Regressão Linear e Regularização

Tiago Mendonça dos Santos

tiagoms.com
tiagomendonca
tiagoms1@insper.edu.br

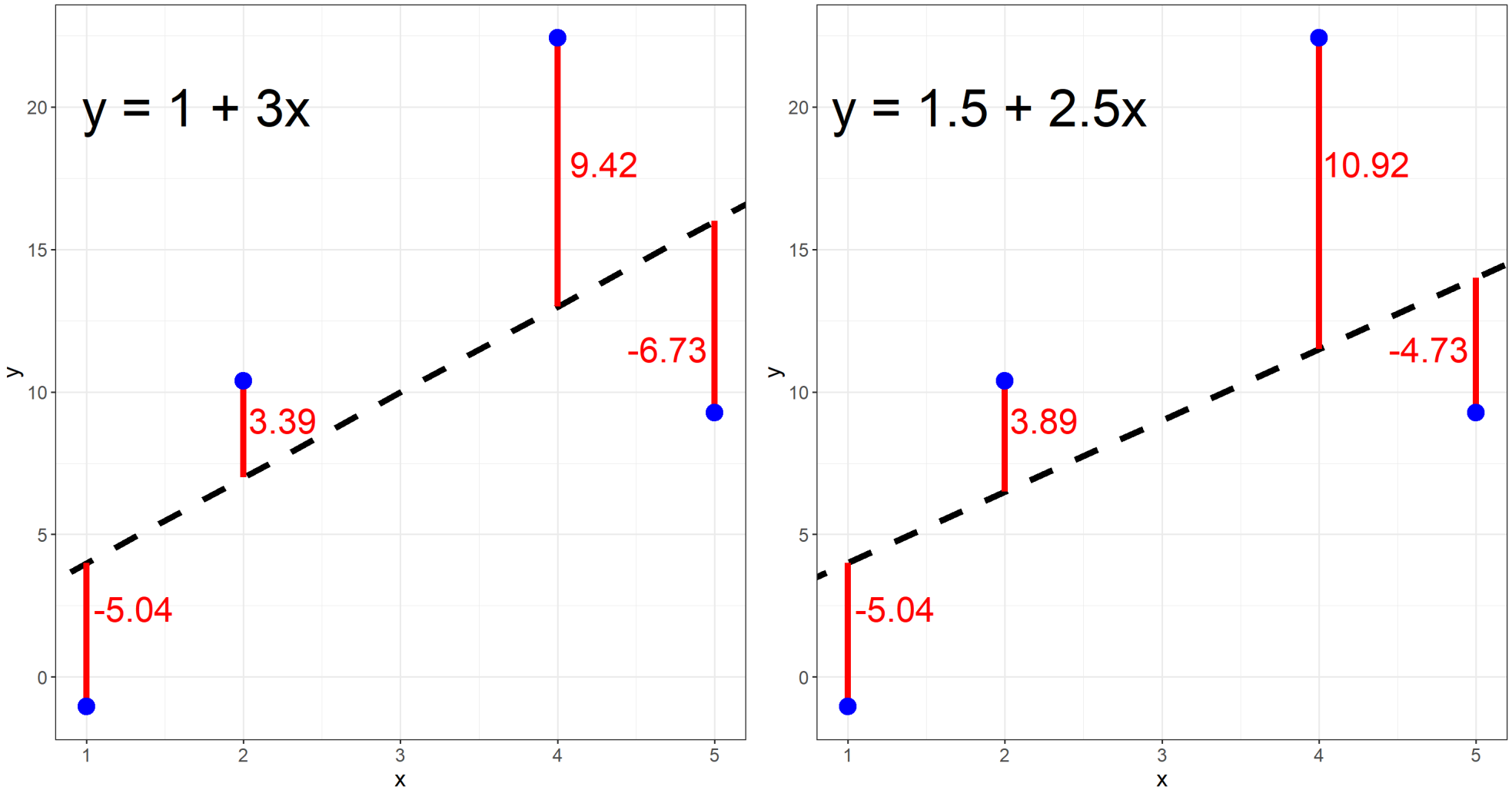
Objetivos

Ao final dessa aula você deverá ser capaz de:

- interpretar, ajustar e aplicar um modelo linear
- compreender e aplicar técnicas de seleção de variável
- relacionar técnicas de regularização com o *trade-off* viés-variância
- ajustar, definir hiperparâmetros e aplicar modelos com técnicas de regularização
- comparar modelos de regressão linear e regularizados

Introdução

Regressão


y = 1 + 3x
9.42
-6.73
3.39
-5.04
y = 1.5 + 2.5x
10.92
-4.73
3.89
-5.04
y
x
0
5
10
15
20
1
2
3
4
5

Resíduo

Chamamos as diferenças apresentadas anteriormente, em vermelho, de **resíduos**. Essa quantidade é definida, de forma matemática, como:

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i).$$

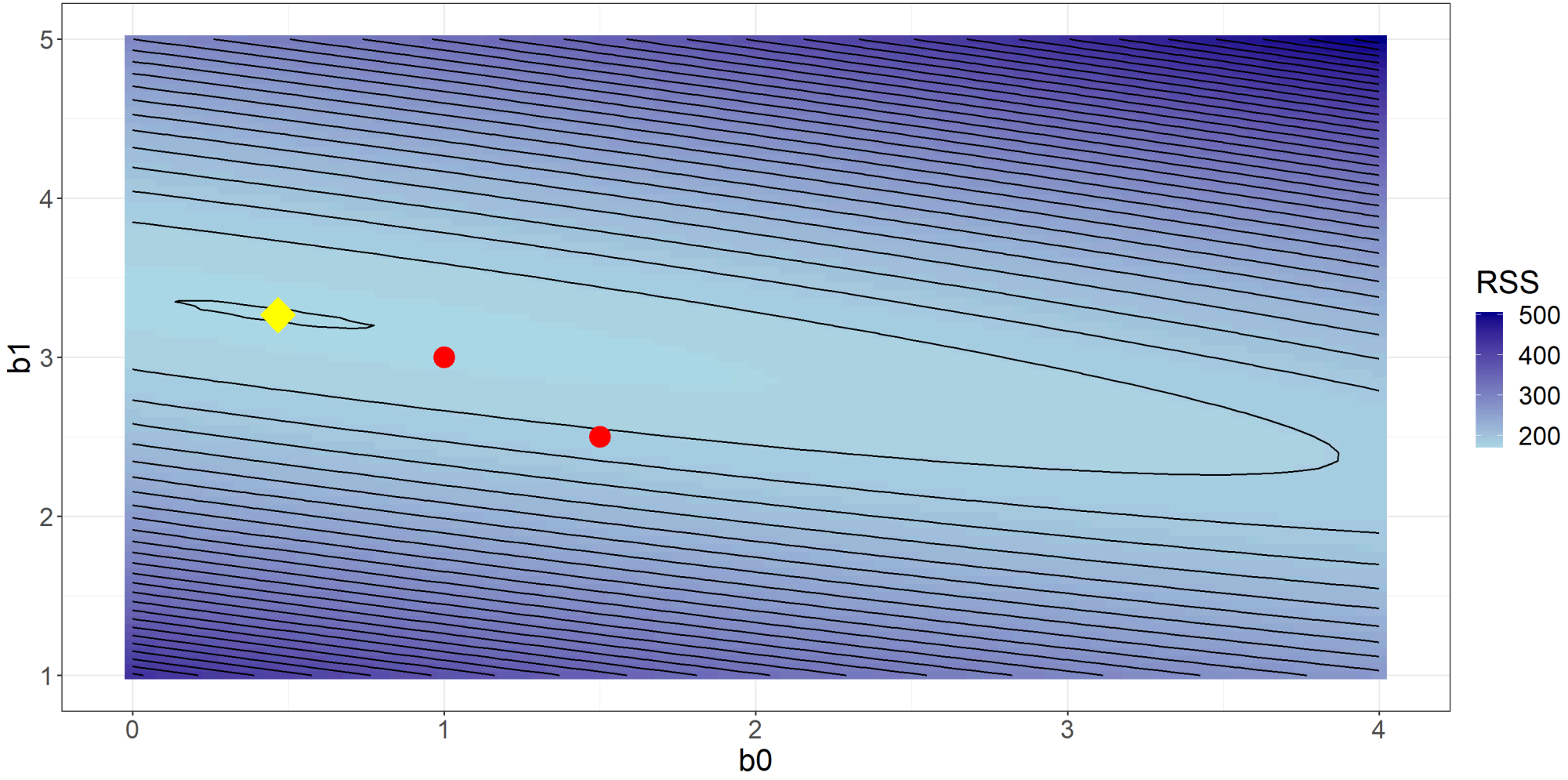
Podemos definir a soma dos quadrados dos resíduos (RSS - *residual sum of squares*) da seguinte forma:

$$\text{RSS} = e_1^2 + \cdots + e_n^2$$

$$\text{RSS} = \left(y_1 - (\hat{\beta}_0 + \hat{\beta}_1 x_1)\right)^2 + \cdots + \left(y_n - (\hat{\beta}_0 + \hat{\beta}_1 x_n)\right)^2$$

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2$$

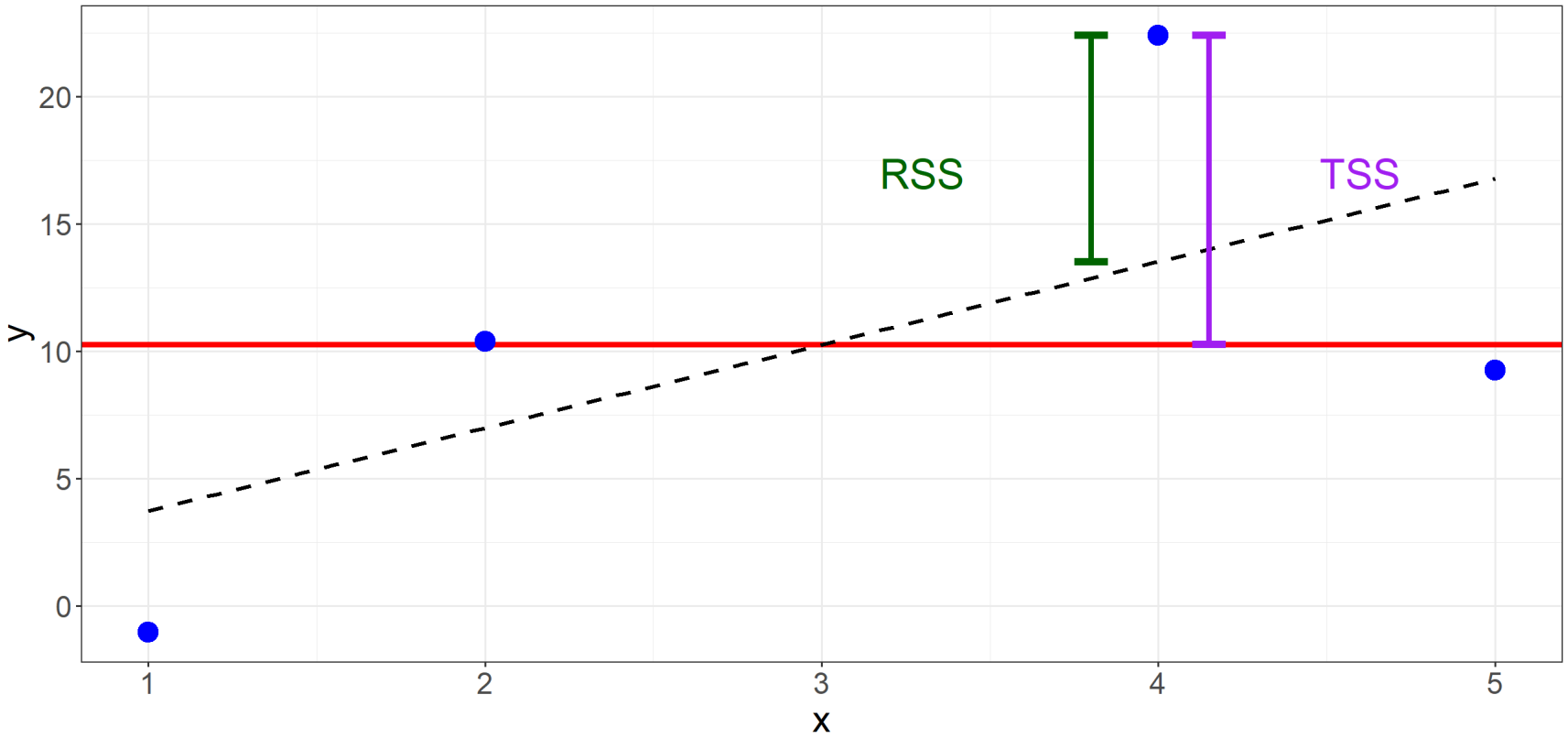
Regressão



R^2

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

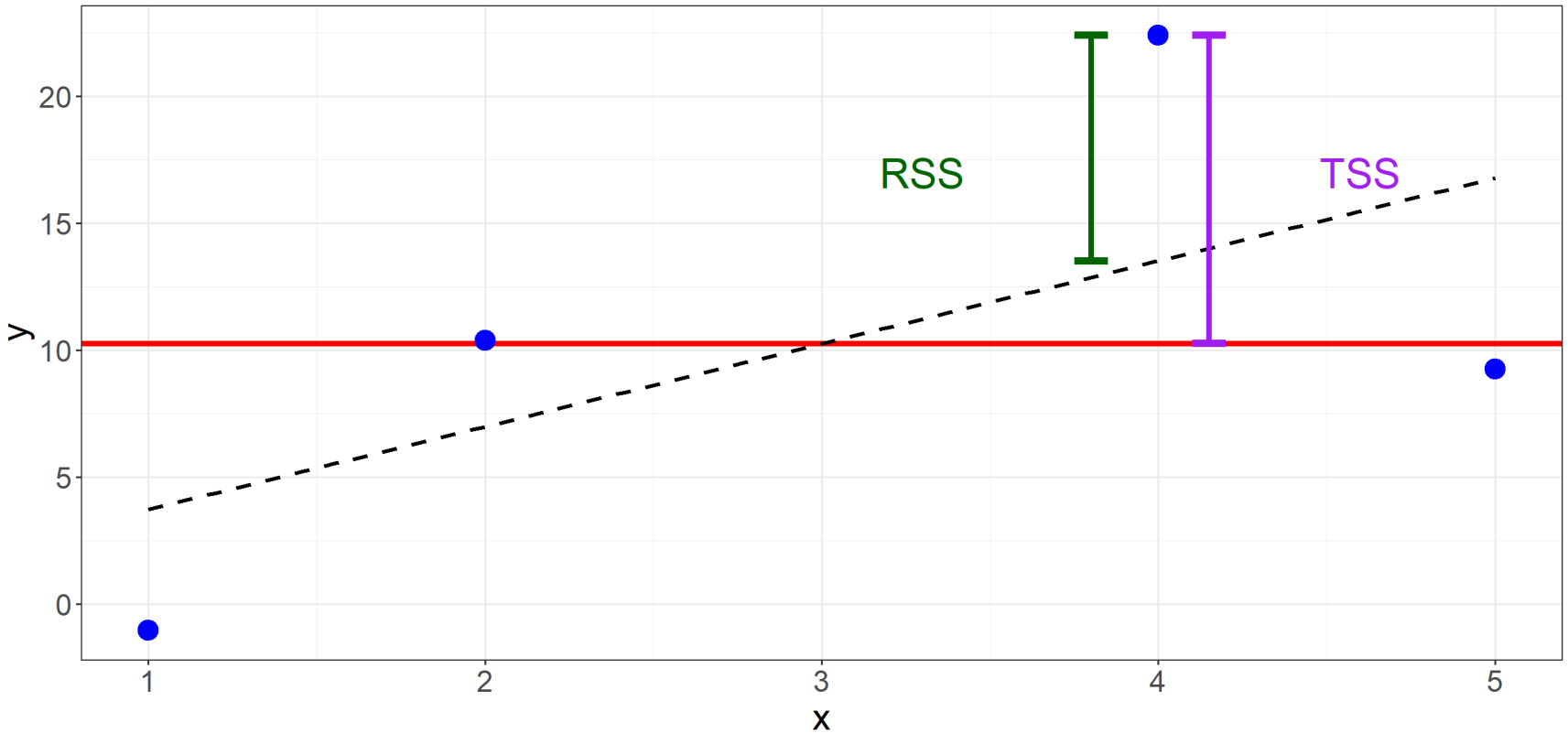
em que $\text{RSS} = \sum(y_i - \hat{y})^2$ e $\text{TSS} = \sum(y_i - \bar{y})^2$.



