Modelos Lineares

Caio Graco-Roza

3/3/2021

Vamos usar a base de dados `iris` . A base de dados contem o medições em centímetro de comprimento e largura da pétala e sépala de 50 flores de 3 espécies de Íris.

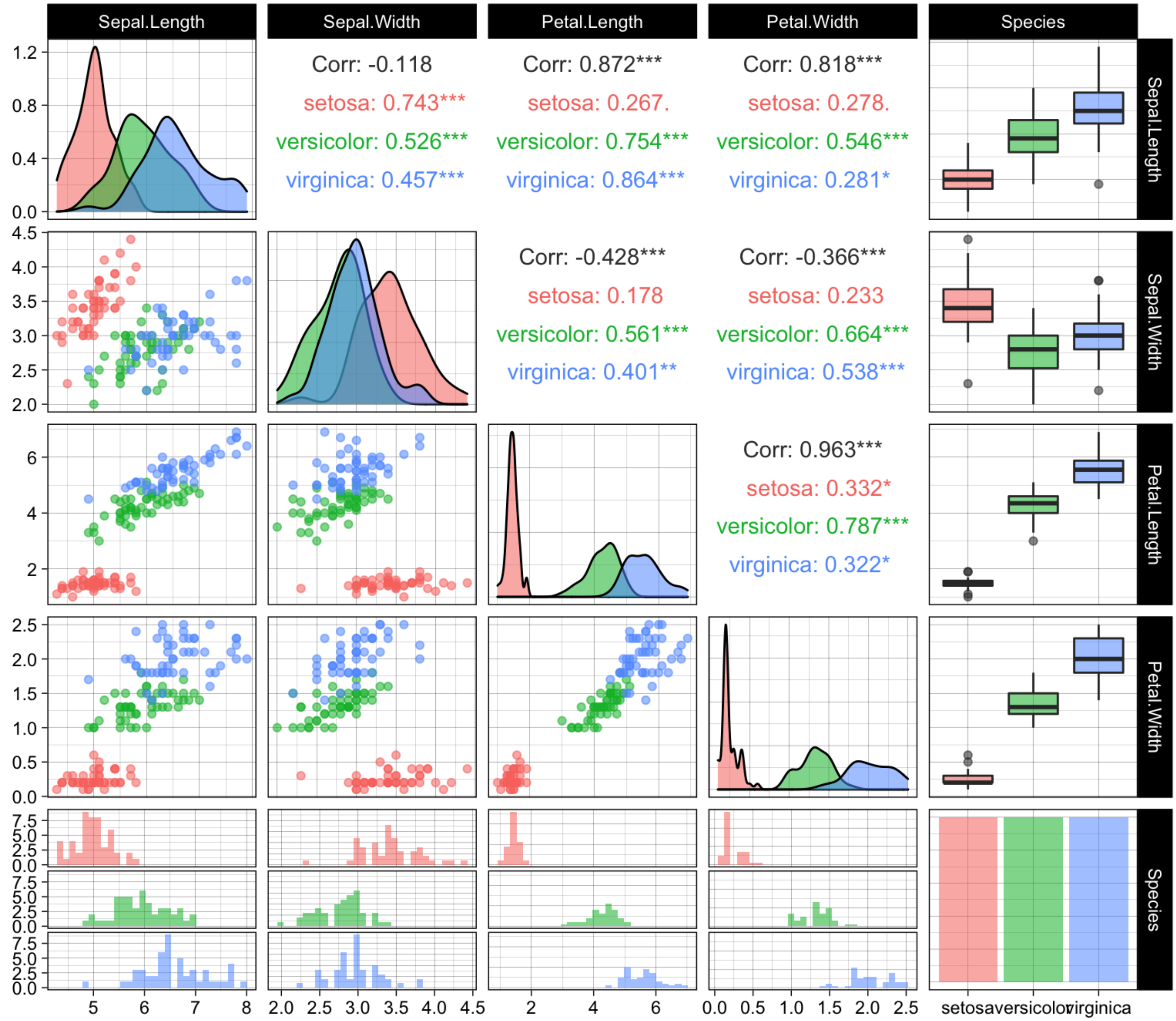
```
data(iris)

summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

Vamos dar uma olhada nos dados com o nosso método favorito. **Gráficos**

