ME111 - Laboratório de Estatística

Aula 15 - Regressão

Profa. Larissa Avila Matos

Regressão

- Considere duas variáveis X e Y. Tome n pares (X_1, Y_1), (X_2, Y_2), $\ldots$, (X_n, Y_n) dessas variáveis.

- Se Y é uma função linear de X, é possivél estabelecer uma regressão linear simples cujo modelo estatístico é

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n,$$

  - Y_i é uma variável aleatória e representa o valor da variável resposta (variável dependente) para a i-ésima observação,

  - x_i representa o valor da variável explicativa (variável independente, variável regressora) para a i-ésima observação,

  - ϵ_i é uma variável aleatória que representa o erro experimental,

  - β_0 e β_1 são os parâmetros desconhecidos do modelo.

- Uma vez que X_i é uma variável determinística (constante conhecida), substituímos X_i por x_i.

Interpretação dos parâmetros do modelo

- O parâmetro β_0 é chamado intercepto ou coeficiente linear e representa o ponto em que a reta regressora corta o eixo dos y's, quando $x = 0$.

- O parâmetro β_1 representa a inclinação da reta regressora e é dito coeficiente de regressão ou coeficiente angular. Além disso, temos que para um aumento de uma unidade na variável x, o valor $E(Y|x)$ aumenta β_1 unidades.

Suposições para o modelo

1. A relação entre Y e X é linear e os valores de x são fixos.
2. A média do erro é nula, $E(\epsilon_i) = 0, \quad i = 1. \ldots, n.$

$$E(Y_i) = E(\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 + \beta_1 x_i + E(\epsilon_i) = \beta_0 + \beta_1 x_i$$

$$\Rightarrow E[Y|x] = \beta_0 + \beta_1 x.$$

3. O erro é homocedástico (tem variância constante).

$$Var(\epsilon_i) = E(\epsilon_i^2) - [E(\epsilon_i)]^2 = E(\epsilon_i^2) = \sigma^2, \ \forall i$$

$$\Rightarrow Var(Y_i) = E[Y_i - E(Y_i|x_i)]^2 = E(\epsilon_i^2) = \sigma^2.$$

4. Os erros são não correlacionados.

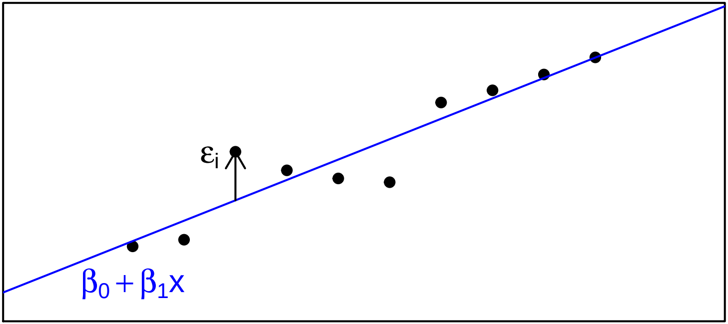
$$Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i, \varepsilon_j) - E(\varepsilon_i)E(\varepsilon_j) = E(\varepsilon_i, \varepsilon_j) = 0, \quad \text{para} \quad i \neq j.$$

5 Vamos assumir que os erros tem distribuição Normal, ou seja, $\epsilon_i \sim N(0, \sigma^2)$.

- Se $\epsilon_i \sim N(0, \sigma^2) \Rightarrow Y_i \sim N(\beta_0 + \beta_1 x_i; \sigma^2)$.

6 Por (4) e (5) temos que Y_i e Y_j são independentes ($Y_i \perp Y_j$.)

- O primeiro passo na análise de regressão é obter as estimativas dos parâmetros do modelo, ou seja, encontrar as estimativas de $\widehat{\beta}_0$ e $\widehat{\beta}_1$.
- O objetivo é estimar os parâmetros β_0 e β_1 de modo que os desvios $(\epsilon_i = Y_i - [\beta_0 + \beta_1 x_i])$ entre os valores observados e estimados sejam mínimos. Isso equivale a minimizar o comprimento do vetor de erros, $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)'$.



