Módulo de Estatística

aula 3

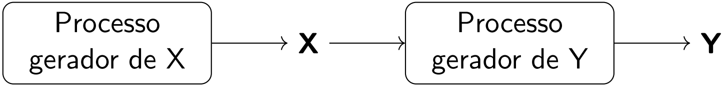
Guilherme Gomes

FGV MANAGEMENT

1. Introdução
2. Mínimos Quadrados Ordinários
3. Qualidade do ajuste
4. Múltiplos regressores
5. Regressão Linear
6. Regressão linear generalizada

Até o momento consideramos que os processos geradores de uma variável X são **independentes** de qualquer outra variável.

Agora iremos considerar modelos que o processo gerador de uma variável Y **depende** da variável X.



As variáveis X e Y recebem nomes especiais:

- **Variável** Y: <u>dependente</u>, de resposta, explicada, prevista ou regressando;
- **Variável** X: <u>independente</u>, de controle, explicativa, previsora ou regressor.

Utilizaremos nessa aula a base de dados `survey` do pacote `MASS`.

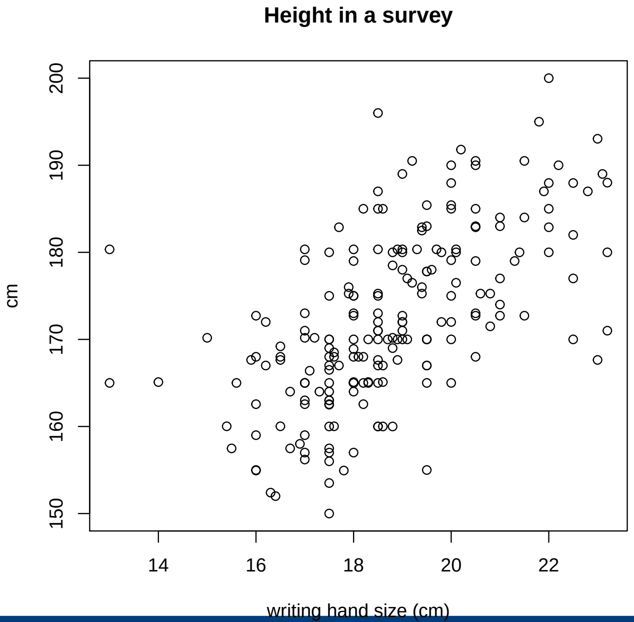
1. Instale o pacote `MASS`;
2. Em gráficos de boxplot com as médias, apresente a diferença existente entre a variável `Height` entre homens e mulheres;
3. Teste a hipótese de que as médias são diferentes a um nível de significância de 5%.

Note que algumas das observações são `NA` para a variável `Height`:

```
> sum(is.na(survey$Height))
[1] 28
```

Qual a melhor previsão que podemos fazer para esses valores?

- Podemos usar uma medida de centralidade global;
- Podemos usar uma medida de centralidade para cada grupo.


Height in a survey
cm
200
190
180
170
160
150
14
16
18
20
22
writing hand size (cm)

Essas são **medidas unitárias** para calcular a relação entre duas variáveis;

Uma outra forma de calcular essa relação é por meio de uma **função afim**;

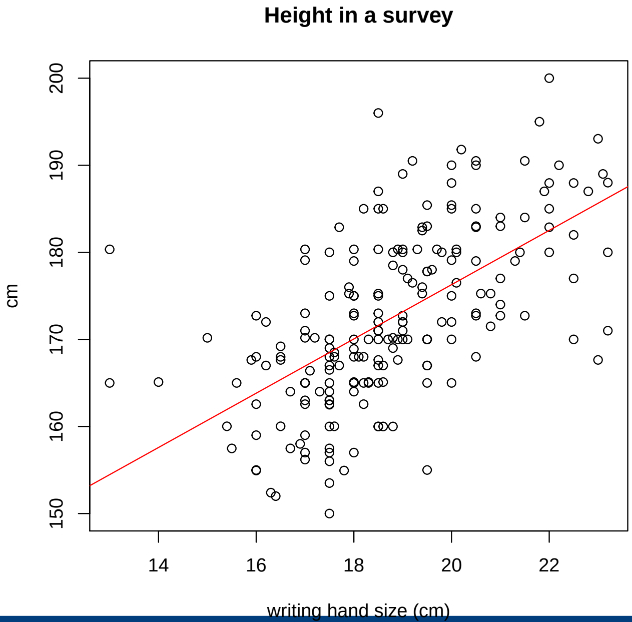
Queremos encontrar uma reta que resuma a informação do gráfico.

Equação da reta:

$$\hat{y} = b_0 + b_1.x$$

onde:
b_0 é o coeficiente linear (ou intercepto), e
b_1 é o coeficiente angular.