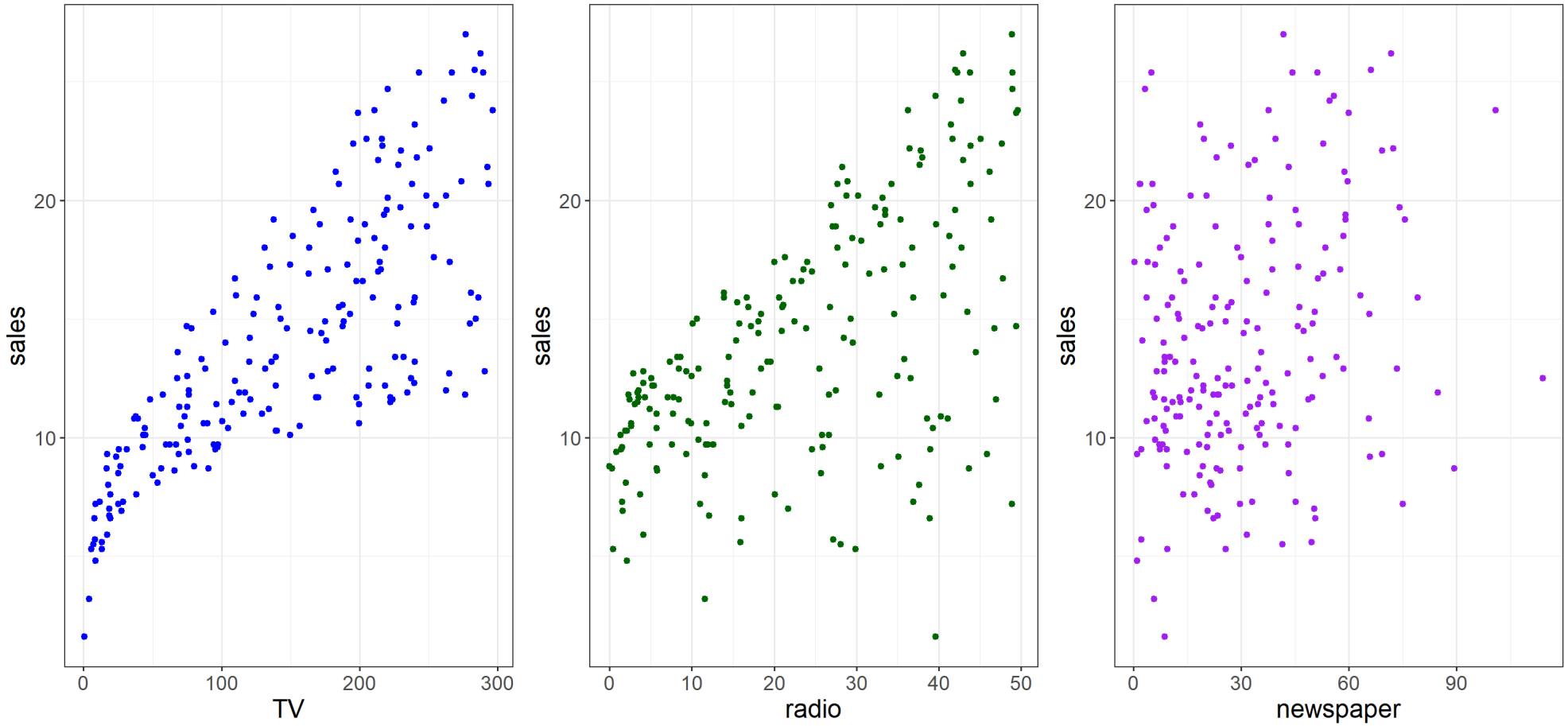
R^2 Ajustado

$$\mathrm{R}^2\text{Ajustado} = 1 - \frac{\mathrm{RSS}/(n-d-1)}{\mathrm{TSS}/(n-1)}$$

em que $\mathrm{RSS} = \sum(y_i - \hat{y})^2$, $\mathrm{TSS} = \sum(y_i - \bar{y})^2$ e d é o número de variáveis no modelo.

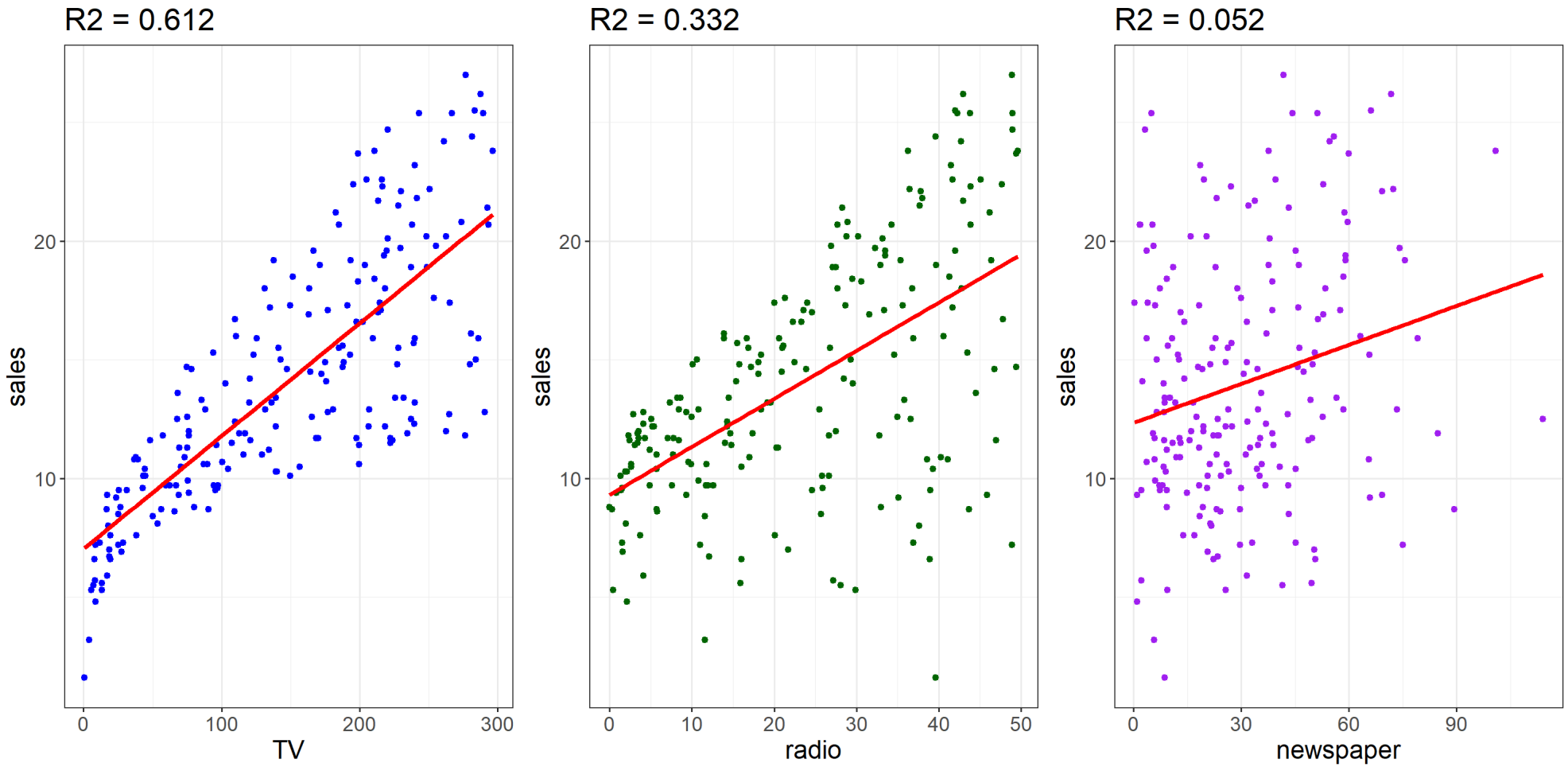


Advertising[1]



[1] Exemplo retirado do livro *An Introduction to Statistical Learning with Applications in R*.

Advertising[1]



[1] Exemplo retirado do livro *An Introduction to Statistical Learning with Applications in R*.

R^2

Insper

```
fit1 <- lm(sales ~ TV, data = advertising)

y_pred <- predict(fit1, advertising)
y_bar  <- mean(advertising$sales)

RSS <- sum((advertising$sales - y_pred)^2)
TSS <- sum((advertising$sales - y_bar)^2)

1 - RSS/TSS
```

```
## [1] 0.6118751
```

```
summary(fit1)$r.squared
```

```
## [1] 0.6118751
```

R^2 Ajustado

```
y_pred <- predict(fit1, advertising)
y_bar  <- mean(advertising$sales)

RSS <- sum((advertising$sales - y_pred)^2)
TSS <- sum((advertising$sales - y_bar)^2)

1 - (RSS/(nrow(advertising) - 1 - 1))/(TSS/(nrow(advertising) - 1))
```

```
## [1] 0.6099148
```

```
summary(fit1)$adj.r.squared
```

```
## [1] 0.6099148
```

Regressão Linear Múltipla

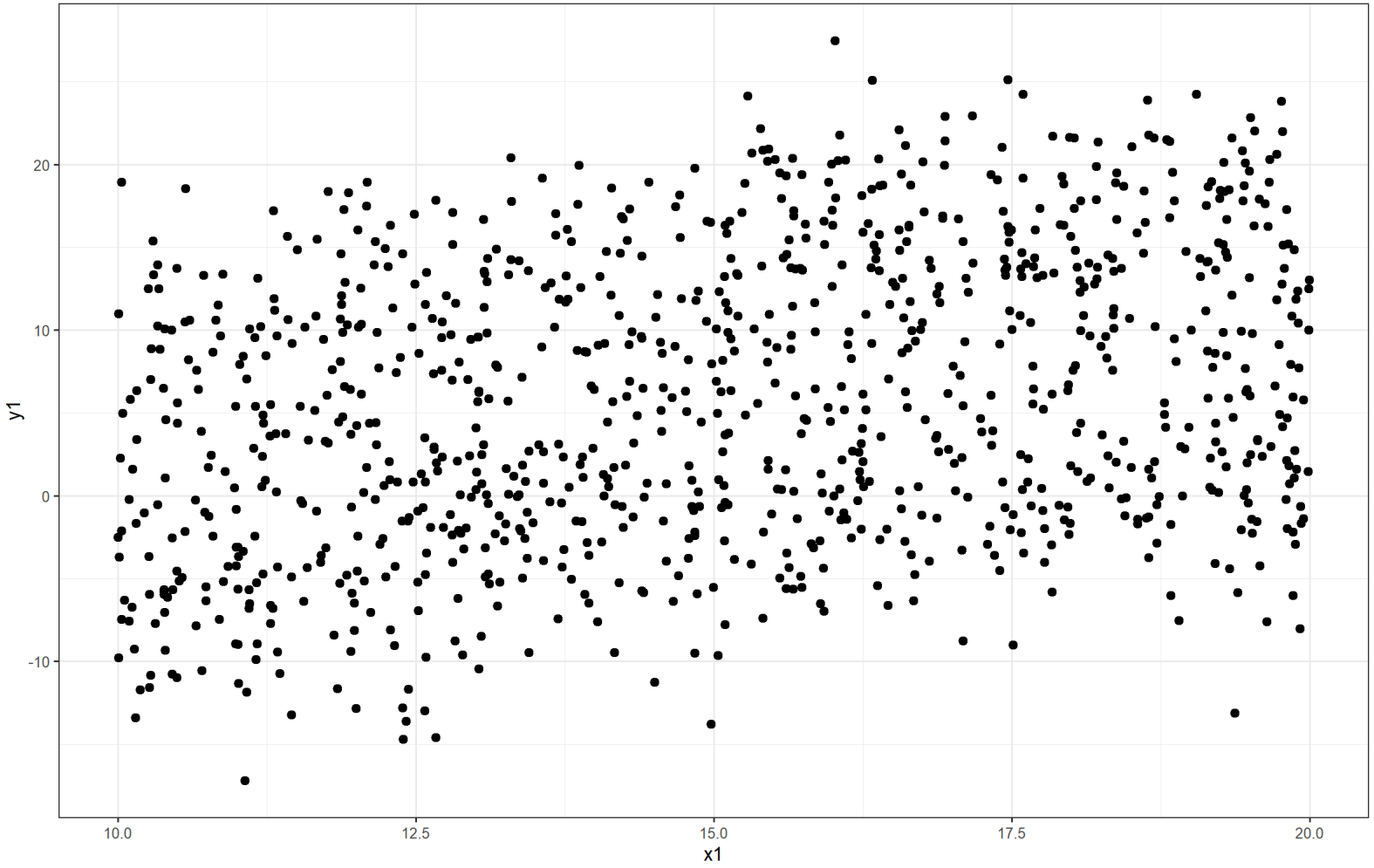
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Vamos considerar o seguinte exemplo:

```
library(readxl)

dados <- read_xlsx("dados/dados.xlsx")
```

