
# 7. Viés por Variável Omitida (base3)
# ======================================

# Seleciona colunas e remove NAs
base3_sel <- dplyr::select(base3, wage, educ, abil) %>% na.omit()

# Modelo verdadeiro com abil
modelo_verdadeiro <- lm(wage ~ educ + abil, data = base3_sel)

# Modelo onde abil é função de educ
modelo_abil <- lm(abil ~ educ, data = base3_sel)

# Modelo omitindo abil
modelo_omitido <- lm(wage ~ educ, data = base3_sel)

# Cálculo do viés
beta2 <- coef(modelo_verdadeiro)["abil"]
delta1 <- coef(modelo_abil)["educ"]
bias <- beta2 * delta1
cat("Viés estimado:", round(bias, 4), "\n")
```

```
## Viés estimado: 0.2388
```

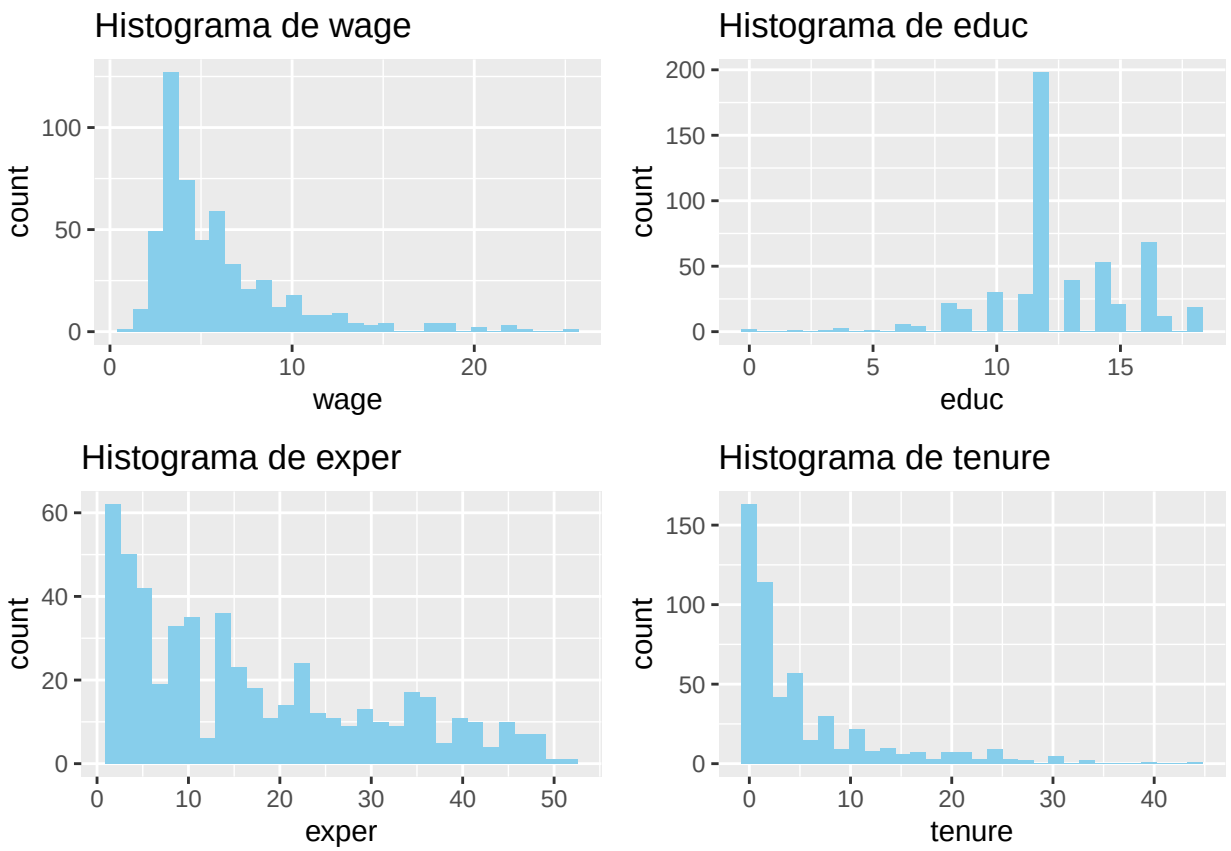
```r
# ======================================================
# 8. Heteroscedasticidade: Resíduos vs Variáveis
# ======================================================