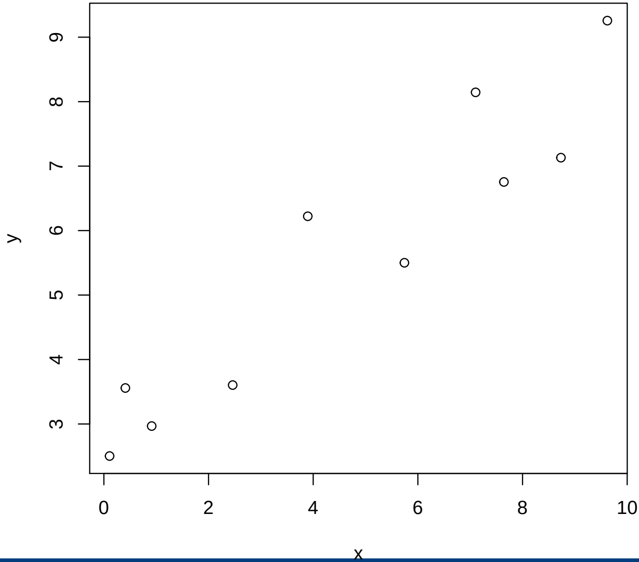
```
> lm(Height ~ Wr.Hnd,data = survey)

Call:
lm(formula = Height ~ Wr.Hnd, data = survey)

Coefficients:
(Intercept)       Wr.Hnd
    113.954        3.117
```


Height in a survey
cm
200
190
180
170
160
150
14
16
18
20
22
writing hand size (cm)

A título de simplificação vamos utilizar dados simulados.

Simulei um conjunto de dados com duas variáveis x e y e 10 observações.

Tais dados são desenhados em um gráfico de pontos, note que eles estão espalhados em torno de uma linha imaginária.



y
3
4
5
6
7
8
9
0
2
4
6
8
10
x

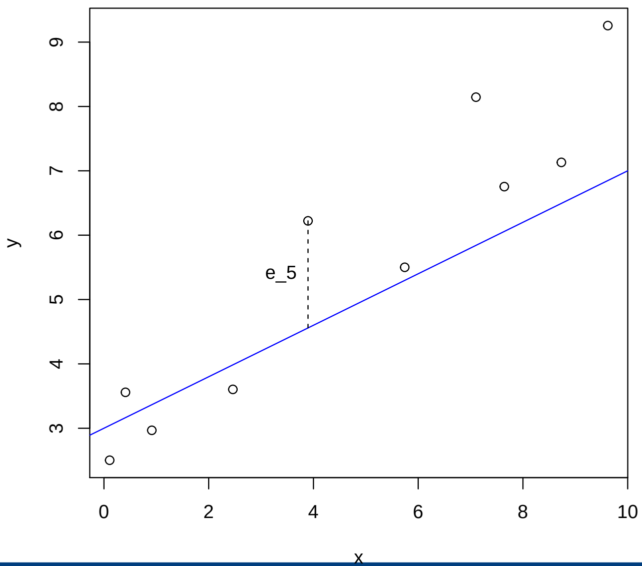
Qual a melhor reta que podemos encontrar?

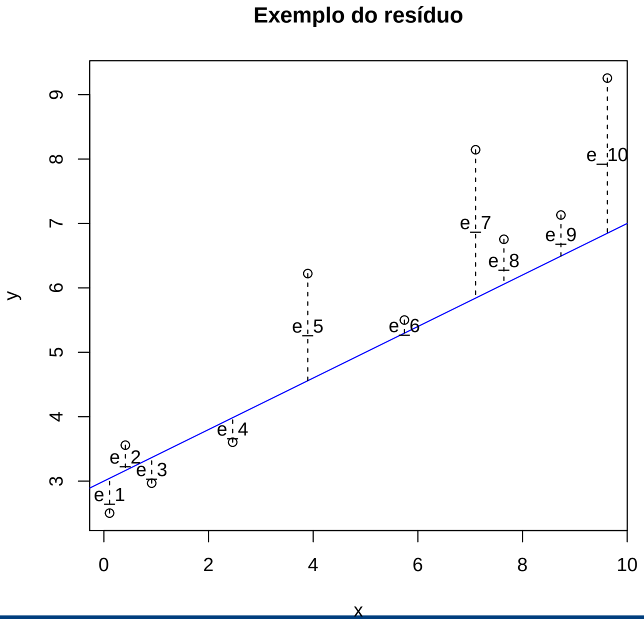
Aquela que está mais próxima de todos os pontos.

Defina a distância da reta ao ponto i como sendo e_i, chamado de **resíduo**.

Desse modo temos que $e_i = y_i - \hat{y}_i$.