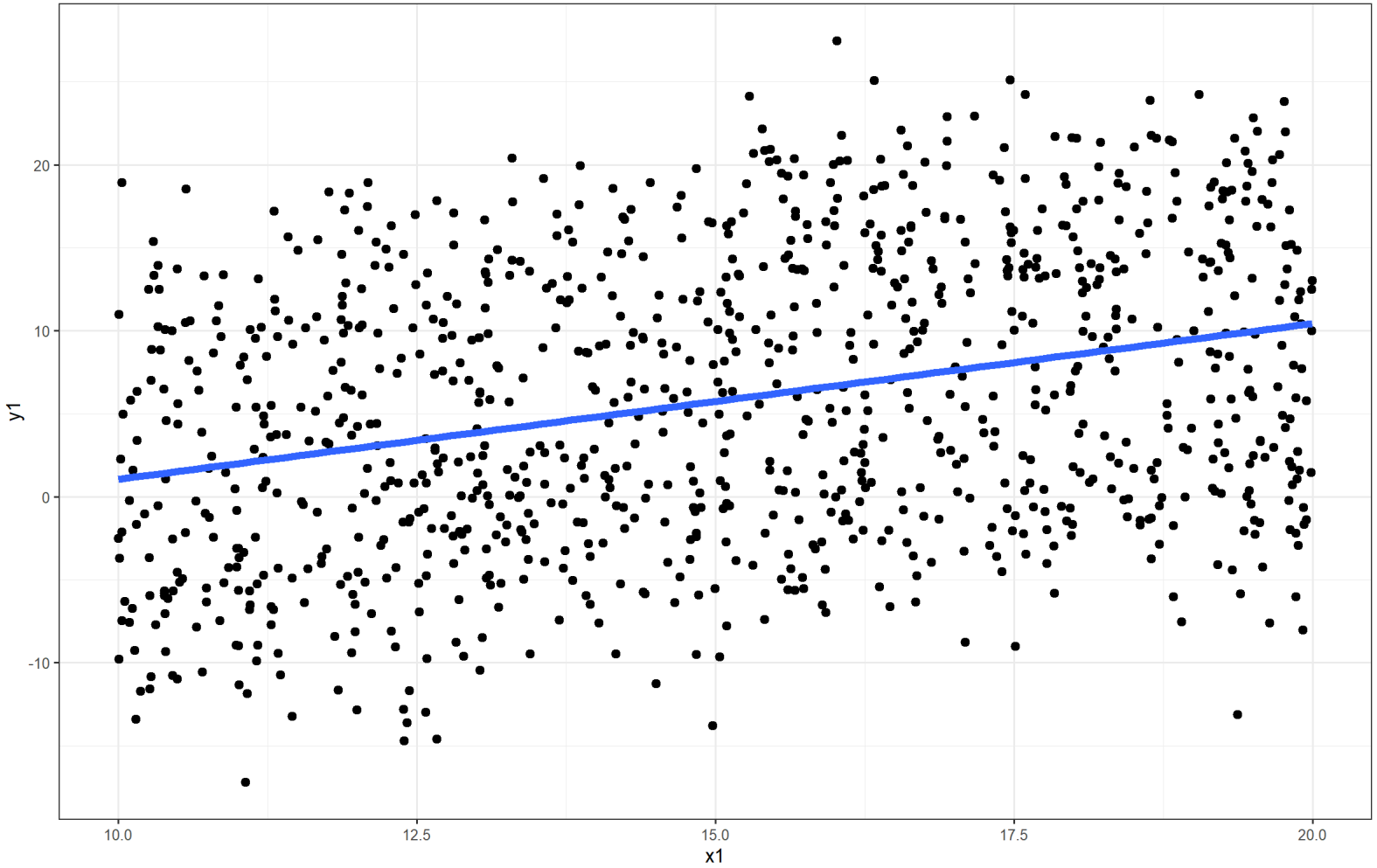
x1	x2	y1	y2
11.14	B	2.88	4.04
16.22	A	1.49	15.70
16.09	B	5.18	12.32
16.23	A	3.16	22.18
18.61	B	18.40	6.61
16.40	B	3.59	7.89

Exemplo



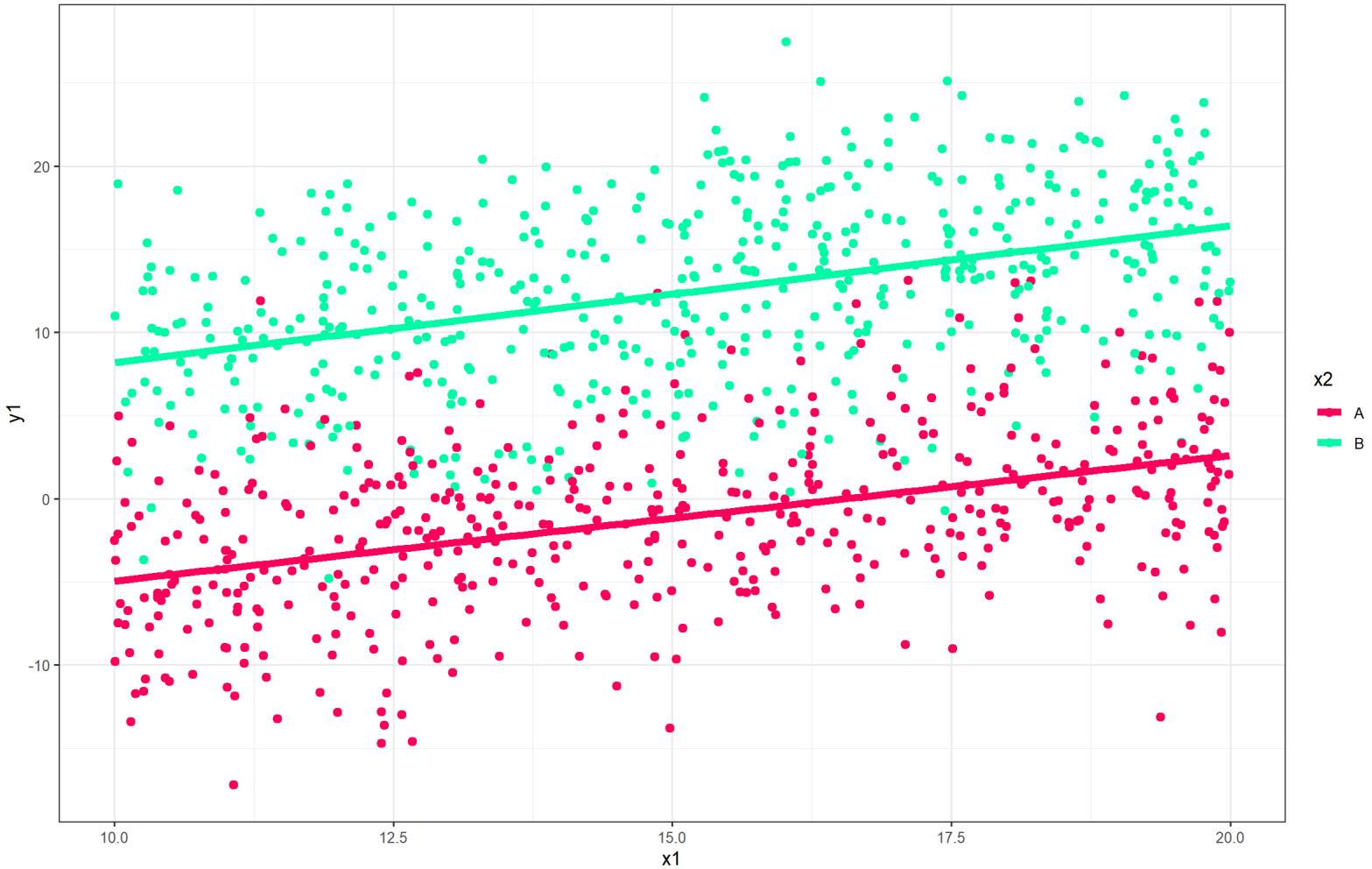
Exemplo

$$Y = \beta_0 + \beta_1 X_1$$



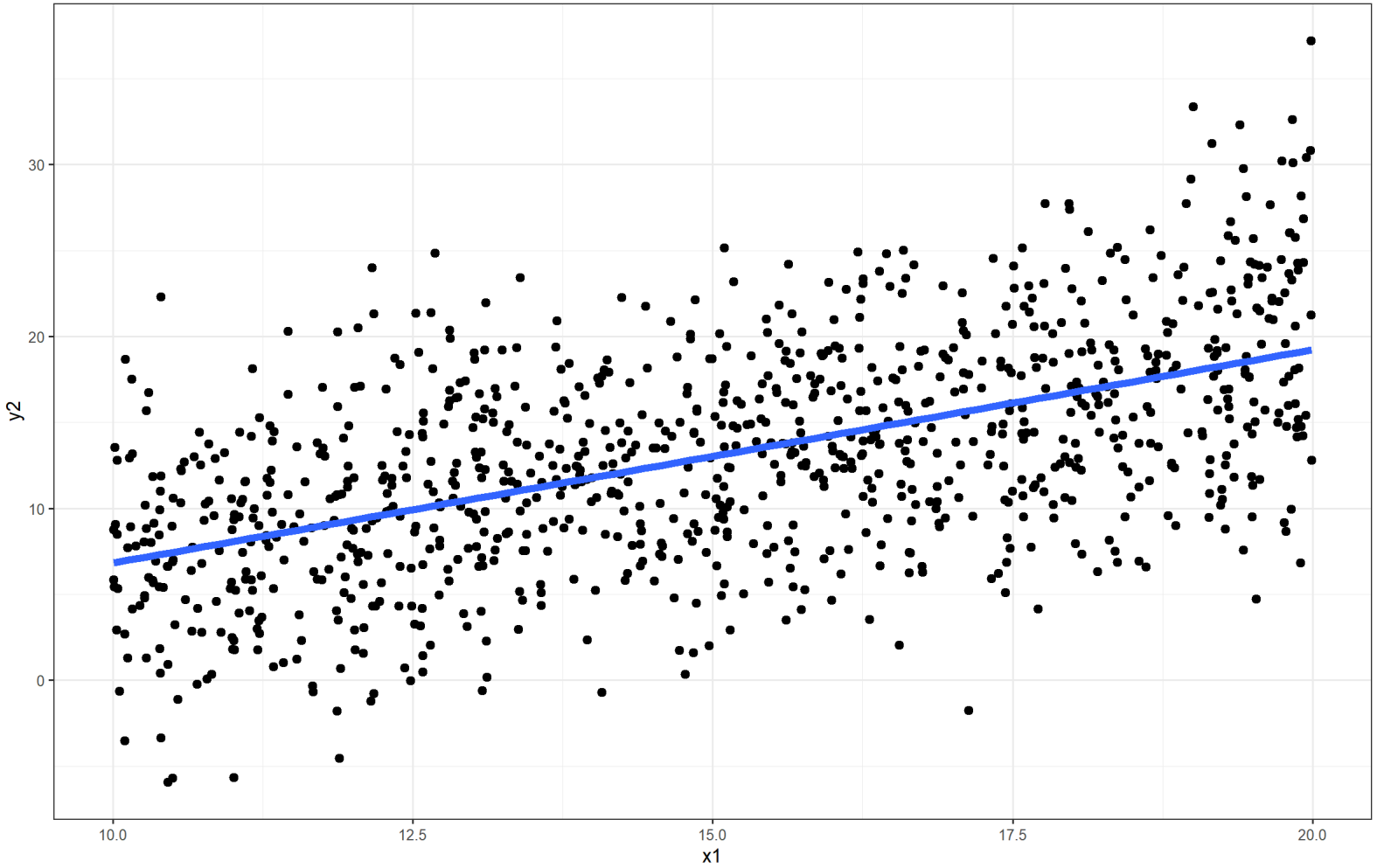
Exemplo

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$



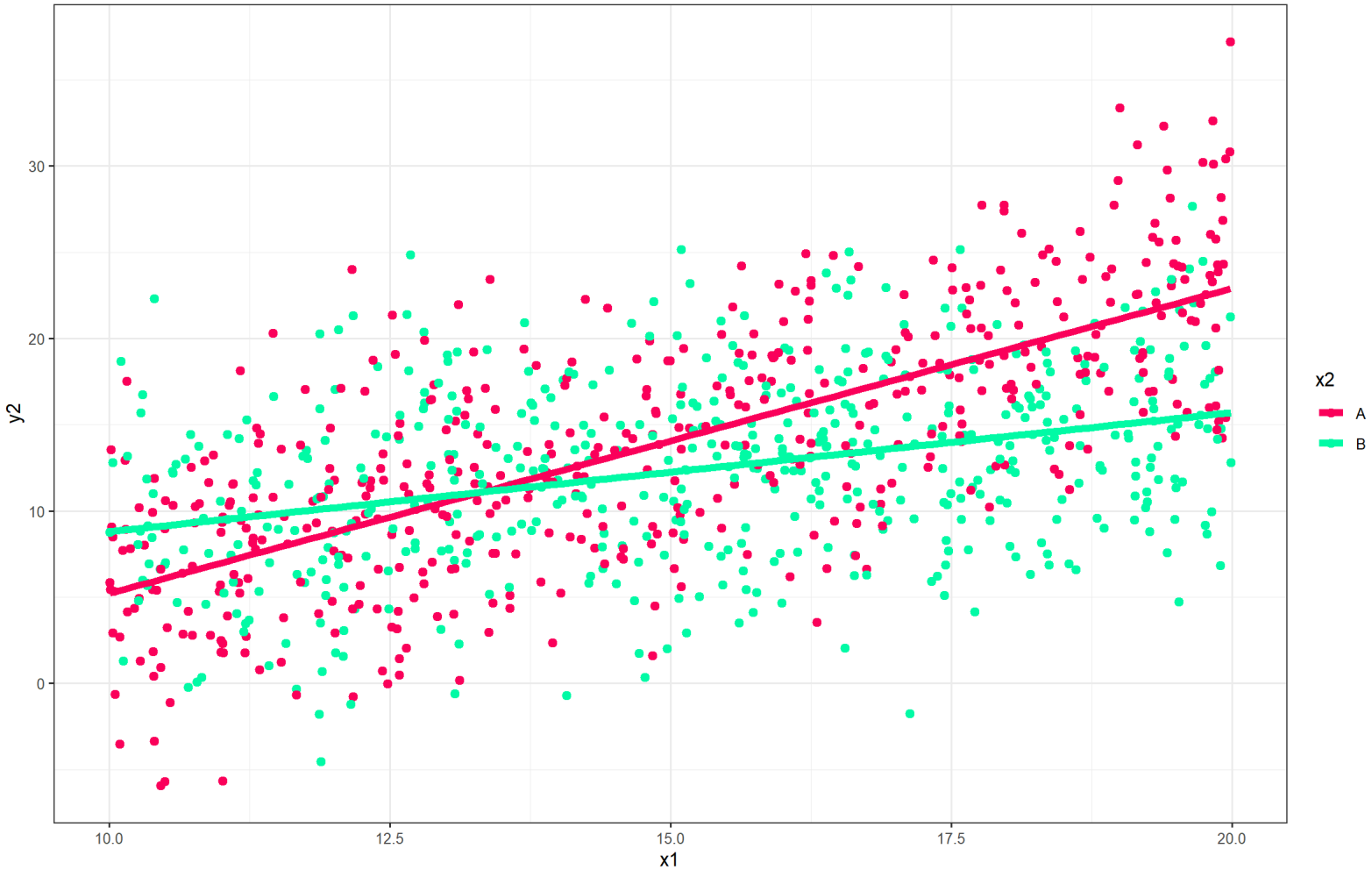
Exemplo

$$Y = \beta_0 + \beta_1 X_1$$



Exemplo

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$



Advertising

```
fit <- lm(sales ~ ., data = advertising)

summary(fit)
```

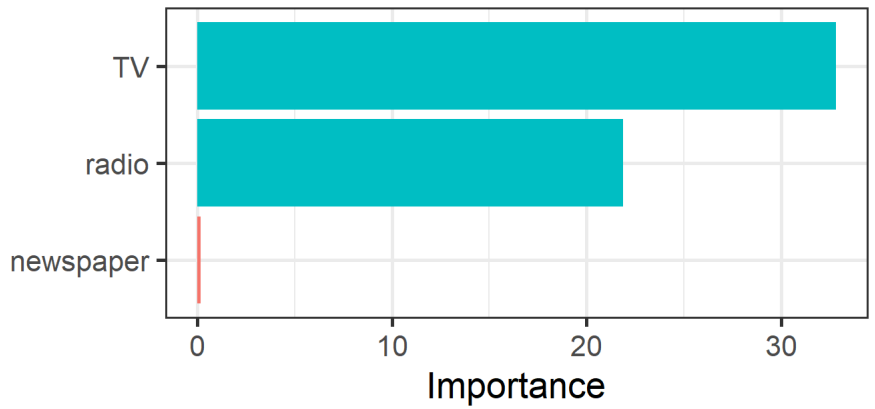
```
##
## Call:
## lm(formula = sales ~ ., data = advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## radio        0.188530   0.008611  21.893   <2e-16 ***
## newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Importância das variáveis