```
ggpairs(iris, aes(colour = Species, alpha = 0.4))
```



Modelo linear simples

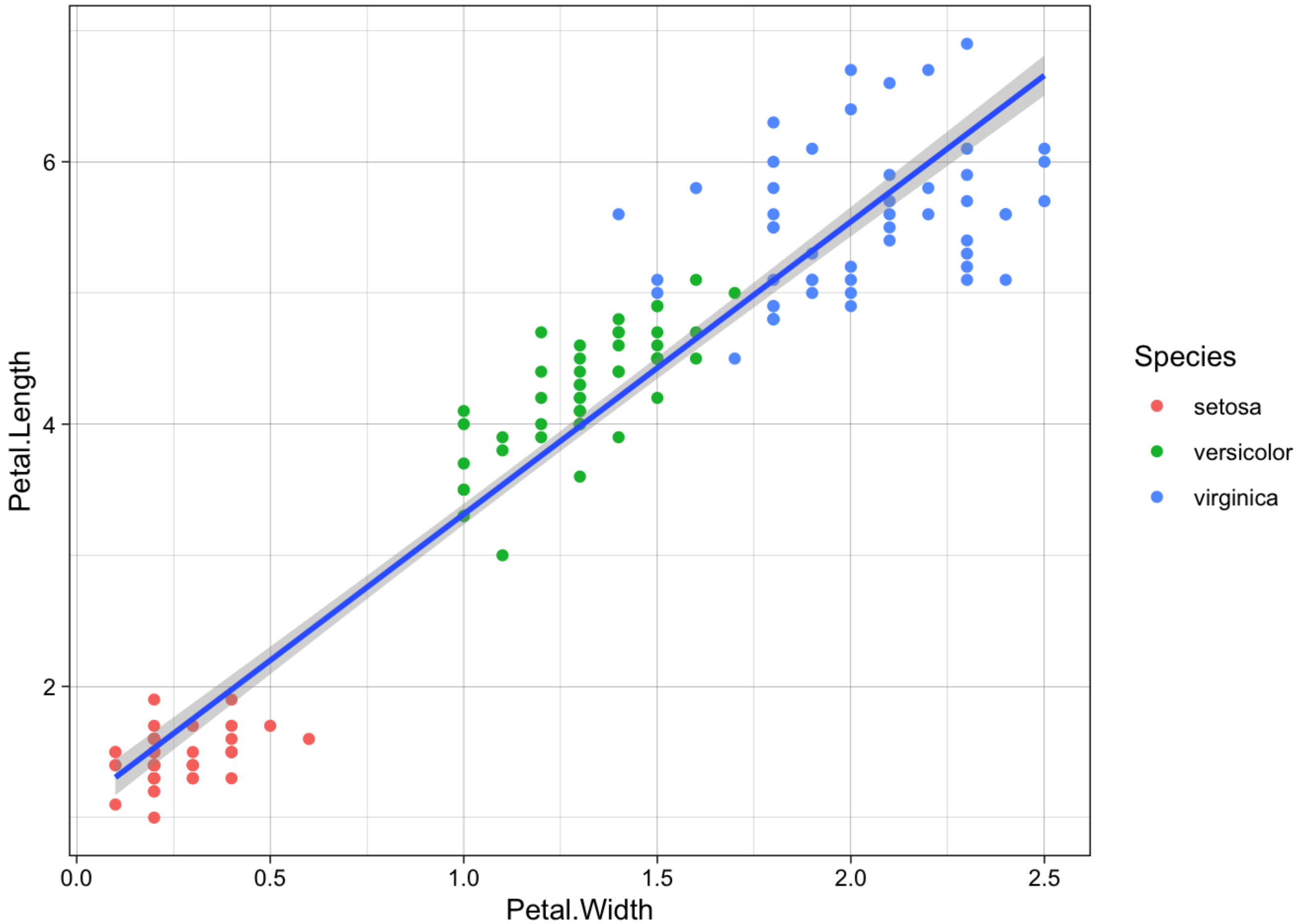
O modelo linear simples pode ser descrito através da equação:

$$Y = \beta_0 + \beta_1 X$$

Essa equação pode ser lida como "pegue o valor da variável X, multiplique por β_1 e some esse valor a β_0 o resultado é o valor da variável Y.

Numa linguagem mais simples podemos escrever a equação como $Y = intercepto + inclinação \times observação$.

Para estudarmos a nossa regressão linear vamos usar a relação entre tamanho e largura da pétala.



Modelos lineares são estimados baseados no método dos *minimos quadrados ordinarios*. Esse método é reconhecido por *minimizar a soma dos quadrados dos resíduos* (SQR). O método pode ser dividido em dois processos, o de estimar a inclinação da curva (β_1) e o de estimar o intercept (β_0).

Como estimar a inclinação da curva?

A inclinação da curva é simplesmente a covariância entre as variável resposta (Y) e a variável dependente (X) dividido pela variancia (s^2) de X.

$$inclinação = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}))}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Transformando isso em R

