# Aula 8 - Introdução muito básica e rápida a análise estatística e modelos lineares

Vitor Rios

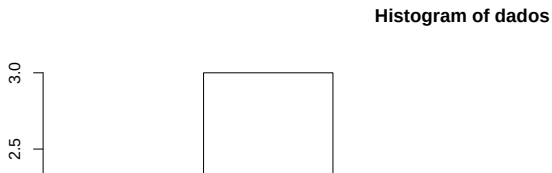
11 de novembro de 2017

# Lembrando o básico

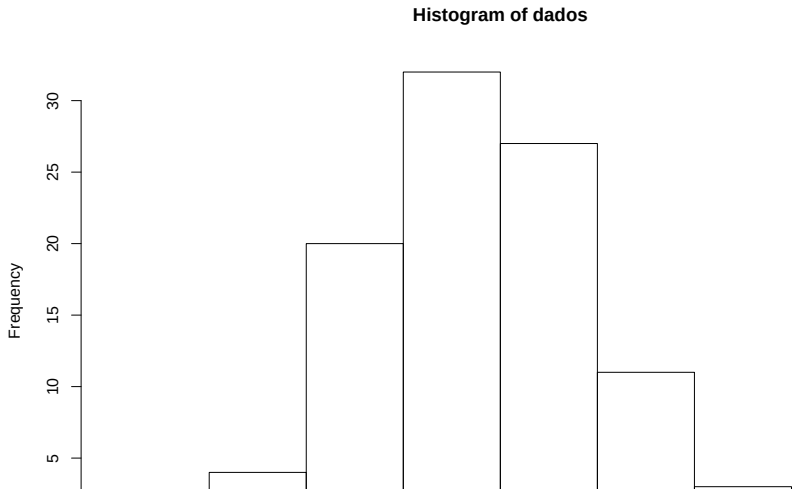
Ao coletarmos dados, eles tem uma determinada distribuição, isto é alguns valores podem ser mais frequentes que outros, podemos ter valores mais ou menos distantes da média, etc.

Podemos ver isso fazendo um gráfico da frequencia de cada valor dos dados. Supondo que medimos o tamanho de alguns bichos.

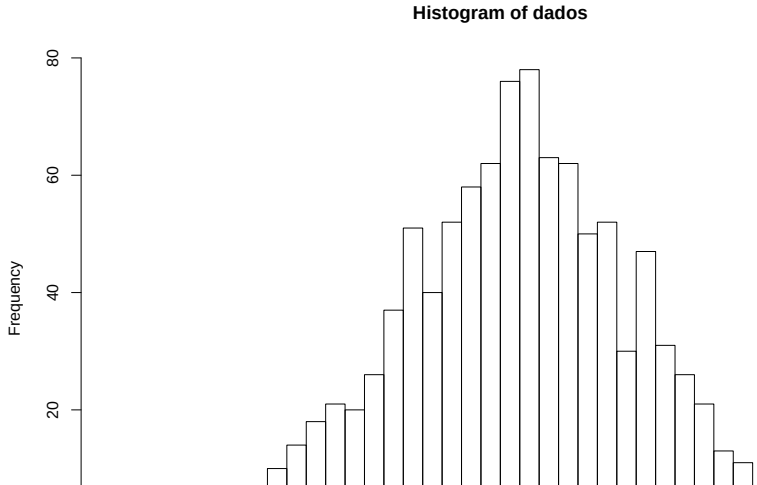
```
#primeiro, vamos gerar os dados usando rnorm(), sorteando 10 individuos de

dados = rnorm (10, mean = 1.6, sd = 0.3 )
#agora um histograma
hist(dados)
```



```r
#a amostra é muito pequena para entendermos o que está acontecendo,refazen
dados = rnorm (100, mean = 1.6, sd = 0.3 )
#agora um histograma
hist(dados)
```



```r
#e de novo
dados = rnorm (1000, mean = 1.6, sd = 0.3 )
#agora um histograma
hist(dados,breaks = 50)
```