```
library(vip)

vi(fit)
```

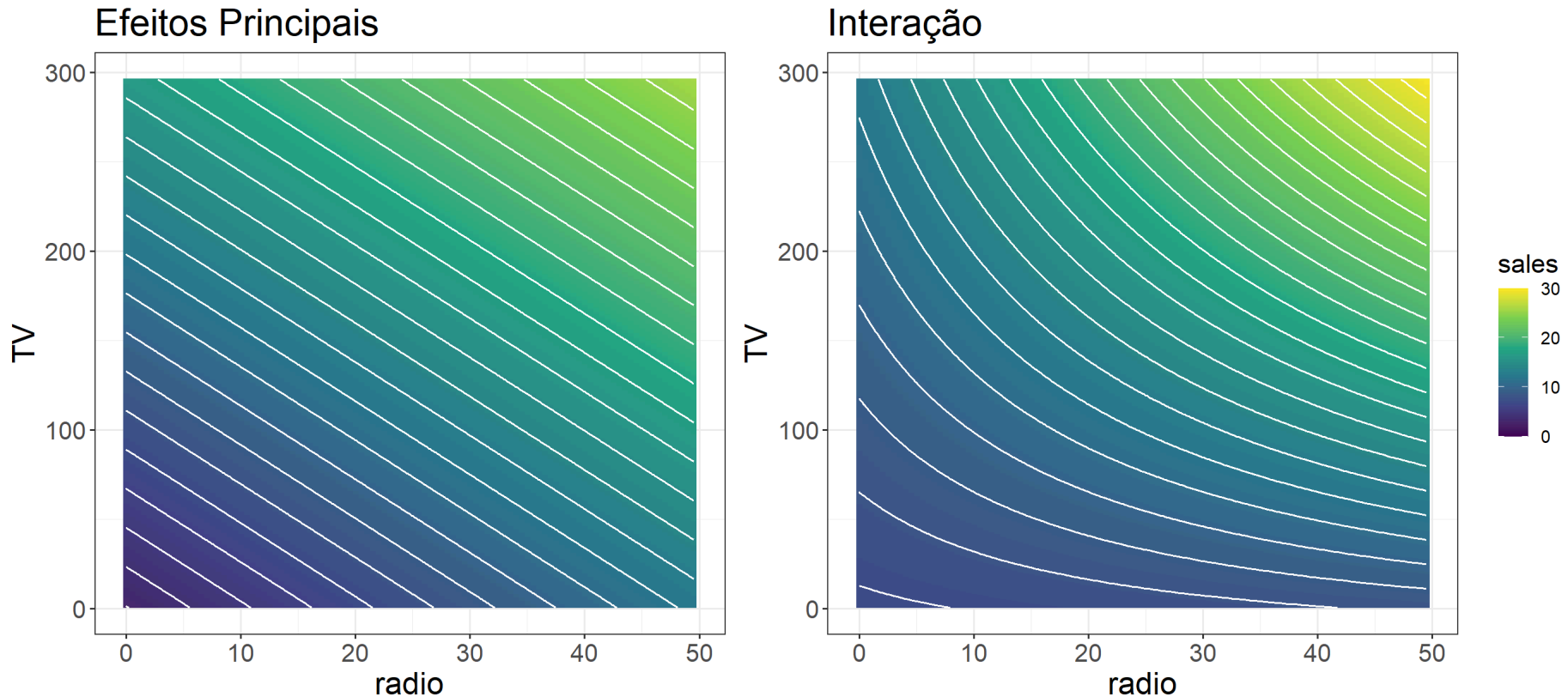
```
## # A tibble: 3 × 3
##   Variable  Importance Sign
##   <chr>          <dbl> <chr>
## 1 TV            32.8   POS
## 2 radio         21.9   POS
## 3 newspaper      0.177 NEG
```

```
vip(fit, mapping = aes(fill = Sign))
```

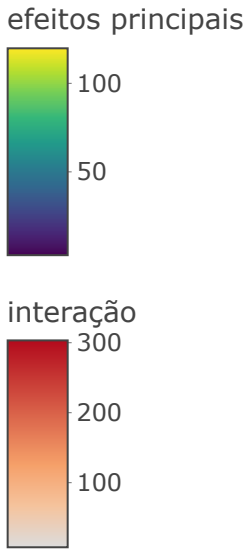


Interação

```
fit <- lm(sales ~ TV + radio, advertising)

fit_interacao <- lm(sales ~ TV*radio, advertising)
```



Interação



Seleção de Modelos

- **Seleção de subconjuntos**: considera um subconjunto das p preditoras.
- **Regularização**: ajusta-se um modelo com as p preditoras e os coeficientes estimados são encolhidos em direção a zero. Essa abordagem reduz a variância.
- **Redução de dimensão**: considera a utilização de uma combinação das p preditoras numa dimensão M tal que $M < p$.

Critérios

- $\mathrm{C}_p = \frac{1}{n}(\mathrm{RSS} + 2p\hat{\sigma}^2)$
- Akaike Information Criteria - $\mathrm{AIC} = \frac{1}{n\hat{\sigma}^2}(\mathrm{RSS} + 2p\hat{\sigma}^2)$
- Bayesian Information Criteria - $\mathrm{BIC} = \frac{1}{n}(\mathrm{RSS} + \log(n)p\hat{\sigma}^2)$

em que p é o número de preditoras utilizadas no modelo e $\hat{\sigma}^2$ é uma estimativa da variância do erro ϵ baseado em

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon.$$

Best Subset Selection[1]

- Seja $\mathcal{M}_0$ o modelo nulo, ou seja, o modelo sem preditoras. Esse modelo prevê com base na média amostral das observações.
- Para $k = 1, \ldots, p$:
  - Ajuste todos os $\binom{p}{k}$ modelos com k preditoras.
  - Selecione o melhor entre os $\binom{p}{k}$ modelos e denote por $\mathcal{M}_k$. O *melhor* modelo pode ser definido de acordo com RSS ou R^2.
- Selecione o melhor modelo para cada $\mathcal{M}_0, \ldots, \mathcal{M}_p$ utilizando validação cruzada para o erro de previsão, C_p, AIC, BIC ou R^2 ajustado.

[1] descrição apresentada no livro *An Introduction to Statistical Learning with Applications in R*.

Stepwise

Para contornar o problema do número de modelos relativo ao *best subset selection*, os métodos *stepwise* exploram um espaço restrito de modelos.

Número de variáveis	Best subset	Stepwise
2	4	4
4	16	11
8	256	37
16	65.536	137
32	4.294.967.296	529

Forward stepwise selection[1]

- Seja $\mathcal{M}_0$ o modelo nulo, ou seja, o modelo sem preditoras.
- Para $k = 0, \ldots, p-1$:
  - Considere todos os $p - k$ modelos que aumentam as preditoras em $\mathcal{M}_k$ com uma preditora.
  - Escolha o *melhor* modelo entre os $p - k$ modelos e denote por $\mathcal{M}_{k+1}$. O melhor pode ser definido como a menor RSS ou maior R^2.
- Selecione o melhor modelo entre $\mathcal{M}_0, \ldots, \mathcal{M}_p$ utilizando validação cruzada para o erro de previsão, C_p, AIC, BIC ou R^2 ajustado.

Esse método pode ser aplicado para os cenários de alta dimensão em que $n < p$. No entanto, para esses casos, é possível construir os modelos $\mathcal{M}_0, \ldots \mathcal{M}_{n-1}$. Isso porque o método dos mínimos quadrados não possui solução única para os casos em que $p \geq n$.

[1] descrição apresentada no livro *An Introduction to Statistical Learning with Applications in R*.

Backward stepwise selection[1]

- Seja $\mathcal{M}_p$ o modelo completo, ou seja, o modelo contendo as p preditoras.
- Para $k = p, p-1, \ldots, 1$:
  - Considere todos os k modelos que contenham todas as preditoras em $\mathcal{M}_k$ menos uma para um total de $k-1$ preditoras.
  - Escolha o *melhor* modelo entre k modelos e denote por $\mathcal{M}_{k-1}$. O melhor pode ser definido como a menor RSS ou maior R^2.
- Selecione o melhor modelo entre $\mathcal{M}_0, \ldots, \mathcal{M}_p$ utilizando validação cruzada para o erro de previsão, C_p, AIC, BIC ou R^2 ajustado.

[1] descrição apresentada no livro *An Introduction to Statistical Learning with Applications in R*.

Credit[1]

- **ID**: id
- **Income**: renda (em $10,000)
- **Limit**: limite de crédito
- **Rating**: rating de crédito
- **Cards**: número de cartões de crédito
- **Age**: idade em anos
- **Education**: anos de escolaridade
- **Gender**: Male / Female
- **Student**: Yes / No
- **Married**: Yes / No
- **Ethnicity** : African American / Asian / Caucasian
- **Balance**: saldo médio do cartão de crédito em $

[1] dados contidos no pacote *ISLR*.

Credit

Faça uma análise exploratória dos dados (EDA - *exploratory data analysis*). Quais variáveis você acredita que mais se relacionam com *Balance*?

```
library(ISLR)

data(Credit)
```

ID	Income	Limit	Rating	Cards	Age	Education	Gender	Student	Married	Ethnicity
29	186.634	13414	949	2	41	14	Female	No	Yes	African American
86	152.298	12066	828	4	41	12	Female	No	Yes	Asian
140	107.841	10384	728	3	87	7	Male	No	No	African American
192	124.29	9560	701	3	52	17	Female	Yes	No	Asian
294	140.672	11200	817	7	46	9	Male	No	Yes	African American
324	182.728	13913	982	4	98	17	Male	No	Yes	Caucasian

Credit

```
library(ISLR)

data(Credit)

fit <- lm(Balance ~ ., data = Credit[,-1])

summary(fit)$coefficients
```

```
##                      Estimate  Std. Error     t value      Pr(>|t|)
## (Intercept)       -479.2078706 35.77393717 -13.3954468  6.730600e-34
## Income              -7.8031018  0.23423191 -33.3135727 7.372312e-116
## Limit                0.1909067  0.03277862   5.8241238  1.205974e-08
## Rating               1.1365265  0.49089445   2.3152157  2.112213e-02
## Cards               17.7244836  4.34103295   4.0830106  5.401200e-05
## Age                 -0.6139088  0.29398941  -2.0882005  3.743127e-02
## Education           -1.0988553  1.59795129  -0.6876651  4.920746e-01
## GenderFemale       -10.6532477  9.91399990  -1.0745660  2.832368e-01
## StudentYes         425.7473595 16.72258016  25.4594300  8.854521e-85
## MarriedYes          -8.5339006 10.36287466  -0.8235071  4.107256e-01
## EthnicityAsian      16.8041792 14.11906302   1.1901767  2.347047e-01
## EthnicityCaucasian  10.1070252 12.20992331   0.8277714  4.083088e-01
```

Stepwise

Forward