- No método de Mínimos Quadrados não é necessário conhecer a forma da distribuição dos erros.
- Então, o método de MQ consiste em minimizar a soma dos quadrados dos desvios $S(\beta_0, \beta_1)$,

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 x_i]^2.$$

- **Observação:** A distância entre a reta e os valores observados poderia ser calculada de diferentes formas. A escolha do quadrado está na simplicidade dos cálculos envolvidos.
- Para encontrarmos as estimativas dos parâmetros, devemos minimizar $S(\beta_0, \beta_1)$ em relação aos parâmetros β_0 e β_1.

- Assim, derivamos $S(\beta_0, \beta_1)$ em relação aos parâmetros β_0 e β_1.
- As derivadas são dadas por

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i); \quad \text{e}$$

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) x_i.$$

- Substituindo β_0 e β_1 por $\widehat{\beta}_0$ e $\widehat{\beta}_1$, para indicar valores particulares dos parâmetros que minimizam $S(\beta_0, \beta_1)$, e igualando as derivadas parciais a zero, obtemos

$$-2\sum_{i=1}^{n}(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0; \quad \text{e}$$

$$-2\sum_{i=1}^{n}(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)x_i = 0.$$

- Portanto, para encontrar as estimativas dos parâmetros devemos resolver o seguinte sistema de equações

$$\begin{cases} n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} Y_i \\ \widehat{\beta}_0 \sum_{i=1}^{n} x_i + \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i Y_i. \end{cases}$$

- Resolvendo o sistema, temos que

$$\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{x}; \quad \text{e}$$

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}.$$

- $\widehat{\beta}_0$ e $\widehat{\beta}_1$ são chamados de Estimadores de Mínimos Quadrados (EMQ) e os valores de $\widehat{\beta}_0$ e $\widehat{\beta}_1$ são chamados de estimativas de Mínimos Quadrados.

- Portanto, o modelo de regressão linear simples ajustado é dado por

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i, \quad i = 1, \ldots, n,$$

onde $\widehat{Y}$ é um estimador pontual da média da variável Y para um valor de x.

Resíduos

- A diferença entre o valor observado Y_i e o correspondente valor ajustado $\widehat{Y}_i$ é chamada de resíduo(r_i) e é denotado por

$$r_i = Y_i - \widehat{Y}_i = Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i).$$

- Essa medida é importante já que por meio dela verificamos o ajuste do modelo.

Propriedades

1. A soma dos resíduos é sempre nula ($\sum_{i=1}^{n} r_i = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i) = 0$).

2. A soma dos valores observados Y_i é igual a soma dos valores ajustados $\widehat{Y}_i$ ($\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \widehat{Y}_i$).

3. A reta de regressão de mínimos quadrados passa pelo ponto $(\bar{x}, \bar{Y})$.

4. A soma dos resíduos ponderado pelo correspondente valor da variável regressora é sempre nula ($\sum_{i=1}^{n} x_i r_i = 0$).

5. A soma dos resíduos ponderado pelo correspondente valor ajustado é sempre zero ($\sum_{i=1}^{n} \widehat{Y}_i r_i = 0$).

Exemplo cars

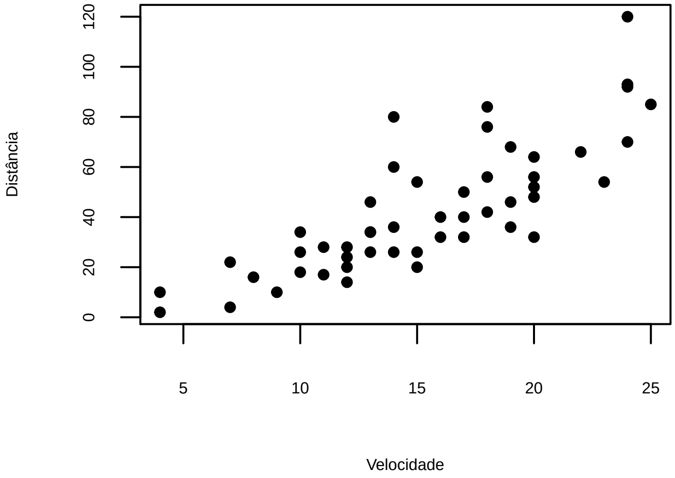
- Para uma análise inicial, usaremos o conjunto de dados de **cars**.
- Esse conjunto consiste de 50 observaçõese e 2 variáveis (**dist** e **speed**).

```
data(cars)
str(cars)
```

```
'data.frame':   50 obs. of  2 variables:
 $ speed: num  4 4 7 7 8 9 10 10 10 11 ...
 $ dist : num  2 10 4 22 16 10 18 26 34 17 ...
```