Exemplo do resíduo




Exemplo do resíduo
e_1
e_2
e_3
e_4
e_5
e_6
e_7
e_8
e_9
e_10
0
2
4
6
8
10
3
4
5
6
7
8
9
x
y

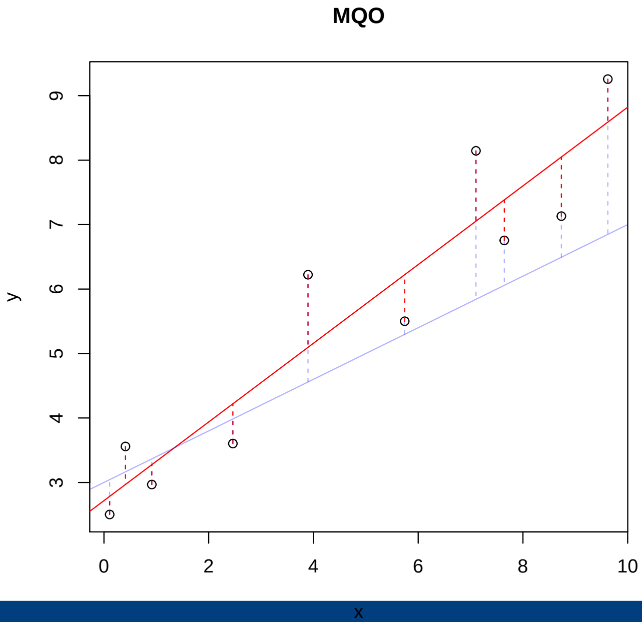
Os coeficiêntes linear e angular - (b_0^*, b_1^*) - para melhor reta são dados pela **minimização dos quadrados dos resíduos**:

$$[b_0^*, b_1^*] = argmin_{b0,b1} SQR = argmin_{b0,b1} \sum_{i=1}^{n} e_i^2$$
$$= argmin_{b0,b1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$= argmin_{b0,b1} \sum_{i=1}^{n} (y_i - b0 - b1.x_i)^2$$

A solução do problema de minimização acima é dada por:

$$b_1^* = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{cov(x,y)}{var(x)}$$

$$b_0^* = \bar{y} - b_1^*.\bar{x}$$


MQO
y
3
4
5
6
7
8
9
0
2
4
6
8
10
x

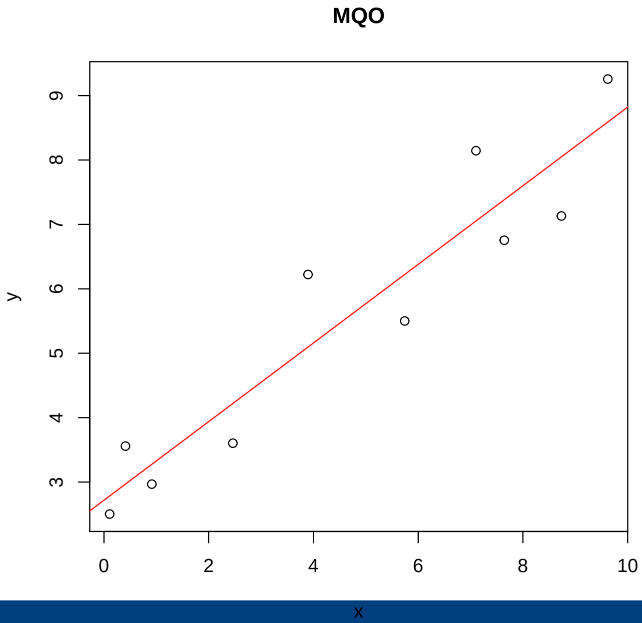
ANOVA é uma sigla para análise de variância.

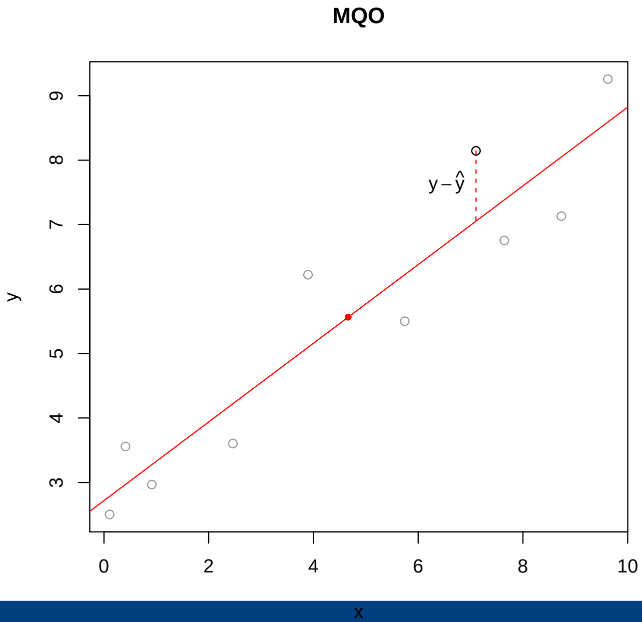
Queremos identificar quanto da variância nos dados pode ser explicada pelo modelo.