```
library(MASS)

fit <- lm(Balance ~ 1, data = Credit[,-1])

stepAIC(fit, direction = "forward",
  scope = list(lower = ~ 1,
               upper = ~ Income + Limit + Rating + Cards + Age + Education +
                         Gender + Student + Married + Ethnicity))
```

Backward

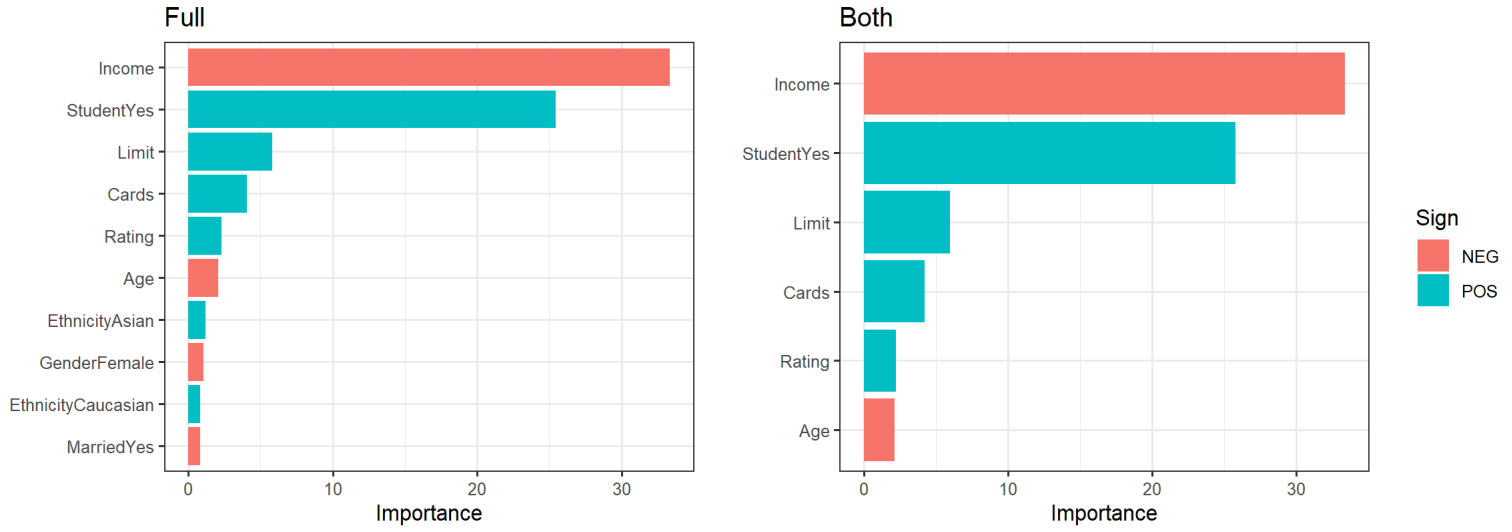
```
fit <- lm(Balance ~ ., data = Credit[,-1])

stepAIC(fit, direction = "backward")
```

Both

```
stepAIC(fit, direction = "both")
```

Variable Importance

```
library(patchwork)

fit1 <- lm(Balance ~ ., data = Credit[, -1])

fit2 <- lm(Balance ~ Income + Limit + Rating + Cards + Age + Student, data = Credit[, -1])

g1 <- vip(fit1, mapping = aes(fill = Sign)) + labs(title = "Full")

g2 <- vip(fit2, mapping = aes(fill = Sign)) + labs(title = "Both")

g1 + g2 + plot_layout(guides = "collect")
```



Shrinkage Methods

ou

Métodos de Encolhimento

Regressão Ridge

Antes o interesse era minimizar a seguinte quantidade:

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Agora consideramos uma penalização para os coeficientes (o que acontece se $\lambda = 0$? E se $\lambda \to \infty$?)

$$\text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

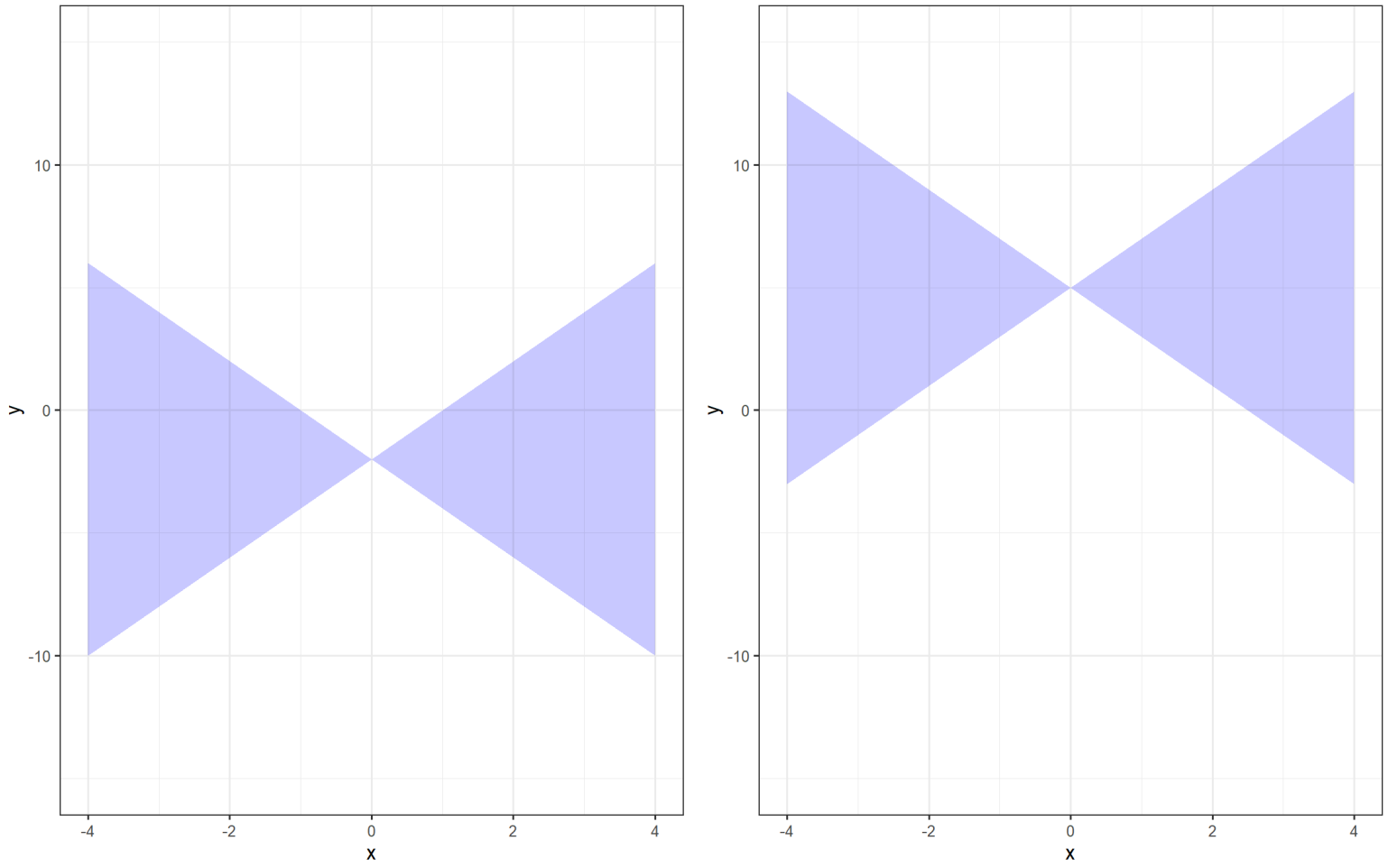
É possível mostrar que minimizar a quantidade acima é equivalente a

$$\arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \text{ sujeito a } \sum_{j=1}^{p} \beta_j^2 \leq s$$

obs: note que β_0 não é regularizado.

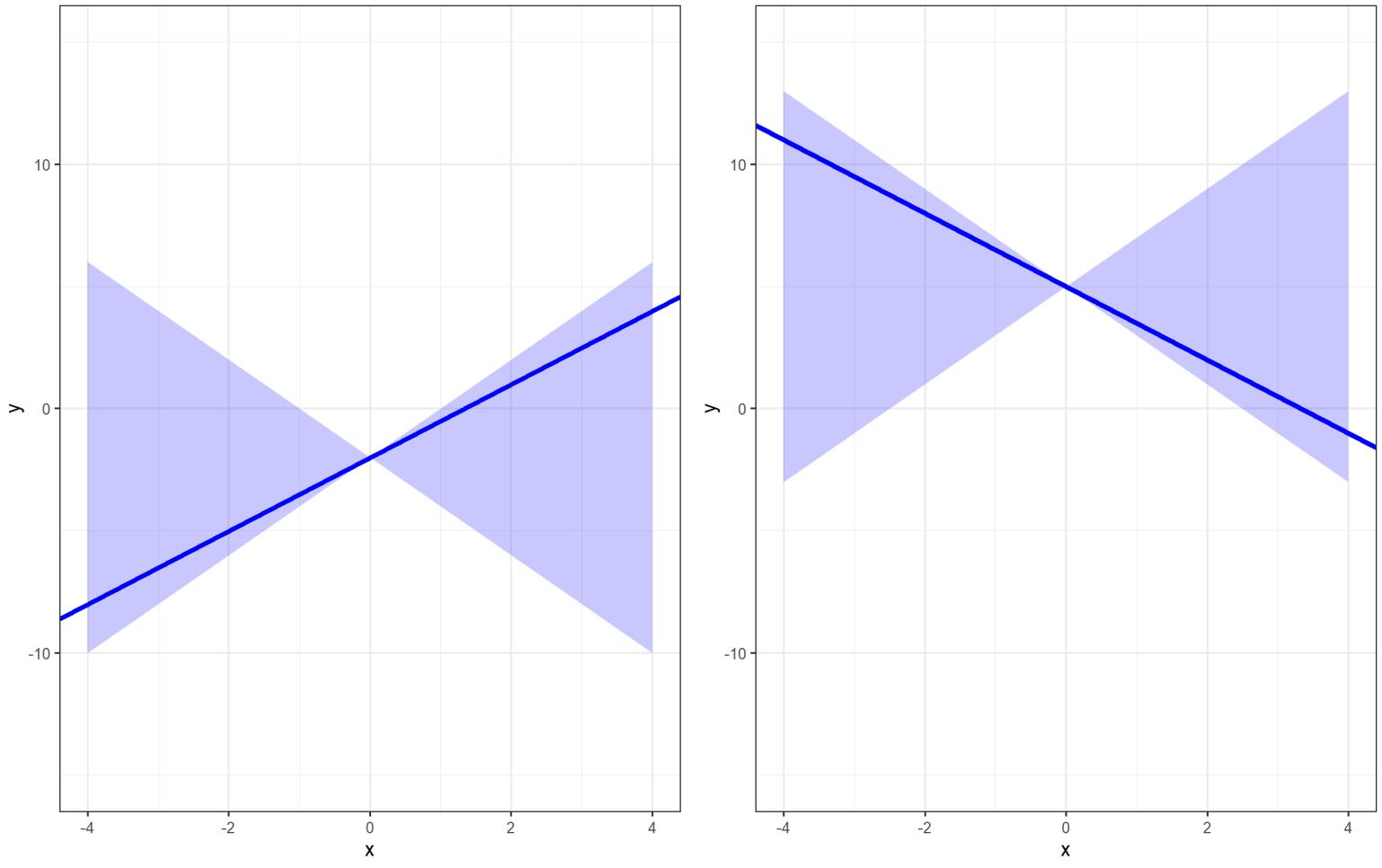
Regressão Ridge

Vamos pensar num caso simples em que $Y = \beta_0 + \beta_1 X_1$. Assim, $|\beta_1| \leq s$. Nesse caso, consideraremos $-2 \leq \beta_1 \leq 2$.



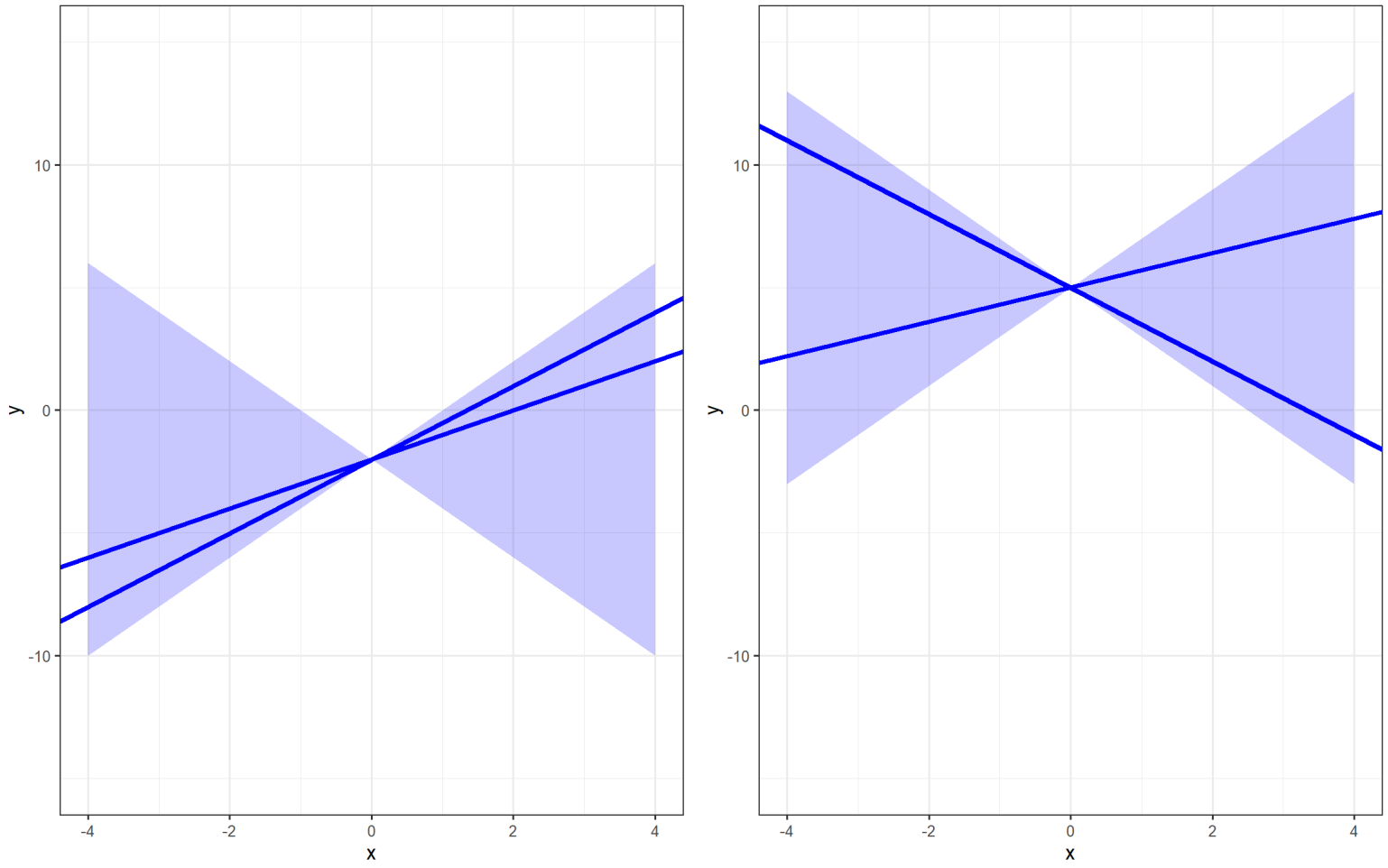
Regressão Ridge

Vamos pensar num caso simples em que $Y = \beta_0 + \beta_1 X_1$. Assim, $|\beta_1| \leq s$. Nesse caso, consideraremos $-2 \leq \beta_1 \leq 2$.



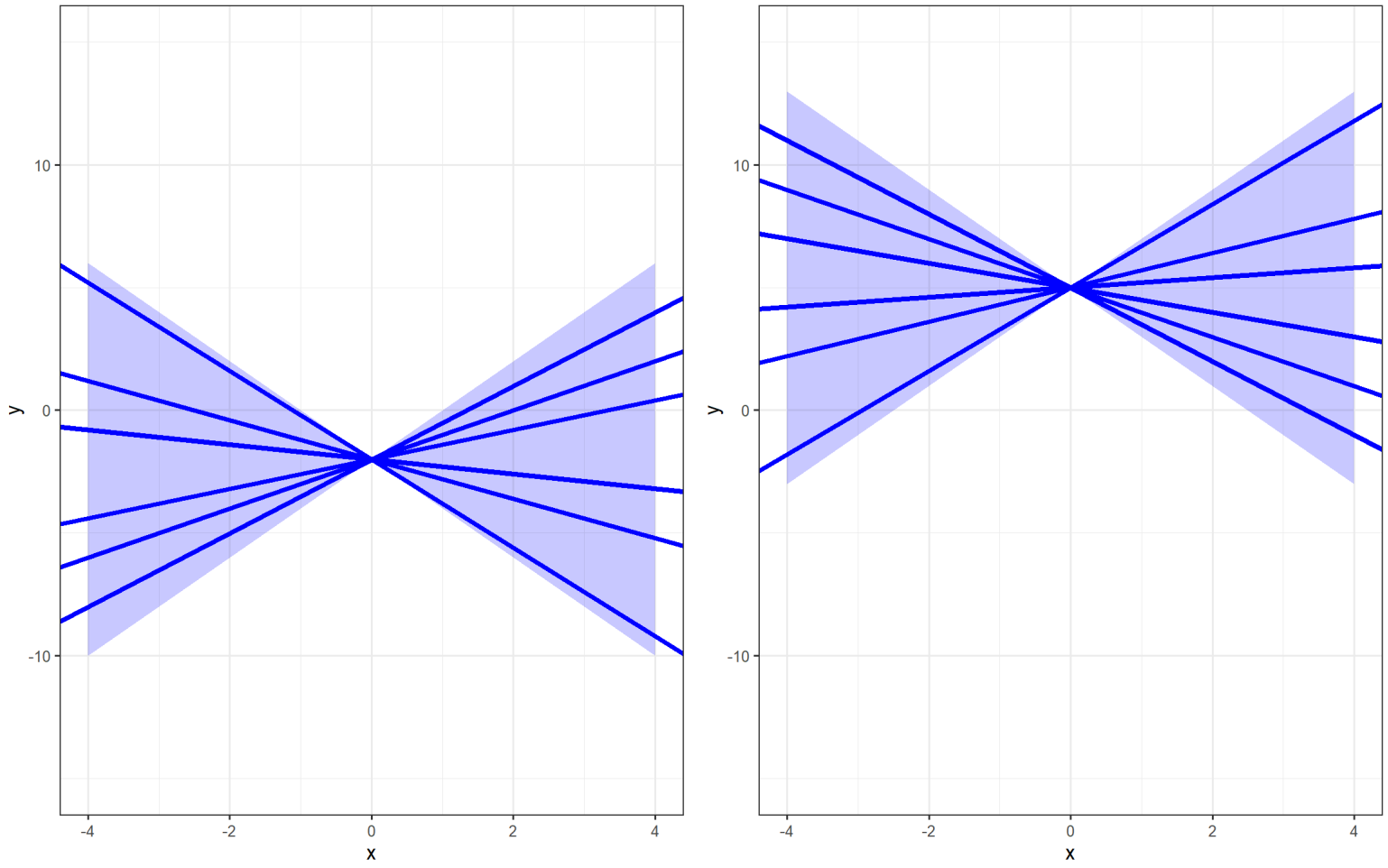
Regressão Ridge

Vamos pensar num caso simples em que $Y = \beta_0 + \beta_1 X_1$. Assim, $|\beta_1| \leq s$. Nesse caso, consideraremos $-2 \leq \beta_1 \leq 2$.



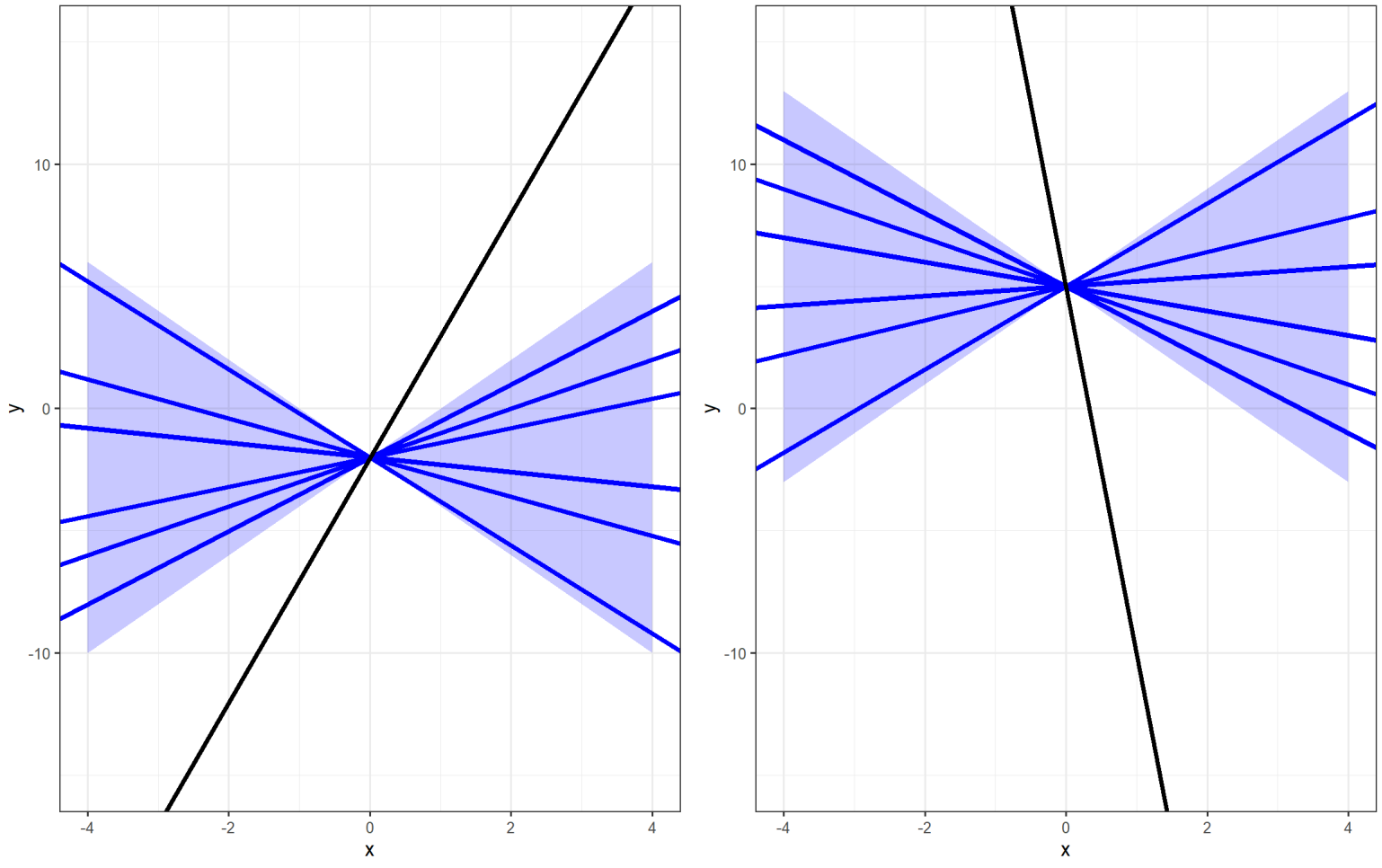
Regressão Ridge

Vamos pensar num caso simples em que $Y = \beta_0 + \beta_1 X_1$. Assim, $|\beta_1| \leq s$. Nesse caso, consideraremos $-2 \leq \beta_1 \leq 2$.



Regressão Ridge

Vamos pensar num caso simples em que $Y = \beta_0 + \beta_1 X_1$. Assim, $|\beta_1| \leq s$. Nesse caso, consideraremos $-2 \leq \beta_1 \leq 2$.

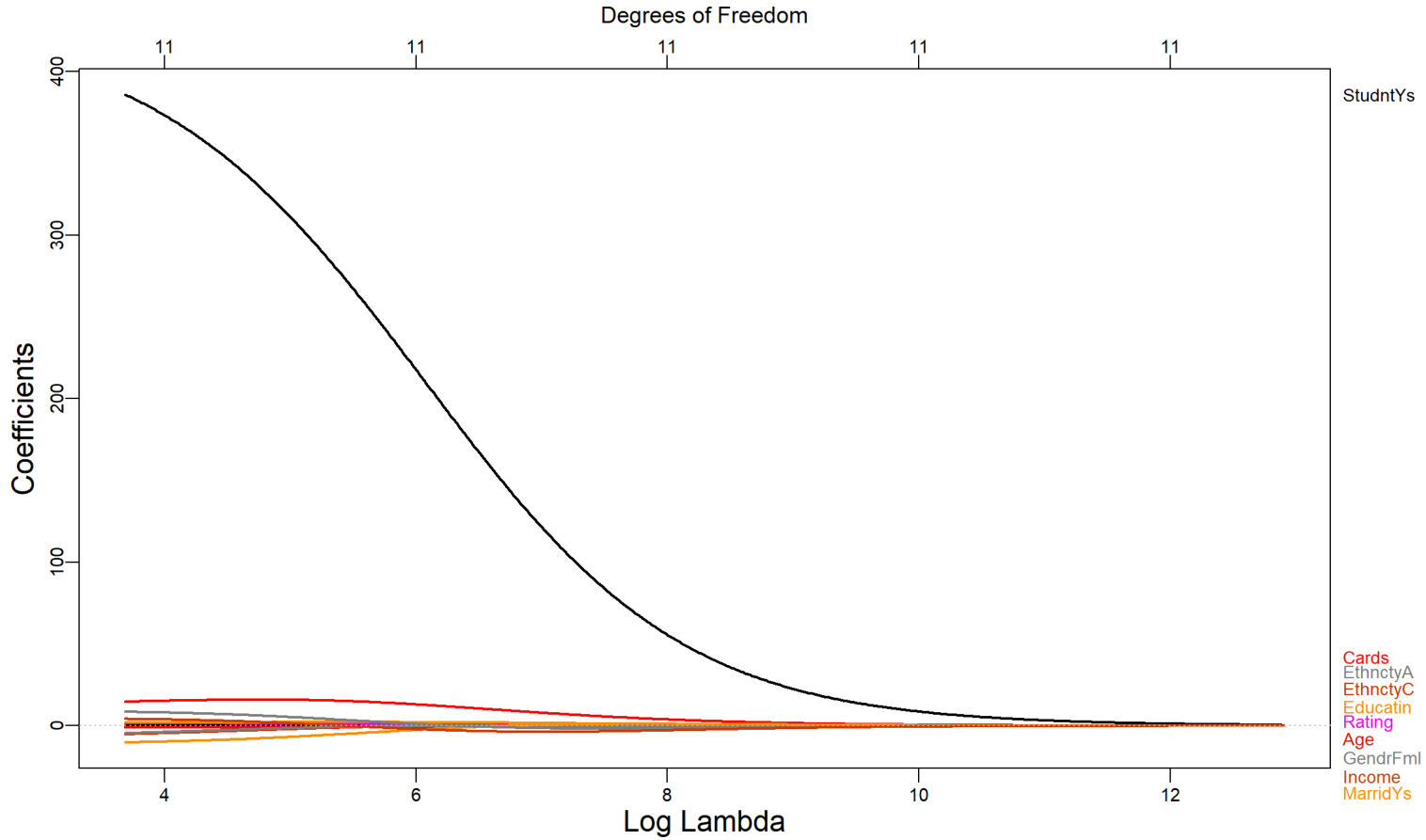


Regressão Ridge

Vamos utilizar a regressão ridge para os dados *Credit*.