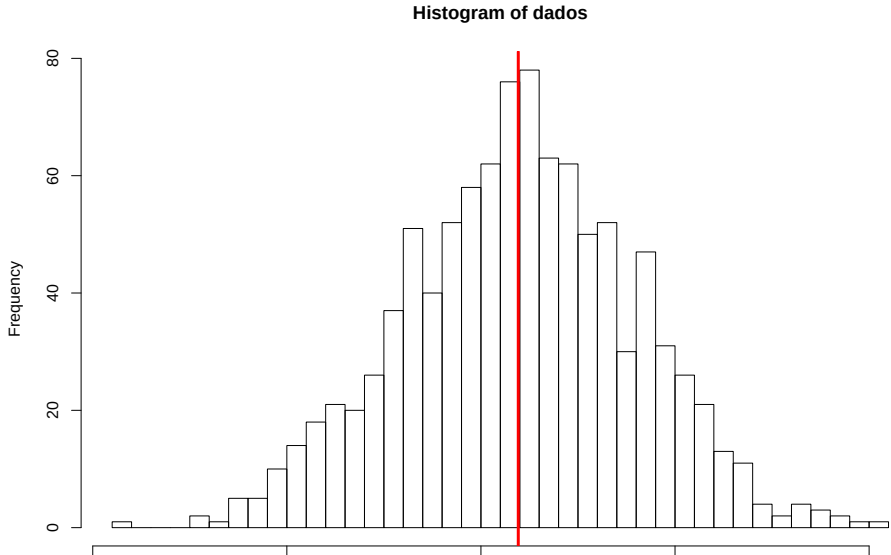


Acima, vimos uma característica de todo e qualquer conjunto de dados: quanto maior o `n` amostral, mais aprendemos sobre nossos dados

Além diso, percebemos que um determinado valor tem uma frequência maior que os outros, e que por coincidencia fica no meio da distribuição. Vemos também que a distribuição é aproximadamente simétrica em torno deste valor central.

Se somarmos todos os valores e dividirmos pelo `n`, teremos a `média`, que para uma distribuição gaussiana (também chamada de "normal") descreve o valor mais provável. Em outras palavras, se colocarmos todos os bichos que medimos num pote e pegarmos um ao acaso, provavelmente ele terá tamanho próximo da média

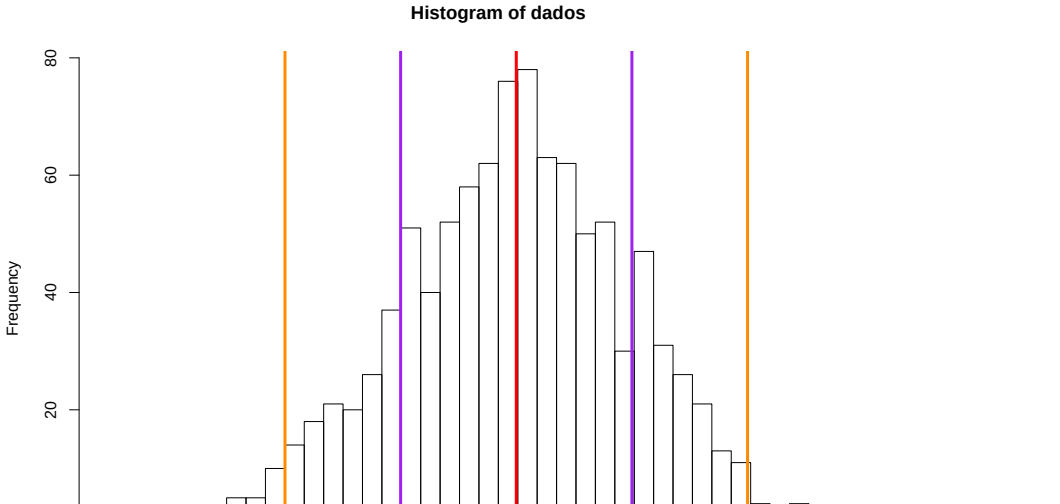
```r
hist(dados,breaks = 50)
abline(v=mean(dados),col="red",lwd=3)
```