- Diagrama de dispersão

```
plot(x=cars$speed,y=cars$dist,xlab="Velocidade",ylab="Distância",main="",pch=20,
     cex.axis=0.5,cex.lab=0.5)
```



- **Objetivo:** obter uma medida que permita quantificar a dependência que pode existir entre duas variáveis (positiva, negativa, muita ou pouca).
- Dado n pares de observações $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$:

$$Corr(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

  onde s_x é o desvio padrão de X e s_y é o desvio padrão de Y.
- Essa medida leva em consideração todos os desvios $(x_i - \bar{x})$ e $(y_i - \bar{y})$ padronizados da forma $z_{x_i} = \frac{x_i - \bar{x}}{s_x}$ e $z_{y_i} = \frac{y_i - \bar{y}}{s_y}$.
- Interpretação: z_{x_i} indica o número de desvios-padrão que a observação x_i está afastada da média de X.

Propriedades

- $-1 \leq Corr(X, Y) \leq 1$
- $Corr(X, Y)$ próxima de 1: X e Y estão positivamente associadas e o tipo de associação entre as variáveis é linear.
- $Corr(X, Y)$ próxima de -1: X e Y estão negativamente associadas e o tipo de associação entre as variáveis é linear.
- Se z_x e z_y têm o mesmo sinal, estamos somando um termo positivo na expressão da correlação.
- Se z_x e z_y têm sinais opostos, estamos somando um termo negativo na expressão da correlação.

- Para as variáveis **dist** e **speed** nós temos que a correlação é

```r
round(cor(cars$speed,cars$dist),3)
```

```
[1] 0.807
```

- Ou seja, existe uma correlação linear positiva entre essas variáveis.

```r
ajuste <- lm(cars$dist ~ cars$speed)
ajuste
```

```
Call:
lm(formula = cars$dist ~ cars$speed)

Coefficients:
(Intercept)   cars$speed
    -17.579        3.932
```

```r
summary(ajuste)
```

```

Call:
lm(formula = cars$dist ~ cars$speed)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.069  -9.525  -2.272   9.215  43.201 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -17.5791     6.7584  -2.601   0.0123 *  
cars$speed    3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,	Adjusted R-squared:  0.6438 
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

```r
ajuste$coef[1]
```

```
(Intercept)
  -17.57909
```

```r
ajuste$coef[2]
```

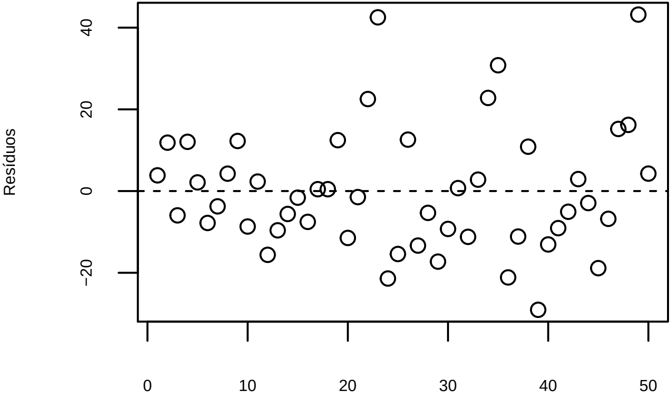
```
cars$speed
  3.932409
```

- Então, o modelo é dado por

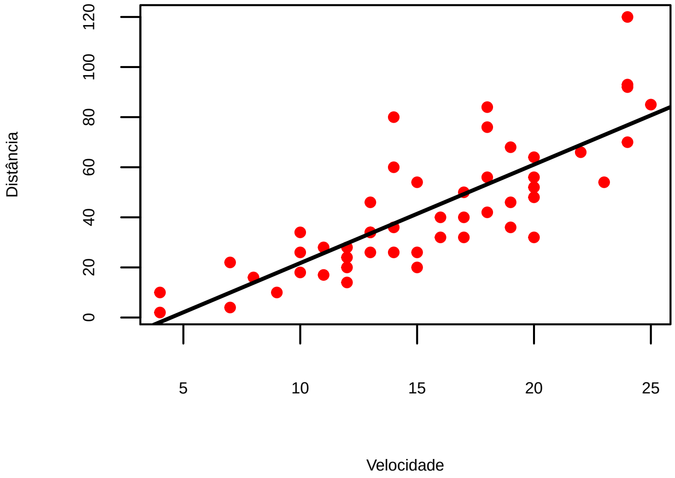
  `dist` = -17.5790949 + 3.9324088*`speed`

- Verificação das suposições do modelo