```
library(ISLR)    # base de dados
library(glmnet)  # LASSO, ridge e elasticnet
library(plotmo)  # gráficos

X <- model.matrix(Balance ~ ., data = Credit[,-1])[,-1]

y <- Credit$Balance

set.seed(12)

idx <- sample(nrow(Credit), size = .75 * nrow(Credit)) # indice treinamento

ridge <- glmnet(X[idx,], y[idx], alpha = 0, nlambda = 500)

plot_glmnet(ridge, lwd = 2, cex.lab = 1.3)

ridge$a0

ridge$beta

ridge$lambda
```

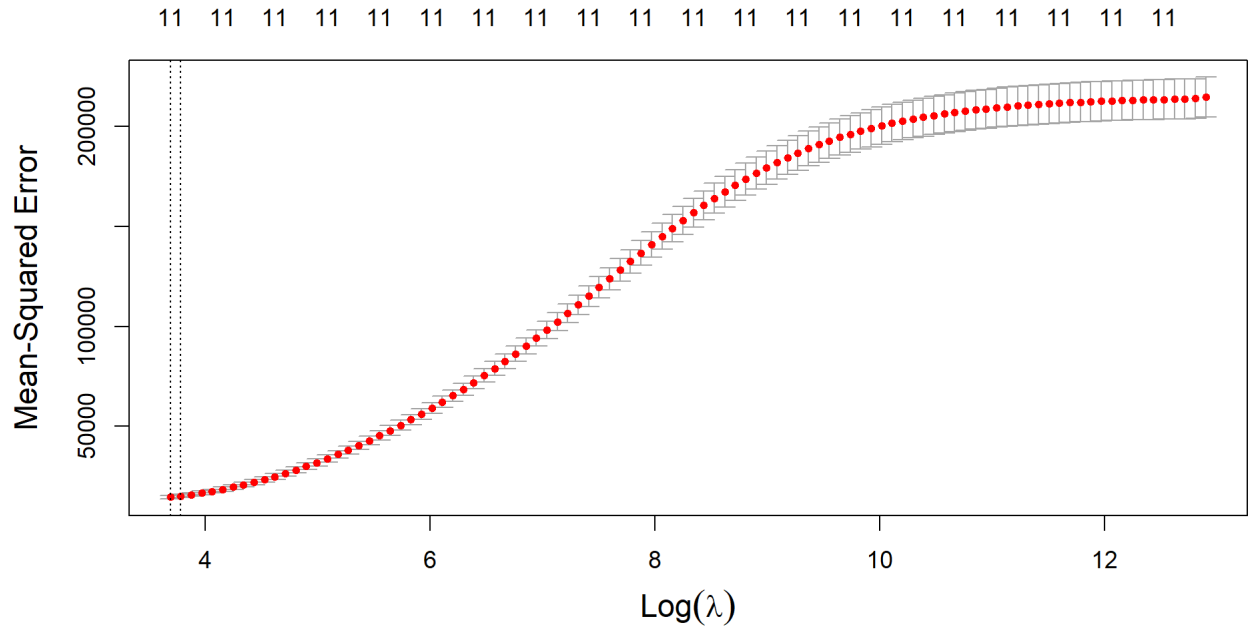
Regressão Ridge



Regressão Ridge

Utilizaremos validação cruzada para determinar o valor ótimo de λ. Esse gráfico apresenta a estimativa do erro e o desvio-padrão. Essa função utiliza 10-folds como padrão. Os números na parte superior indicam quantos coeficientes são diferentes de zero.