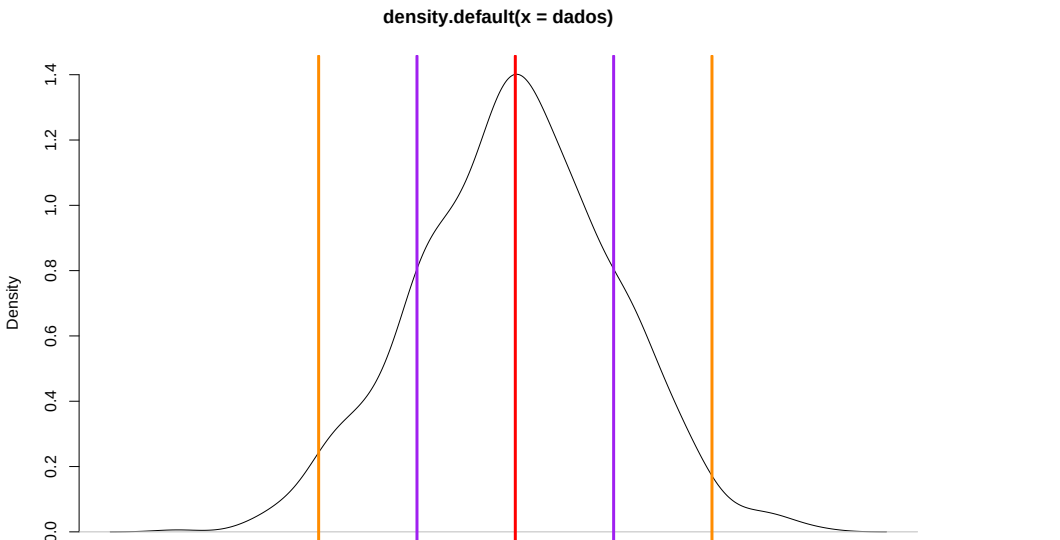
Podemos também calcular o quanto os dados estão distribuídos em torno da média, em outras palavras o quanto de nossa distribuição está mais ou menos perto da média. Para isso usamos o `desvio padrão` (*standard deviation*, `sd`).

Numa distribuição com desvio padrão alto, a probabilidade de encontrar um bicho muito maior ou menor que a média é grande, enquanto que se o desvio padrão for baixo, temos o contrário, a maioria dos bichos estará próximo à média. O desvio padrão não é um valor dentro da distribuição, mas sim um descrição dela.

Podemos ter uma noção melhor escolhendo dois pontos na nossa distribuição: um igual à média mais o desvio padrão, e um igual à média menos o desvio padrão, e destacando eles. Note que a maior parte dos dados (68%) fica nesse intervalo. Numa curva normal, 95% dos dados ficam no intervalo `média $+-$ 2*sd`

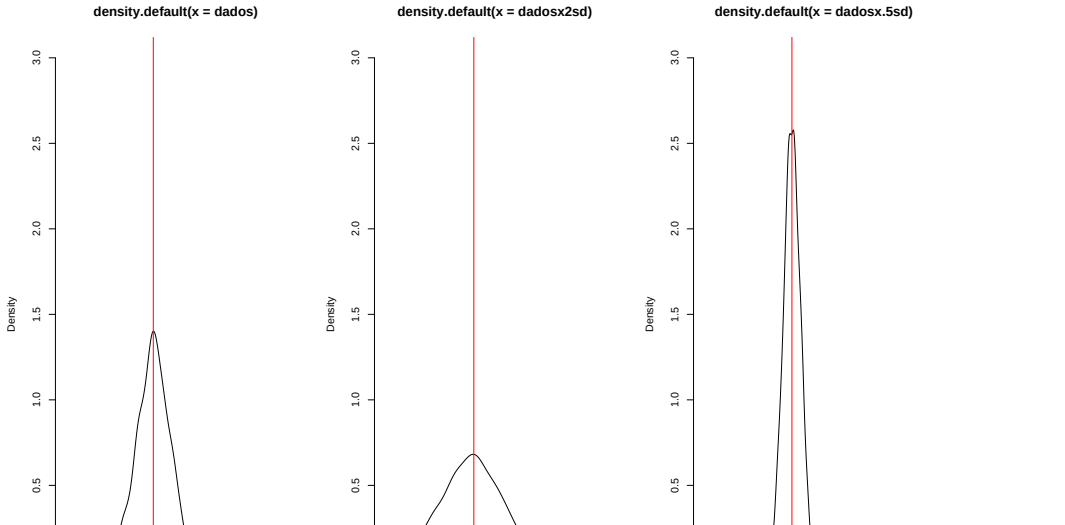


Geralmente, quando temos muitos dados, representamos a distribuição na forma de uma curva, que representa a probabilidade dos valores (tecnicamente, a densidade probabilística), ao invés de suas frequências



# Desvio padrão maior ou menor

Vamos comparar 3 distribuições, a nossa original, uma com 2x o desvio padrão, e uma com 0.5x o desvio padrão, todas com a mesma média



# Variância

Também podemos descrever a abertura da curva normal usando a `variância`, que é igual ao quadrado do desvio padrão (tecnicamente, o desvio é a raiz da variância)

O conceito intuitivo de variância é o mesmo do desvio padrão padrão: mantendo a mesma média, uma variância maior significa que podemos esperar mais valores longe da média, e uma variância menor significa que podemos esperar mais valores perto da média

# Pra quê isso serve?

Uma vez que descrevemos nossos dados com média e variância, podemos usar isso para fazer inferências, isto é prever resultados futuros e entender a releção entre mais de uma variável.