```
plot(ajuste$residuals,xlab="",ylab="Resíduos",main="",cex.axis=0.5,cex.lab=0.5)
abline(h=0,lty=2)
```



```r
plot(x=cars$speed,y=cars$dist,xlab="Velocidade",ylab="Distância",main="",pch=20,
     cex.axis=0.5,cex.lab=0.5,col="red")
abline(a = ajuste$coef[1], b = ajuste$coef[2], col = "black", lwd=2)
```



Exemplo: `SleepStudy`

- Os dados foram obtidos de uma amostra de alunos que fizeram testes de habilidades para medir a função cognitiva. Todos os alumos na pesquisa registraram o tempo e a qualidade do sono em um diário do sono durante um período de duas semanas.

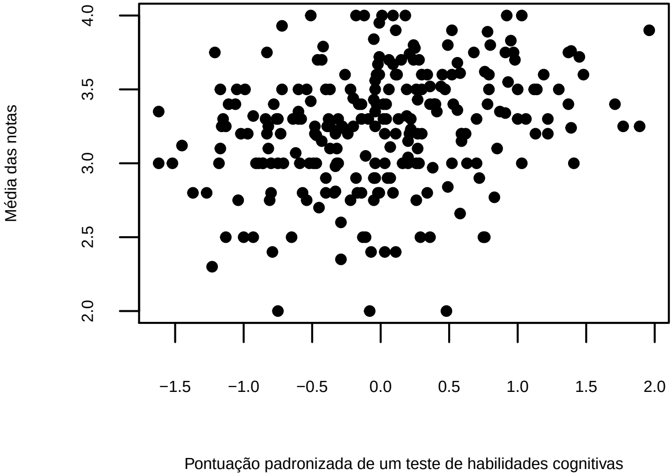
- Objetivo: Analisar a associação entre a variável **`GPA`** e a variável **`CognitionZscore`**, onde
  - **`GPA`**: Média das notas (escala 0-4); e
  - **`CognitionZscore`**: Pontuação padronizada em um teste de habilidades cognitivas.

```r
library(Lock5Data)
data(SleepStudy)
dados<-SleepStudy
str(dados)
```

```
'data.frame':   253 obs. of  27 variables:
 $ Gender          : int  0 0 0 0 0 1 1 0 0 0 ...
 $ ClassYear       : int  4 4 4 1 4 4 2 2 1 4 ...
 $ LarkOwl         : Factor w/ 3 levels "Lark","Neither",..: 2 2 3 1 3 2 1 1 2 2 ...
 $ NumEarlyClass   : int  0 2 0 5 0 0 2 0 2 2 ...
 $ EarlyClass      : int  0 1 0 1 0 0 1 0 1 1 ...
 $ GPA             : num  3.6 3.24 2.97 3.76 3.2 3.5 3.35 3 4 2.9 ...
 $ ClassesMissed   : int  0 0 12 0 4 0 2 0 0 0 ...
 $ CognitionZscore : num  -0.26 1.39 0.38 1.39 1.22 -0.04 0.41 -0.59 1.03 0.72 ...
 $ PoorSleepQuality: int  4 6 18 9 9 6 2 10 5 2 ...
 $ DepressionScore : int  4 1 18 1 7 14 1 2 12 6 ...
 $ AnxietyScore    : int  3 0 18 4 25 8 0 2 16 11 ...
 $ StressScore     : int  8 3 9 6 14 28 1 3 20 31 ...
 $ DepressionStatus: Factor w/ 3 levels "moderate","normal",..: 2 2 1 2 2 1 2 2 1 2 ...
 $ AnxietyStatus   : Factor w/ 3 levels "moderate","normal",..: 2 2 3 2 3 1 2 2 3 1 ...
 $ Stress          : Factor w/ 2 levels "high","normal": 2 2 2 2 2 1 2 2 1 1 ...
 $ DASScore        : int  15 4 45 11 46 50 2 7 48 48 ...
 $ Happiness       : int  28 25 17 32 15 22 25 29 29 30 ...
 $ AlcoholUse      : Factor w/ 4 levels "Abstain","Heavy",..: 4 4 3 3 4 1 4 3 3 4 ...
 $ Drinks          : int  10 6 3 2 4 0 6 3 3 6 ...
 $ WeekdayBed      : num  25.8 25.7 27.4 23.5 25.9 ...
 $ WeekdayRise     : num  8.7 8.2 6.55 7.17 8.67 8.95 8.48 9.07 8.75 8 ...
 $ WeekdaySleep    : num  7.7 6.8 3 6.77 6.09 9.05 7.73 9.02 8.25 6.6 ...
 $ WeekendBed      : num  25.8 26 28 27 23.8 ...
 $ WeekendRise     : num  9.5 10 12.6 8 9.5 ...
 $ WeekendSleep    : num  5.88 7.25 10.09 7.25 7 ...
 $ AverageSleep    : num  7.18 6.93 5.02 6.9 6.35 9.04 7.52 9.01 8.54 6.68 ...
 $ AllNighter      : int  0 0 0 0 0 0 1 0 0 0 ...
```

- Diagrama de dispersão:

```
plot(x=dados$CognitionZscore,y=dados$GPA, xlab="Pontuação padronizada de um teste de habil
     ylab="Média das notas",main="",pch=20,cex.axis=0.5,cex.lab=0.5)
```



- Para as variáveis `GPA` e `dados$CognitionZscore` nós temos que a correlação é

```
round(cor(dados$GPA,dados$CognitionZscore),3)
```

```
[1] 0.267
```

- Ou seja, existe uma correlação linear positiva entre essas variáveis.

```r
ajuste <- lm(GPA ~ CognitionZscore, data=SleepStudy)
ajuste
```

```

Call:
lm(formula = GPA ~ CognitionZscore, data = SleepStudy)

Coefficients:
    (Intercept)  CognitionZscore
         3.2438           0.1526
```

```r
summary(ajuste)
```

```
Call:
lm(formula = GPA ~ CognitionZscore, data = SleepStudy)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.3171 -0.2377  0.0472  0.2752  0.8340 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      3.24380    0.02454 132.160  < 2e-16 ***
CognitionZscore  0.15261    0.03479   4.386  1.7e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3904 on 251 degrees of freedom
Multiple R-squared:  0.07119,	Adjusted R-squared:  0.06749 
F-statistic: 19.24 on 1 and 251 DF,  p-value: 1.698e-05
```

```
ajuste$coef[1]
```

```
(Intercept)
     3.2438
```

```
ajuste$coef[2]
```

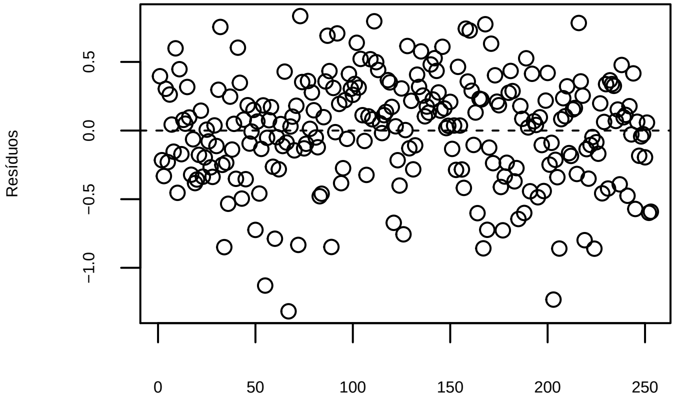
```
CognitionZscore
      0.1526142
```

- Então, o modelo é dado por

  `GPA` $= 3.2438005 + 0.1526142^{*}$`CognitionZscore`

- Verificação das suposições do modelo

```
plot(ajuste$residuals,xlab="",ylab="Resíduos",main="",cex.axis=0.5,cex.lab=0.5)
abline(h=0,lty=2)
```



```r
plot(x=dados$CognitionZscore,y=dados$GPA,
     xlab="Pontuação padronizada de um teste de habilidades cognitivas",
     ylab="Média das notas",main="",pch=20,cex.axis=0.5,cex.lab=0.5, col="red")
abline(a = ajuste$coef[1], b = ajuste$coef[2], col = "black", lwd=2)
```