```
cv_ridge <- cv.glmnet(X[idx,], y[idx], alpha = 0)

plot(cv_ridge, cex.lab = 1.3)
```



Regressão Ridge

```
y_ridge <- predict(ridge, newx = X[-idx,], s = cv_ridge$lambda.1se)

tab <- tibble(metodo = c("lm", "ridge", "lasso", "elastic"),
              mse = NA)

tab$mse[tab$metodo == "ridge"] <- mean((y[-idx] - y_ridge)^2)

# modelo linear sem regularização

fit_lm <- lm(Balance ~ ., Credit[idx, -1])

y_lm <- predict(fit_lm, Credit[-idx,])

tab$mse[tab$metodo == "lm"] <- mean((y[-idx] - y_lm)^2)

tab
```

```
## # A tibble: 4 × 2
##   metodo     mse
##   <chr>    <dbl>
## 1 lm      11205.
## 2 ridge   14606.
## 3 lasso      NA
## 4 elastic    NA
```

Regressão LASSO

LASSO (Least Absolute Shrinkage and Selection Operator)

Agora consideraremos uma penalização para os coeficientes (o que acontece se $\lambda = 0$? E se $\lambda \to \infty$?)

$$\text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$

É possível mostrar que minimizar a quantidade acima é equivalente a

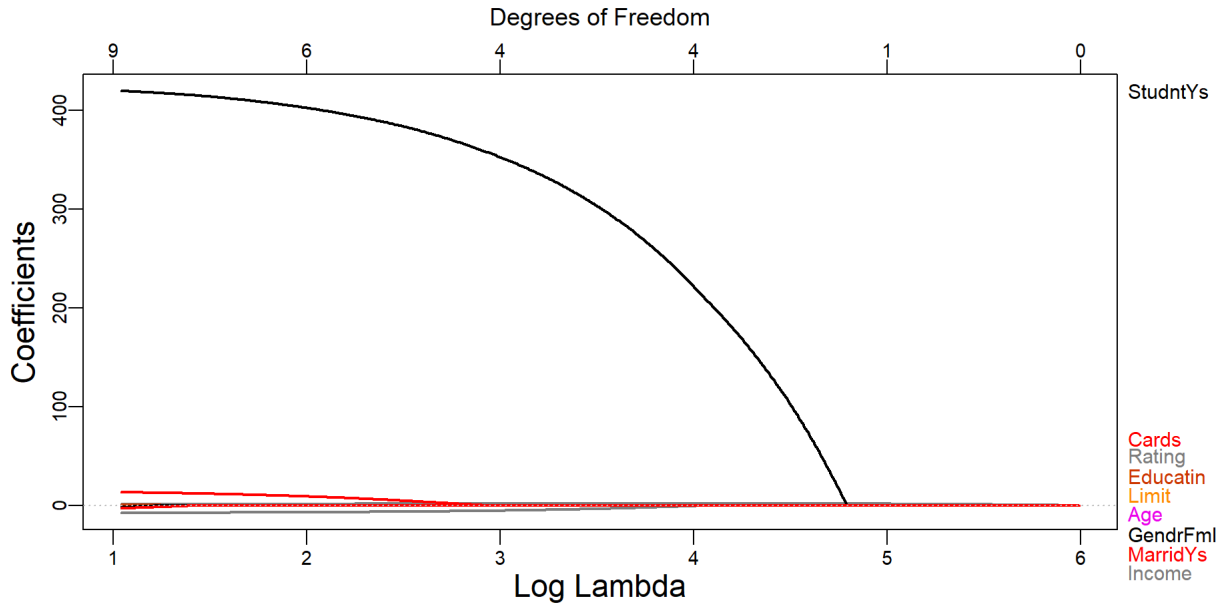
$$\arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \text{ sujeito a } \sum_{j=1}^{p} |\beta_j| \leq s$$

É comum ser descrito como penalização l_1. A norma l_1 de um vetor β é dada por $||\beta||_1 = \sum |\beta_j|$. A norma l_2 é dada por $||\beta||_2 = \sqrt{\sum \beta_j^2}$.

obs: note que β_0 não é regularizado.

Regressão LASSO

```
lasso <- glmnet(X[idx,], y[idx], alpha = 1, nlambda = 1000)

plot_glmnet(lasso, lwd = 2, cex.lab = 1.3, xvar = "lambda")
```

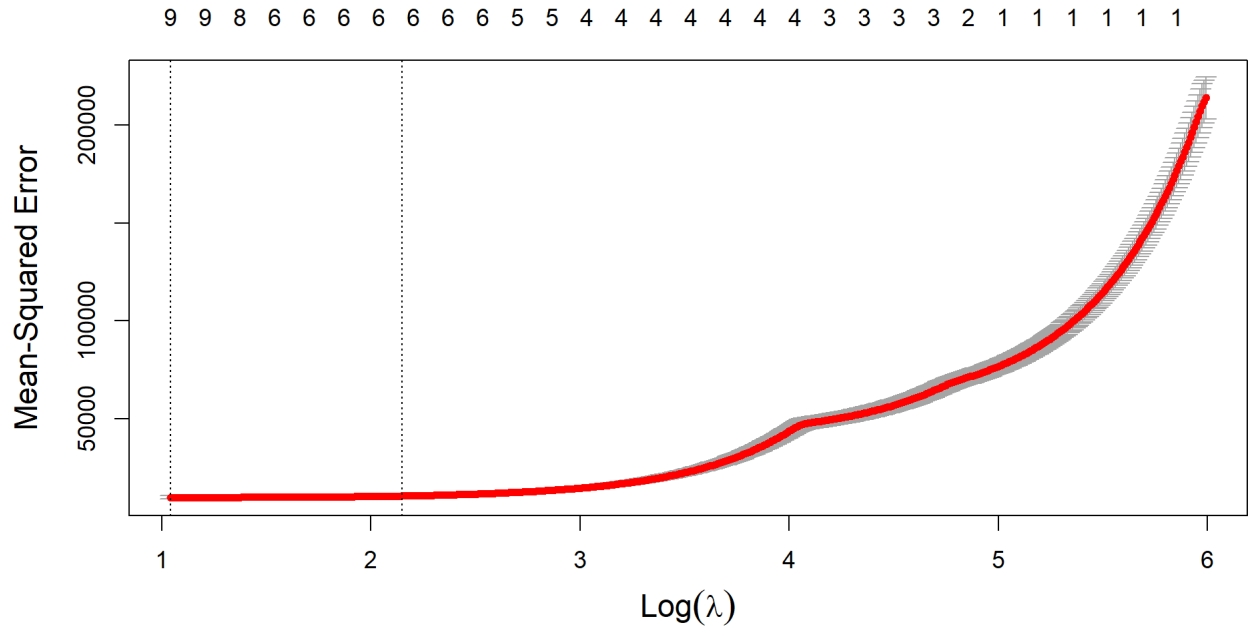



```
# lasso$beta[1:11, 58:67]
```

Regressão LASSO

Utilizaremos validação cruzada para determinar o valor ótimo de λ. Esse gráfico apresenta a estimativa do erro e o desvio-padrão.

```
cv_lasso <- cv.glmnet(X[idx,], y[idx], alpha = 1, lambda = lasso$lambda)

plot(cv_lasso, cex.lab = 1.3)
```

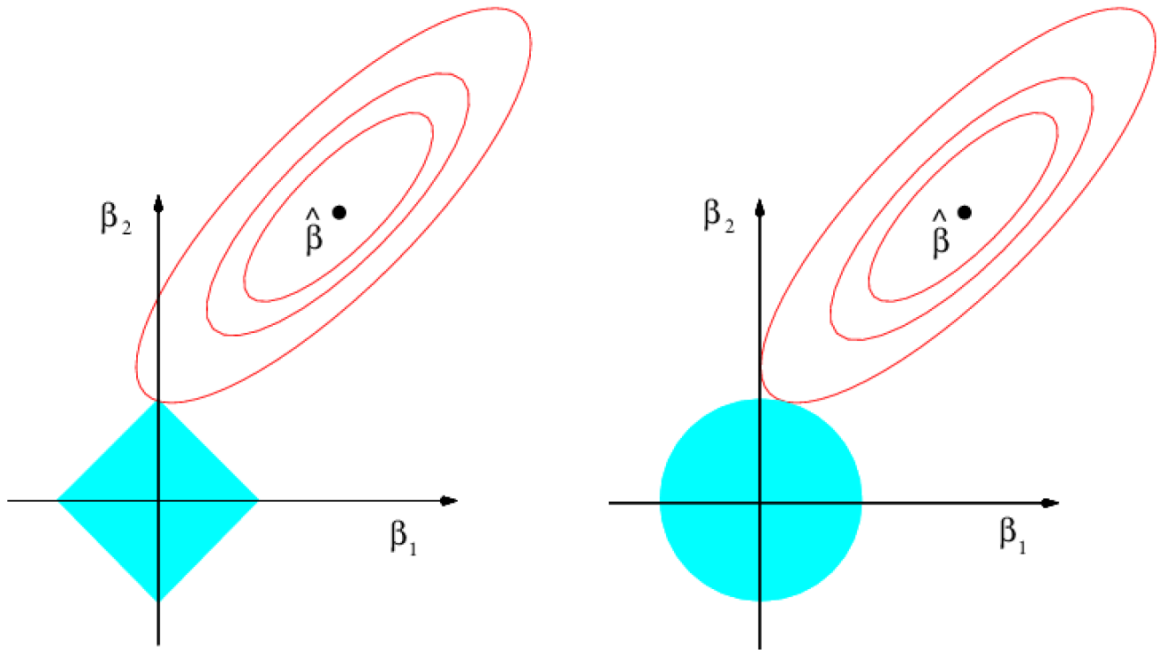


Regressão LASSO

```
y_lasso <- predict(lasso, newx = X[-idx,], s = cv_lasso$lambda.min)

tab$mse[tab$metodo == "lasso"] <- mean((y[-idx] - y_lasso)^2)

tab
```

```
## # A tibble: 4 × 2
##   metodo     mse
##   <chr>    <dbl>
## 1 lm      11205.
## 2 ridge   14606.
## 3 lasso   11181.
## 4 elastic    NA
```

LASSO e Ridge[1]



[1]Figura retirada do livro *An Introduction to Statistical Learning with Applications in R*.

Elastic-net

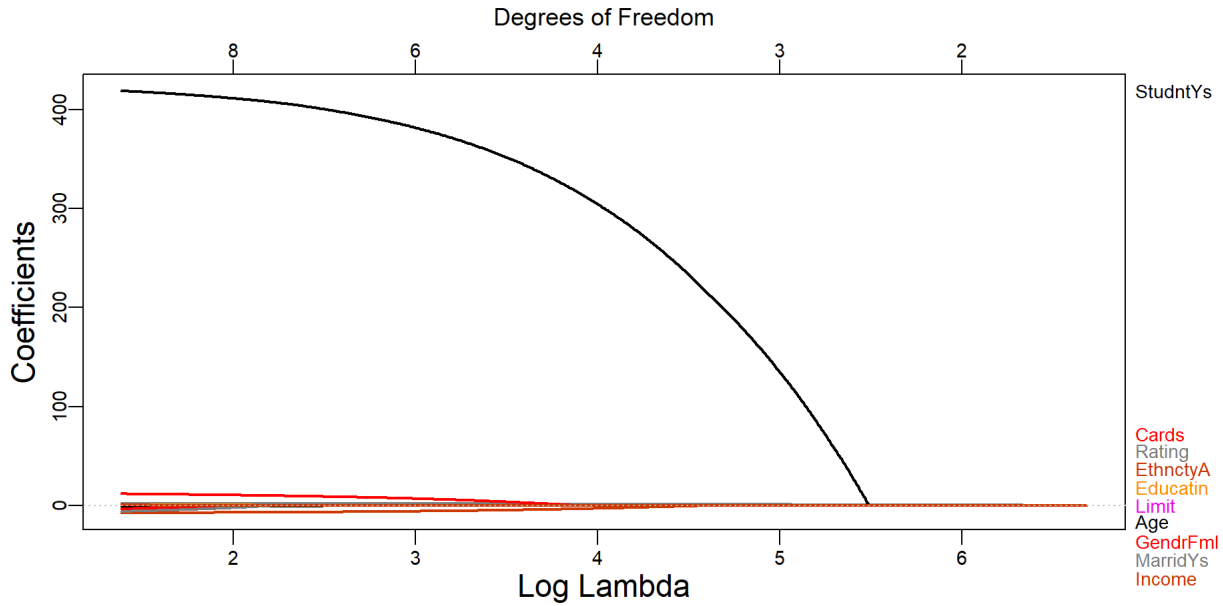
Agora consideraremos uma penalização para os coeficientes (o que acontece se $\lambda = 0$? E se $\lambda \to \infty$?)

$$\text{RSS} + \lambda \sum_{j=1}^{p} \left(\alpha |\beta_j| + \frac{(1-\alpha)}{2} \beta_j^2 \right)$$

Para mais detalhes sobre o pacote **`glmnet`**, acesse a vignettes.

Elastic-net

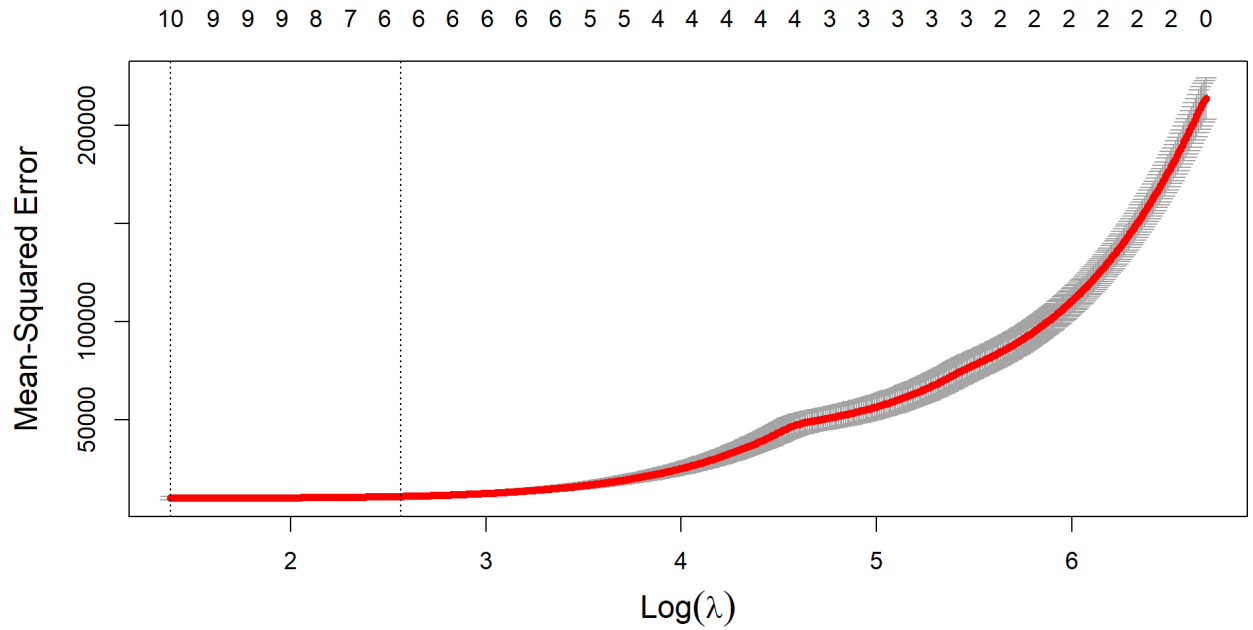
```
elastic <- glmnet(X[idx,], y[idx], alpha = 0.5, nlambda = 1000)

plot_glmnet(elastic, lwd = 2, cex.lab = 1.3, xvar = "lambda")
```



Elastic-net

Utilizaremos validação cruzada para determinar o valor ótimo de λ. Esse gráfico apresenta a estimativa do erro e o desvio-padrão.