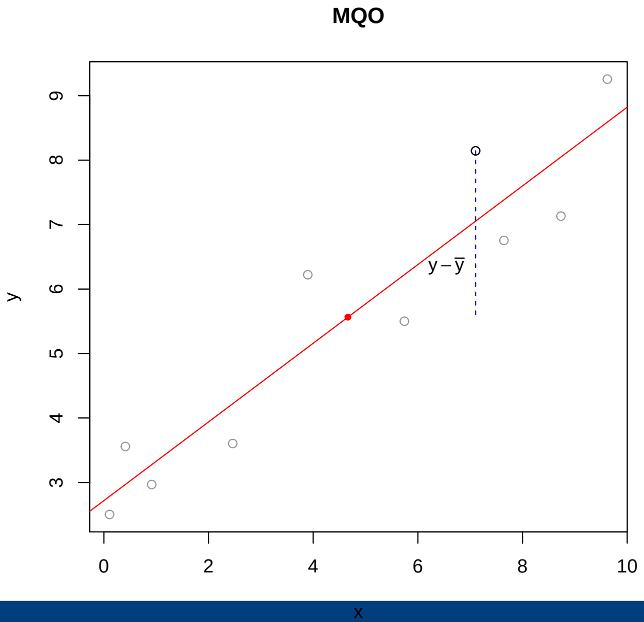
Considere:

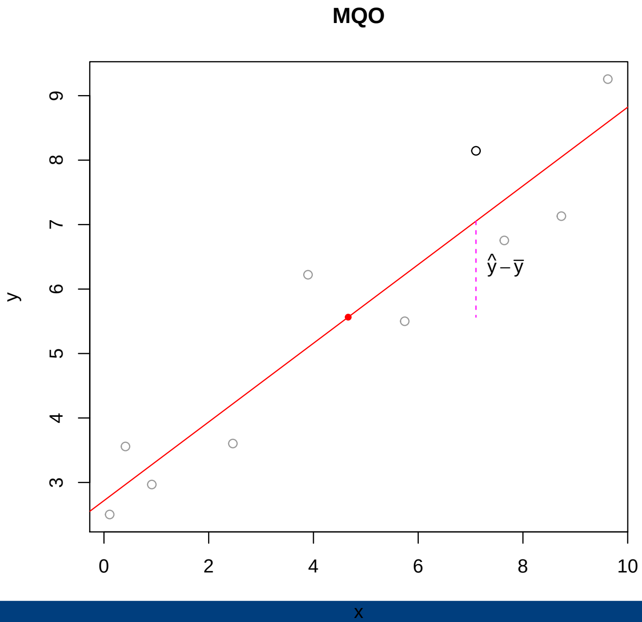
- y_i: o valor observado da variável;
- $\hat{y}_i$: o valor previsto da variável;
- $\bar{y}$: a média dos valores observados.

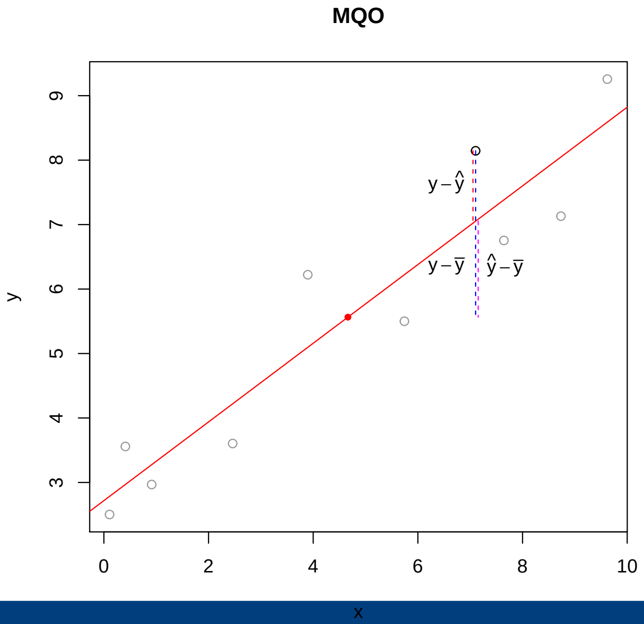
FGV MANAGEMENT




MQO
y
x
0
2
4
6
8
10
3
4
5
6
7
8
9
$y - \hat{y}$


MQO
y
x
$y - \bar{y}$
0
2
4
6
8
10
3
4
5
6
7
8
9


MQO
y
3
4
5
6
7
8
9
0
2
4
6
8
10
$\hat{y} - \bar{y}$
x


MQO
$y-\hat{y}$
$y-\bar{y}$
$\hat{y}-\bar{y}$
y
3
4
5
6
7
8
9
0
2
4
6
8
10
x

A seguinte identidade é válida para todo i:

$$y_i - \bar{y}_i = (\hat{y}_i - \bar{y}_i) + (y_i - \hat{y}_i)$$

O que não é imediato, é que a igualdade abaixo também é válida:

$$\sum_i^n (y_i - \bar{y}_i)^2 = \sum_i^n (\hat{y}_i - \bar{y}_i)^2 + \sum_i^n (y_i - \hat{y}_i)^2$$

Seja:

- Soma dos quadrados totais: $SQT = \sum_i^n (y_i - \hat{y}_i)^2$
- Soma dos quadrados explicados: $SQE = \sum_i^n (\hat{y}_i - \bar{y}_i)^2$
- Soma dos quadrados dos resíduos: $SQR = \sum_i^n (y_i - \hat{y}_i)^2$

$$SQT = SQE + SQR$$

Defino a variável R^2:

$$R^2 = \frac{SQE}{SQT} = 1 - \frac{SQR}{SQT}$$

Quanto o modelo explica dos dados?