## Previsão

Conhecendo a média e a variância de uma ditribuição podemos calcular a probabilidade de obter um determinado valor ao acaso, ou valores a partir de um dado valor limite

Lembre que podeos calcular quantos % da curva estão entre média `média $+-$ sd`, e que 50% da curva estão para cada lado da média. Da mesma forma, podemos escolher um ponto ao acaso e calcular a probabilidade de obter aquele valor, ou um valor menor que ele

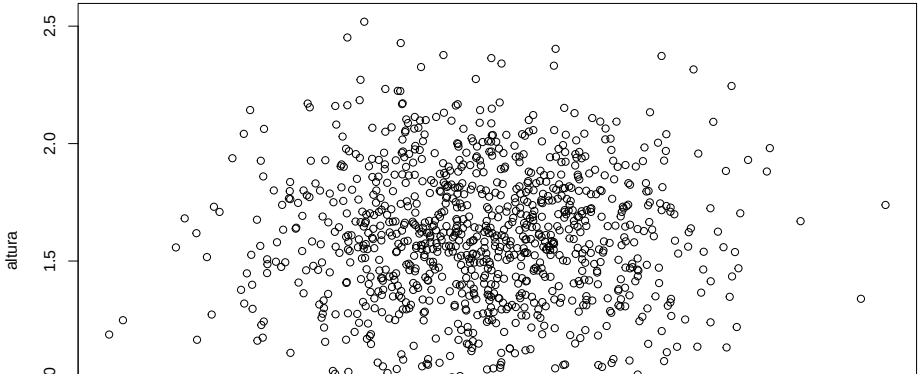
NO R usamos as funções rnom(), pnorm e qnorm para isso. se quisermos saber a probabilidade de um valor menor ou igual à media, podemos fazer:

```
pnorm(2.093456,mean = 1.6,sd = 0.3) #probabilidade de um valor menor ou igu
```

```
## [1] 0.95
```

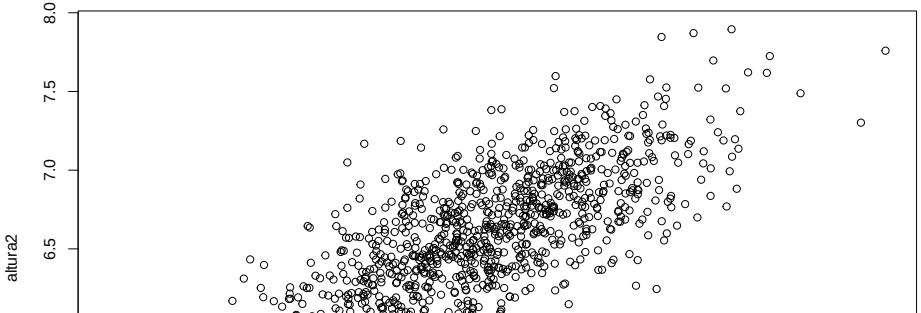
# Relação entre variáveis

Se conhecemos a distribuição de duas variáveis, podemos perguntar como uma afeta a outra. Por exemplo, se medimos a idade e a altura de vária pessoas, a idade influencia na altura? Podemos verificar isso plotando uma contra a outra. Definindo idade como preditora, isto é, idade influencia altura, e não o contrário, temos:



A dispersão dos pontos nos indica que altura e idade são independentes, ou seja um aumento em idade não implica em um aumento de altura. Isto é esperado pois nossos dados foram gerados independemtemente. O que aconteceria se idade de fato influenciasse altura?