```
covariancia <- iris %>% summarise(cov = sum((Petal.Length - mean(Petal.Length)) *
    (Petal.Width - mean(Petal.Width))))

variancia <- iris %>% summarise(var = sum((Petal.Width - mean(Petal.Width))^2))

inclinação <- covariancia$cov/variancia$var

inclinação
```

```
## [1] 2.22994
```

```
# Maneira alternativa como diferença média de x e diferença média de y
diff(range(iris$Petal.Length))/diff(range(iris$Petal.Width))
```

```
## [1] 2.458333
```

```
# Maneira alternativa de escrita no R. Usando quadrados minimos.
cov(iris$Petal.Length, iris$Petal.Width)/var(iris$Petal.Width)
```

```
## [1] 2.22994
```

Como estimar o intercepto?

O intercepto é um parâmetro mais simples de enteder do que a inclinação. Ele é baseado na inclinação e nos valores médios da variável independente (Y) e dependente (X). Podemos descreve-lo como

$$Intercepto = \bar{y} - \beta_1 \times \bar{x}$$

O intercepto representa o nosso valor observado na ausência de efeito da variável X.

```
intercepto <- mean(iris$Petal.Length) - (cov(iris$Petal.Length, iris$Petal.Width)/var(ir
is$Petal.Width)) *
    mean(iris$Petal.Width)

intercepto
```

```
## [1] 1.083558
```

R²

Considerando que a nossa equação da reta é $\hat{y} = 1.08 + 2.22 \times x$ nós podemos tentar predizer valores e ver o quanto eles se distanciam da nossa expectativa.

```
# Aplico a minha formula para os valores observados de X e comparo o Y predito
# com o Y real.

predito <- sapply(iris$Petal.Width, function(x) intercepto + inclinação * x)
```

Assim o nosso R^2 pode ser descrito como a soma dos quadrados das diferenças entre o observado e o predito sobre a variância de y .

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

```
SQD_predito <- sum((iris$Petal.Length - predito)^2)
var_y <- sum((iris$Petal.Length - mean(iris$Petal.Length))^2)

R2 <- 1 - SQD_predito/var_y
```

Erro padrão

O erro padrão da regressão, ou erro padrão do parâmetro, representa a distância média entre os valores observados e a curva de regressão. Em outras palavras, o erro padrão simboliza o quão errado o nosso modelo está em unidades da variável resposta.

O erro padrão da curva pode ser calculado como:

$$ErroPadrão = \frac{\sqrt{\frac{\sum_{i=i}^{n}(y_i - \hat{y})^2}{n-2}}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

```
Erro.inc <- sqrt(sum((iris$Petal.Length - predito)^2)/(150 - 2))/sqrt(sum((iris$Petal.Wi
dth -
    mean(iris$Petal.Width))^2))

Erro.int <- sqrt(1/150) + mean(iris$Petal.Width)^2/sum((iris$Petal.Width - mean(iris$Pet
al.Width))^2)
```

Comparando resultados

```
# Nosso modelo
tribble(~"Parâmetro", ~"Estimado", ~"Erro Padrão", ~"Estatística t", ~"R²", "Intercepto"
,
    intercepto, Erro.int, intercepto/Erro.int, R2, "Inclinação", inclinação,
    Erro.inc, inclinação/Erro.inc, NA) %>% kbl(caption = "Resultados da Regressão linea
r",
    digits = 16) %>% kable_classic(full_width = F, html_font = "Cambria")
```

Resultados da Regressão linear

Parâmetro	Estimado	Erro Padrão	Estatística t	R^2
Intercepto	1.083558	0.09826513	11.02688	0.9271098
Inclinação	2.229940	0.05139623	43.38724	NA

```
summary(lm(Petal.Length ~ Petal.Width, data = iris))
```

```
##
## Call:
## lm(formula = Petal.Length ~ Petal.Width, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.33542 -0.30347 -0.02955  0.25776  1.39453
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.08356    0.07297   14.85   <2e-16 ***
## Petal.Width  2.22994    0.05140   43.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4782 on 148 degrees of freedom
## Multiple R-squared:  0.9271, Adjusted R-squared:  0.9266
## F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16
```