1. R^2 é um valor entre 0 e 1;
2. Quanto mais próximo de 1, mais próximos os pontos estão da reta;
3. Quanto mais próximo de 0, mais distantes os pontos da reta.

obs: Em um modelo linear com apenas uma variável explicativa temos que $R^2 = (\text{correlação})^2$

```
> reg<-lm(Height ~ Wr.Hnd,data=survey)
> SQR<-sum(reg$residuals^2)
> SQT<-sum((survey$Height[!is.na(survey$Height)]
+           -mean(survey$Height,na.rm = T))^2)
> 1-SQR/SQT
```

```
[1] 0.3611947
```

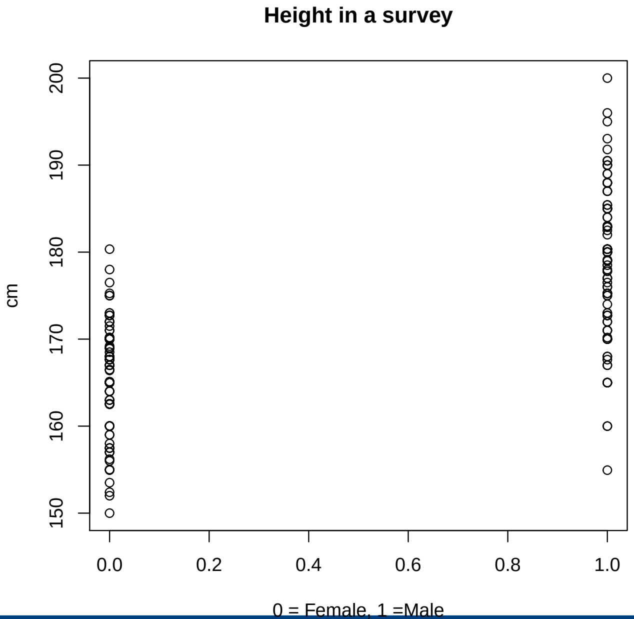
Calculando pela correlação:

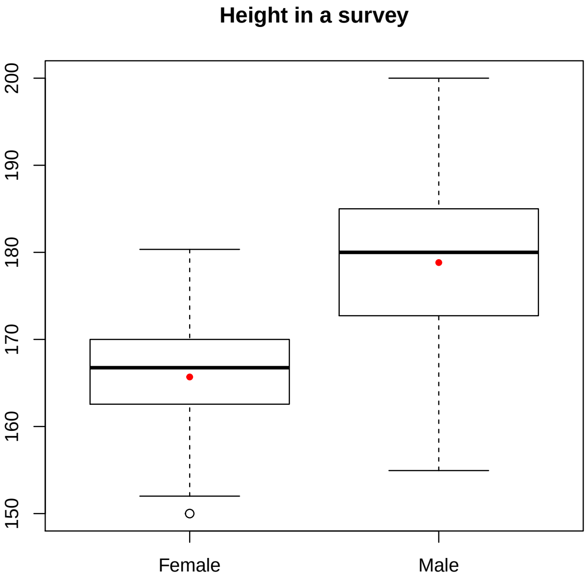
```
> with(survey,cor(Height,Wr.Hnd,use = 'complete.obs'))^2
```

```
[1] 0.3611901
```

Assim aproximadamente 36,12% da variância da altura é explicada em um modelo linear pela variável amplitude da mão.

Vamos olhar para outras variáveis que parecem explicar a altura, uma delas é o sexo.


Height in a survey
cm
200
190
180
170
160
150
0.0
0.2
0.4
0.6
0.8
1.0
0 = Female, 1 =Male



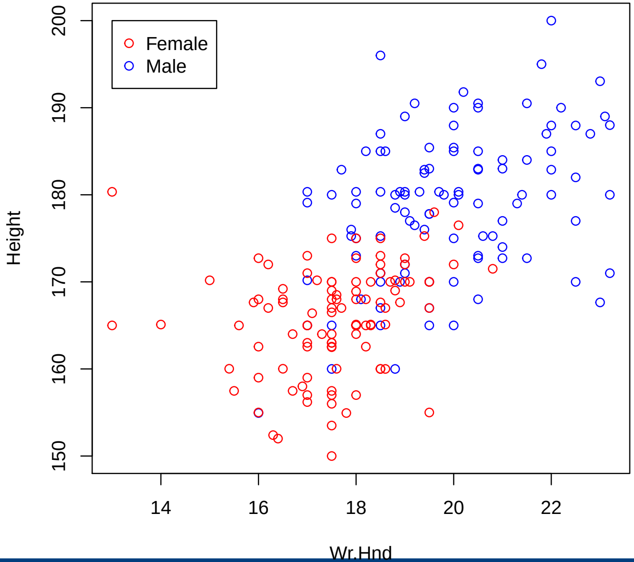
Height in a survey
200
190
180
170
160
150
Female
Male