# Gráficos de resíduos em função de cada explicativa
p1 <- ggplot(base1, aes(x = educ, y = uhat)) + geom_point() +
  geom_smooth(method = "loess") +
  labs(title = "Resíduos vs Educação")

p2 <- ggplot(base1, aes(x = exper, y = uhat)) + geom_point() +
  geom_smooth(method = "loess") +
  labs(title = "Resíduos vs Experiência")

p3 <- ggplot(base1, aes(x = tenure, y = uhat)) + geom_point() +
  geom_smooth(method = "loess") +
  labs(title = "Resíduos vs Tenure")

gridExtra::grid.arrange(p1, p2, p3, ncol = 1)
```



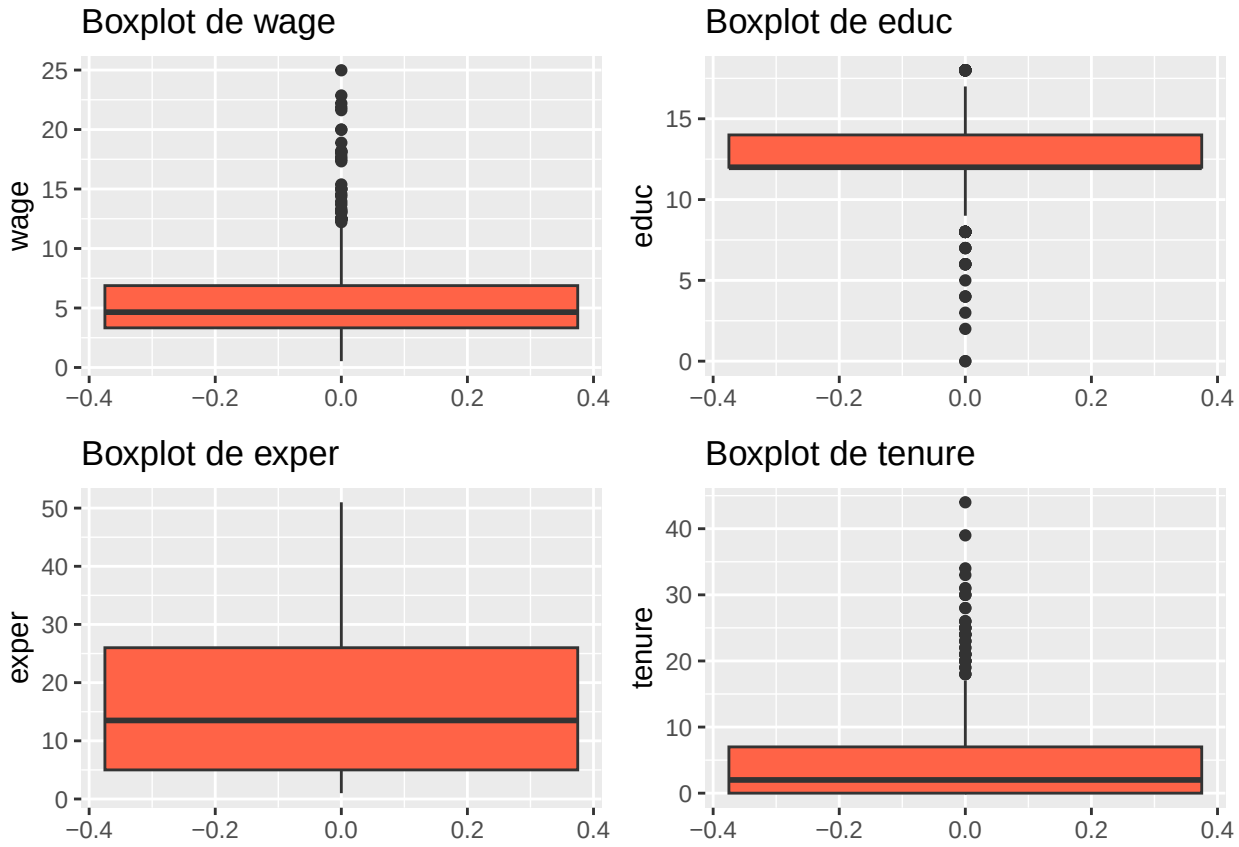





```r
# ======================================
# 9. Histogramas e Boxplots (base1)
# ======================================

# Gera visualizações para variáveis contínuas
vars_cont <- c("wage", "educ", "exper", "tenure")

# Histogramas
plots_hist <- lapply(vars_cont, function(v) {
  ggplot(base1, aes_string(x = v)) +
    geom_histogram(bins = 30, fill = "skyblue") +
    labs(title = paste("Histograma de", v))
})

# Boxplots
plots_box <- lapply(vars_cont, function(v) {
  ggplot(base1, aes_string(y = v)) +
    geom_boxplot(fill = "tomato") +
    labs(title = paste("Boxplot de", v))
})

# Exibição
gridExtra::grid.arrange(grobs = plots_hist, ncol = 2)
```




```
gridExtra::grid.arrange(grobs = plots_box, ncol = 2)
```


Boxplot de wage
wage
25
20
15
10
5
0
−0.4
−0.2
0.0
0.2
0.4
Boxplot de educ
educ
15
10
5
0
Boxplot de exper
exper
50
40
30
20
10
0
Boxplot de tenure
tenure
40
30
20
10
0