```
cv_elastic <- cv.glmnet(X[idx,], y[idx], alpha = 0.5, lambda = elastic$lambda)

plot(cv_elastic, cex.lab = 1.3)
```



Elastic-net

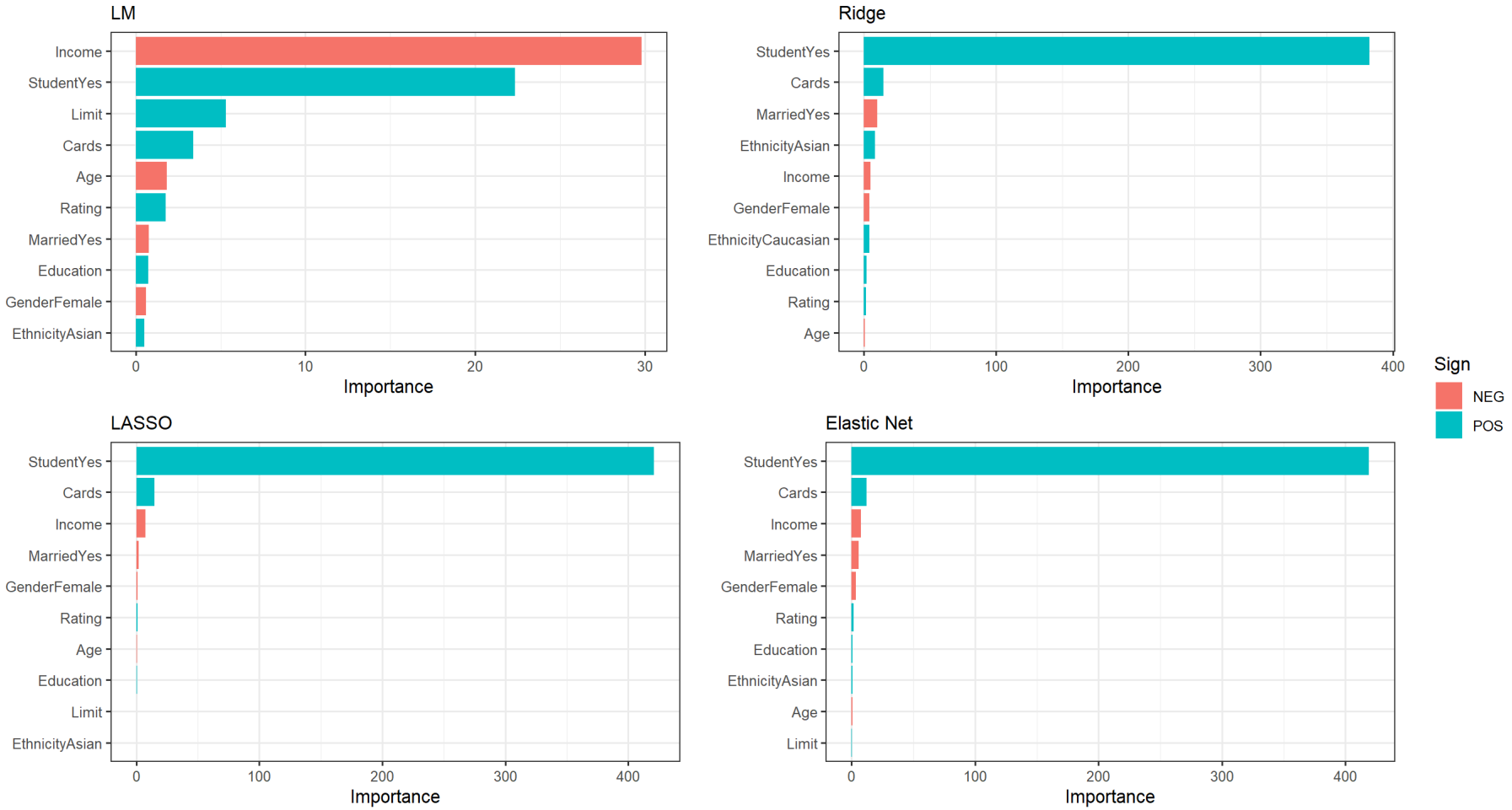
```
y_elastic <- predict(elastic, newx = X[-idx,], s = cv_elastic$lambda.min)

tab$mse[tab$metodo == "elastic"] <- mean((y[-idx] - y_elastic)^2)

tab %>%
  arrange(mse)
```

```
## # A tibble: 4 × 2
##   metodo     mse
##   <chr>    <dbl>
## 1 lasso   11181.
## 2 elastic 11201.
## 3 lm      11205.
## 4 ridge   14606.
```

Variable Importance

```
fit_lm <- lm(Balance ~ ., Credit[idx, -1])

fit_ridge <- glmnet(X[idx,], y[idx], alpha = 0, lambda = cv_ridge$lambda.1se)

fit_lasso <- glmnet(X[idx,], y[idx], alpha = 1, lambda = cv_lasso$lambda.min)

fit_elastic <- glmnet(X[idx,], y[idx], alpha = 0.5, lambda = cv_elastic$lambda.min)

g1 <- vip(fit_lm, mapping = aes(fill = Sign)) +
  labs(subtitle = "LM")

g2 <- vip(fit_ridge, mapping = aes(fill = Sign)) +
  labs(subtitle = "Ridge")

g3 <- vip(fit_lasso, mapping = aes(fill = Sign)) +
  labs(subtitle = "LASSO")

g4 <- vip(fit_elastic, mapping = aes(fill = Sign)) +
  labs(subtitle = "Elastic Net")

(g1 + g2) / (g3 + g4) + plot_layout(guides = "collect")
```

Variable Importance



Conteúdo Extra

Generalized Additive Models - GAMs

Uma forma de generalizar o modelo de regressão linear múltipla

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

para permitir relações não lineares entre cada preditor e a resposta, se dá com a substituição de cada componente linear $\beta_j x_{ij}$ por uma função não linear $f_j(x_{ij})$.

Assim, teríamos

$$\begin{aligned} y_i &= \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i \\ &= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i. \end{aligned}$$

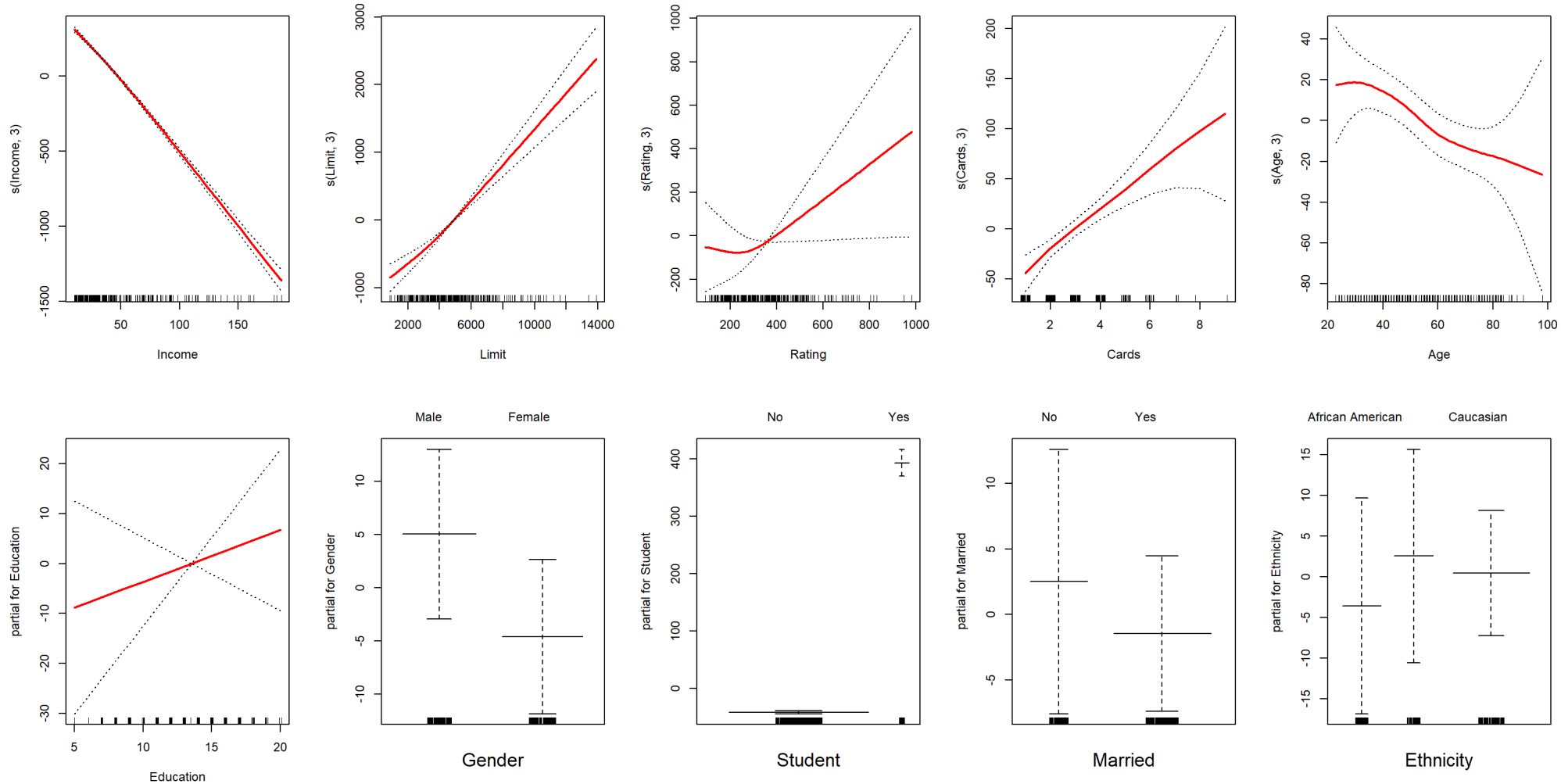
Para modelos de classificação, podemos utilizar

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)$$

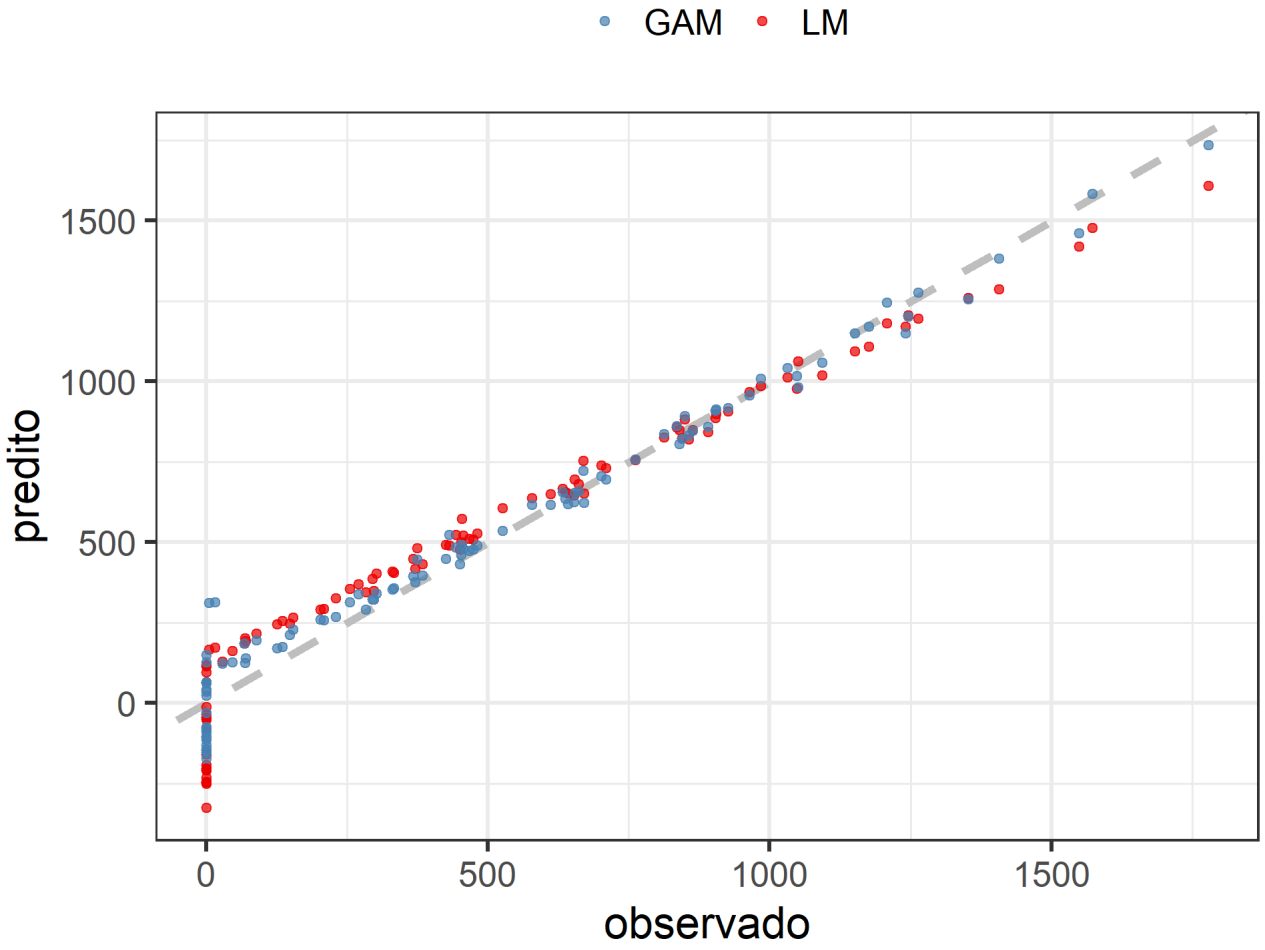
Generalized Additive Models - GAMs

```
library(gam)

set.seed(12)

idx <- sample(nrow(Credit), size = .75 * nrow(Credit)) # indice treinamento

fit <- gam(Balance ~ s(Income, 3) + s(Limit, 3) + s(Rating, 3) + s(Cards, 3) + s(Age, 3) +
                     Education + Gender + Student + Married + Ethnicity, data = Credit[idx,-1])

par(mfrow = c(2, 5))

plot(fit, se = TRUE, col = "red", lwd = 2)
```

Generalized Additive Models - GAMs



Generalized Additive Models - GAMs



Obrigado!

tiagoms.com

tiagomendonca

tiagoms1@insper.edu.br