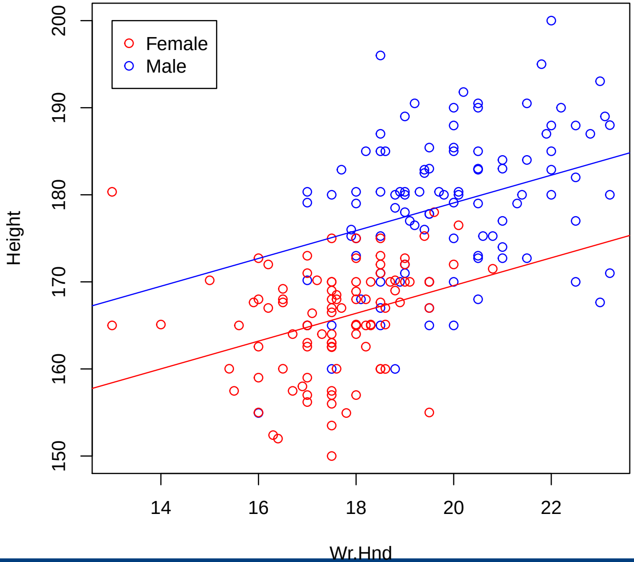
Vamos introduzir a variável sexo como regressor

```
> lm(Height ~ Wr.Hnd + Sex,data = survey)

Call:
lm(formula = Height ~ Wr.Hnd + Sex, data = survey)

Coefficients:
(Intercept)       Wr.Hnd      SexMale
    137.687        1.594        9.490
```

```
> reg<-lm(Height ~ Wr.Hnd + Sex,data=survey)
> SQR<-sum(reg$residuals^2)
> SQT<-sum((survey$Height[!is.na(survey$Height)]
+           -mean(survey$Height,na.rm = T))^2)
> 1-SQR/SQT
[1] 0.5062514
```

O modelo fica claramente melhor, conseguindo explicar 50% da variação dos dados.


Female
Male
Height
Wr.Hnd
150
160
170
180
190
200
14
16
18
20
22



Female
Male
Height
Wr.Hnd
150
160
170
180
190
200
14
16
18
20
22

O processo gerador da variável Y será modelado como sendo uma função "linear" da variável X.

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

Onde ϵ é uma variável aleatória tal que:

- $cov(X, \epsilon) = 0$
- $E[\epsilon] = 0$

Assim a média condicional da variável aleatória Y é modelada como sendo dependente da variável X.

$$\begin{aligned} E[Y|X] &= E[\beta_0 + \beta_1.X + \epsilon|X] \\ &= \beta_0 + \beta_1.E[X|X] \\ &= \beta_0 + \beta_1.X \end{aligned}$$

É possível mostrar que os melhores estimadores para β_0 e β_1 são obtidos pelo MQO. Assim:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1.X$$

Teste de hipótese:

- H0: o parâmetro é igual a zero;
- H1: o parâmetro é diferente de zero.

Utilizando como regressores `Wr.Hnd` e `Sex`, qual a melhor previsão para os dados que estão missings (`NA`).

```
> previsao<-predict(reg
+ ,newdata = survey[is.na(survey$Height),])
> head(previsao,5)
```

```
       3        12        15        25        26
175.8768 180.6601 172.6879 164.7925 176.6740
```

Um laboratório farmaceutico tem interesse em estudar os benefícios da vitamina C. Mais especificamente gostariam de estudar a conjectura de que a ingestão de vitamina C auxilia o crescimento dos ossos.

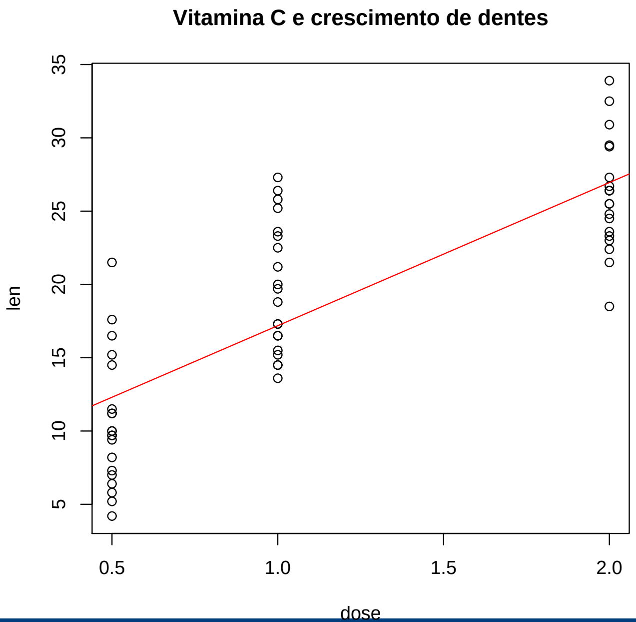
Fonte: http://www.mdig.com.br/index.php?itemid=11670