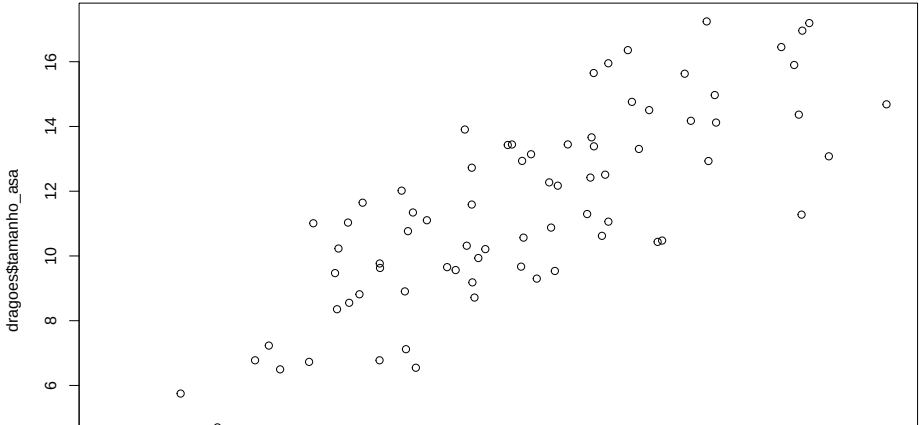
```
altura2= altura + sqrt(idade)
#escolhi somar a raiz da idade ao valor de altura para ilustrar, poderia t
plot(altura2~idade)
```



# Essa relação é forçada, vamos ver dados reais

Dragões!

```
plot(dragoes$tamanho_asa ~ dragoes$idade)
```



# Como encontrar essa relação?

Em outras palavras, qual equação posso usar para prever o tamnho do dragão se eu souber a idade? mesmo com uma relação clara (valores de um lado do gráfico são diferentes de valores no outro), aparementemente um dragão com cerca de 200 anos pode ter um tamanho entre 10 e 16 metros (cada asa). Qual o valor mais provável para um dragão de 200 anos?

## Modelos lineares

Podemos `ajustar um modelo` a nossos dados, isto é, estimar a reta que melhor descreve a relação entre nossas variáveis. Modelos lineares são usados para calcular a influencia de uma variável em outra, desde que algumas premissas sejam cumpridas:

1 - Relação linear: relações quadráticas, logisticas, ou qualquer coisa que não seja uma reta não podem ser analisadas com modelos lineares
2 - Normalidade das variáveis
3 - Homogeneidade das variâncias
4 - Independência

# No R, fazemos isso instantaneamente com a função `lm()`

```
lm(dragoes$tamanho_asa ~ dragoes$idade)

##
## Call:
## lm(formula = dragoes$tamanho_asa ~ dragoes$idade)
##
## Coefficients:
##   (Intercept)  dragoes$idade
##       4.59499         0.05747
```

O modelo linear nos retorna tudo que precisamos para estimar nossa reta: o `intercepto`, isto é o ponto em que ela cruza o eixo y, e a inclinação. Podemos então estimar o valor de y da seguinte forma:

$$\hat{y} = 4.60279 + 0.05753 * x$$

Tecnicamente, temos um termo de erro $\epsilon$, que representa a variação de y em torno do valor previsto pela reta.

summary() nos dá essa informação:

```
summary(lm(dragoes$tamanho_asa ~ dragoes$idade))
```

```
##
## Call:
## lm(formula = dragoes$tamanho_asa ~ dragoes$idade)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7602 -1.2804 -0.0276  1.3044  3.4914
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.594994   0.560899   8.192 4.34e-12 ***
## dragoes$idade 0.057467   0.004564  12.591  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

# Um breve interlúdio sobre valor de p, significância e hipótese nula

Valor de $p$ é probabilidade de encontrarmos, ao acaso, valores da estatítica de interesse iguais ou mais extremos do que o que encontramos nos nossos dados. No modelos lineares, como usamos mais de uma variável, usamos a estatística $F$ para descrever o efeito da variável independente sobre a dependente.

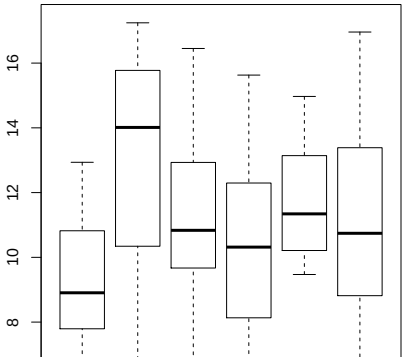
O nível de significância $\alpha$ representa a probabilidade máxima aceitável de encontrar o mesmo resultado por acaso. Um $\alpha = 0.05$ significa que, se repetimos nossas medidias várias vezes, uma vez em cada vinte podemos esperar encontrar o mesmo padrão de nossos dados, mesmo que não haja efeito de uma variável na outra

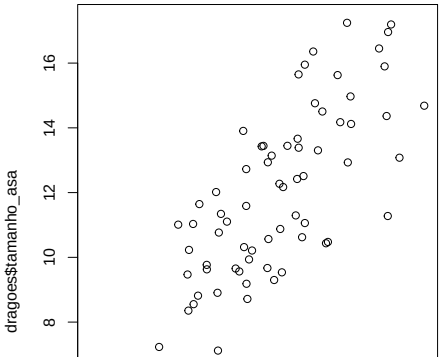
Hipótese nula nada mais é do que um modelo de como a variável dependente agiria se não houvesse efeito da variável independente

# Variável categórica

Não existe diferença prática para se ajustar um modelo com variáveis categóricas ou contínuas no R.

Quando a preditora é categórica, chamamos a análise de *ANOVA*. Quando a preditora é contínua, chamamos de *Regressão*



A interpretação do `lm()` é a mesma:

```
lm(dragoes$tamanho_asa~dragoes$cor)
```

```
##
## Call:
## lm(formula = dragoes$tamanho_asa ~ dragoes$cor)
##
## Coefficients:
##         (Intercept)    dragoes$corbranco   dragoes$cordourado
##               9.022                3.820                2.206
##    dragoes$corpreto     dragoes$corverde  dragoes$corvermelho
##               1.126                2.210                2.230
```

```
summary(lm(dragoes$tamanho_asa~dragoes$cor))

##
## Call:
## lm(formula = dragoes$tamanho_asa ~ dragoes$cor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2041 -1.7150  0.1105  2.0932  5.7071
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            9.0221     0.9835   9.174  8.8e-14 ***
## dragoes$corbranco      3.8199     1.2776   2.990   0.0038 **
## dragoes$cordourado     2.2060     1.4252   1.548   0.1260
## dragoes$corpreto       1.1258     1.3909   0.809   0.4209
## dragoes$corverde       2.2098     1.2622   1.751   0.0842 .
## dragoes$corvermelho    2.2301     1.3142   1.697   0.0940 .
## ---
```

# Mais de uma preditora

Basta incluir a variável na formula usando o sinal de mais, podemos incluir quantas quisermos:

```
lm(dragoes$tamanho_asa ~ dragoes$idade + dragoes$peso)
```

```
##
## Call:
## lm(formula = dragoes$tamanho_asa ~ dragoes$idade + dragoes$peso)
##
## Coefficients:
##   (Intercept)  dragoes$idade   dragoes$peso
##      4.273553       0.057249       0.002634
```

```
summary(lm(dragoes$tamanho_asa ~ dragoes$idade + dragoes$peso))

##
## Call:
## lm(formula = dragoes$tamanho_asa ~ dragoes$idade + dragoes$peso)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6386 -1.3392 -0.0376  1.3347  3.6292
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.273553   0.802122   5.328 9.81e-07 ***
## dragoes$idade 0.057249   0.004601  12.443  < 2e-16 ***
## dragoes$peso  0.002634   0.004679   0.563    0.575
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.945 on 76 degrees of freedom
```

Podemos incluir também interação entre fatores

```
lm(dragoes$tamanho_asa ~ dragoes$idade * dragoes$cor)
```

```
##
## Call:
## lm(formula = dragoes$tamanho_asa ~ dragoes$idade * dragoes$cor)
##
## Coefficients:
##                          (Intercept)                         dragoes$idade
##                             2.482472                             0.070025
##                   dragoes$corbranco                   dragoes$cordourado
##                             2.803526                             3.078620
##                    dragoes$corpreto                    dragoes$corverde
##                             1.366992                             3.444000
##                  dragoes$corvermelho    dragoes$idade:dragoes$corbranco
##                             2.051974                            -0.010953
##  dragoes$idade:dragoes$cordourado     dragoes$idade:dragoes$corpreto
##                            -0.022284                            -0.012897
```

```
summary(lm(dragoes$tamanho_asa ~ dragoes$idade * dragoes$cor))

##
## Call:
## lm(formula = dragoes$tamanho_asa ~ dragoes$idade * dragoes$cor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5820 -1.2640  0.1833  1.4492  2.7301
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        2.482472   2.205317   1.126  0.26432
## dragoes$idade                      0.070025   0.022793   3.072  0.00307
## dragoes$corbranco                  2.803526   2.523969   1.111  0.27064
## dragoes$cordourado                 3.078620   2.921599   1.054  0.29579
## dragoes$corpreto                   1.366992   2.655914   0.515  0.60846
## dragoes$corverde                   3.444000   2.477940   1.390  0.16917
## dragoes$corvermelho                2.051974   2.541789   0.807  0.42235
```