Para desenvolver tal pesquisa foram utilizados porcos da índia. Cada animal recebeu doses diferentes de Vitamina C e foram registrados os tamanhos de seus dentes.
Utilizando a base de dados `ToothGrowth`.:

1. Estime um modelo linear com a variável dependente sendo o tamanho do dente - `len` - e a variável explicativa sendo a quantidade ingerida - `dose`;
2. Verifique se o regressor é significativo a 95% de confiança;
3. Em um gráfico apresente os dados e a melhor reta.

Modelo linear - função `lm`

```
> reg <- lm(len ~ dose, data = ToothGrowth)
```


Vitamina C e crescimento de dentes
len
dose
35
30
25
20
15
10
5
0.5
1.0
1.5
2.0

Voltando ao exemplo anterior, temos outra variável que é a maneira como a Vitamina C é aplicada - `supp`:

1. Estime o modelo linear com a variável dependente sendo o tamanho do dente - `len` - e as variável explicativas sendo a quantidade ingerida - `dose` - e o tipo de ingestão - `supp`;
2. Verifique se o regressor é significativo a 95% de confiança;

Modelo linear - função `lm`

```
> reg <- lm(len ~ dose + supp, data = ToothGrowth)
```

Agora suponha que a variável Y só assume dois valores: tipicamente 0 e 1.
Desse modo queremos modelar Y como sendo uma v.a. que segue distribuição Bernoulli.

$$Y = 1, \quad \text{com prob. } p$$
$$Y = 0, \quad \text{com prob. } 1 - p$$

lembre-se que:

$$E[Y] = 1.p + 0.(1 - p) = p \in [0, 1]$$

Desse modo:

$$Y = \beta_0 + \beta_1.X + \epsilon$$
$$p = E[Y|X] = \beta_0 + \beta_1.X$$

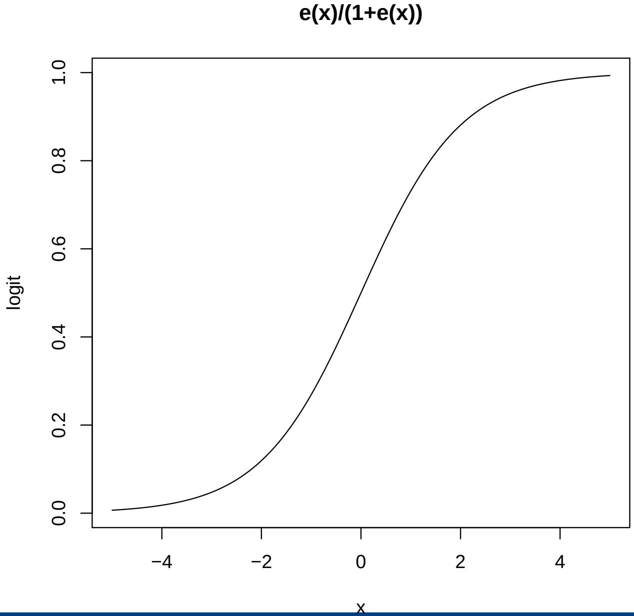
não estará bem especificado.

Seja m uma função tal que $m : \mathbb{R} \rightarrow [0, 1]$.

$$m(x) = \frac{e^x}{1 + e^x}$$

Assim basta definir o modelo tal que:

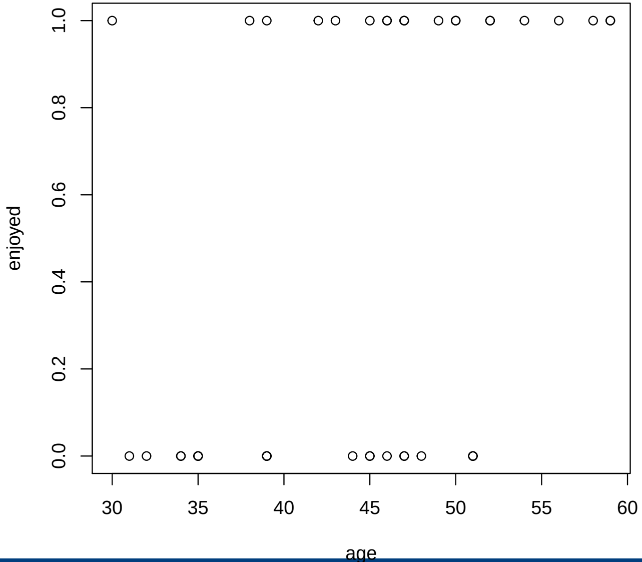
$$Y \sim Bernoulli$$
$$p = E[Y|X] = m(\beta_0 + \beta_1.X)$$


e(x)/(1+e(x))
logit
0.0
0.2
0.4
0.6
0.8
1.0
−4
−2
0
2
4
x

Considere a base de dados `tastesgreat` do pacote `UsingR`.

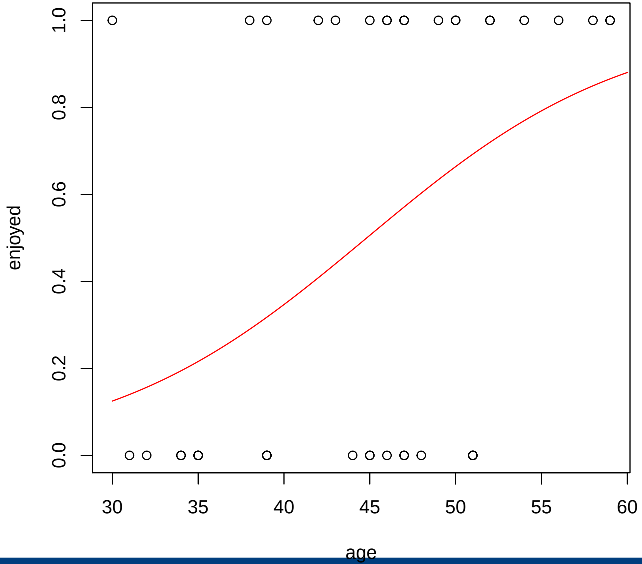
Fonte: http://cdn2.hubspot.net/hub/215841/file-3945428210-jpg/blog-files/new-product-blog-602x347pix.jpg

Estime um modelo logit para a variável `enjoyed` dado a idade (`age`)


enjoyed
1.0
0.8
0.6
0.4
0.2
0.0
30
35
40
45
50
55
60
age

Regressão linear generalizada

FGV MANAGEMENT



```
> library(UsingR)
> reg<-glm(enjoyed ~ age,data = tastesgreat,family=binomial(
> reg

Call:  glm(formula = enjoyed ~ age, family = binomial(), data

Coefficients:
(Intercept)          age
    -5.8876       0.1314

Degrees of Freedom: 39 Total (i.e. Null);  38 Residual
Null Deviance:           55.45
Residual Deviance: 47.36         AIC: 51.36
```

Como interpretar os coeficientes? A chance de um evento A ocorrer é definido como:

$$O_A = \frac{P(A)}{1 - P(A)}$$

A razão de chances entre dois eventos é dado por:

$$\frac{O_A}{O_B} = \frac{P(A)}{1 - P(A)} \cdot \left(\frac{P(B)}{1 - P(B)} \right)^{-1}$$

A razão de chances entre o $Y = 1$ dado o aumento de uma unidade da variável X_i e $Y = 1$ considerando todos os regressores constantes é dado por:

$$exp(\hat{\beta}_i)$$

Se esse valor for maior do que 1, então a variável X_i tem aumenta a chance de $Y_i = 1$ ocorrer, se for negativo, então reduz a chance desse evento ocorrer.

Em nosso exemplo:

$$exp(\hat{\beta}_1) = exp(0.1314) = 1.14$$

Acrescentando mais regressores

```
> library(UsingR)
> reg<-glm(enjoyed ~ age + gender,data = tastesgreat,family=b
> reg

Call:  glm(formula = enjoyed ~ age + gender, family = binomia

Coefficients:
(Intercept)          age   genderMale
    -8.1844       0.1649       2.4224

Degrees of Freedom: 39 Total (i.e. Null);  37 Residual
Null Deviance:          55.45
Residual Deviance: 38.98        AIC: 44.98
```

O fato de ser homem aumenta a chance de gostar do produto em:

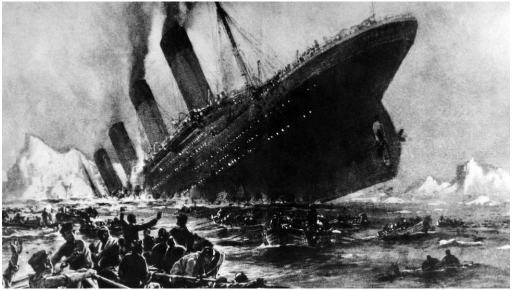
$$exp(\hat{\beta}_2) = 11.27$$

Esse produto parece então ser mais indicado para homens.

Podemos usar essa estimação da amostra para inferir algo sobre a população:

1. Qual a probabilidade de uma mulher de 34 anos gostar do produto?
2. E de um homem de $25,2$ anos?

Considere a base de dados `train.csv` com os dados do Titanic do Kaggle.



Fonte: http://news.bbcimg.co.uk/media/images/59467000/jpg/$_5$9467081¡*llustrationofsinking*96518519.*jpg*

Estime a probabilidade de um indivíduo morrer dado o preço que pagou pela passagem, sua idade e seu sexo.