REGRESSÃO LINEAR

- **Ex. Crescimento da crista em capões recebendo doses diferentes de androsterona (hormona sexual secundária masc.)**

Dose (mg. androst.)	1/2	1	2	4	8
Log_2 dose (x)	-1	0	1	2	3
	8	5	13	17	17
	1	6	7	14	17
Crescimento da crista (Y)	1	9	12	14	20
	3	7	10	19	18
	1	4	11	13	15

- **Em R:**
  ```
  > x=c(-1,-1,-1,-1,-1,0,0,0,0,0,1,1,1,1,1,2,2,2,2,2,
  3,3,3,3,3)
  > y=c(8,1,1,3,1,5,6,9,7,4,13,7,12,10,11,17,14,14,19,
  13,17,17,20,18,15)
  > plot(x,y)
  ```

- **Há uma relação aproximadamente linear. Obviamente o crescimento é a var. dep. e a dose ind.**

- **É claro que para um valor fixo de x, o valor de y varia consideravelmente de ave para ave; pode por isso ser considerada uma var. aleatória com média,**

$E(Y|x)$, variância, $V(Y|x)$, ...

- A função $f(x) = E(Y|x)$ é a regressão de Y em x.
- O nosso propósito consiste em fazer inferências sobre esta função.
- Regressão linear: assumir que $E(Y|x) = \alpha + \beta x$. α e β minimizam $E\{(Y - (\alpha + \beta x))^2\}$
- Agora temos de estimar α e β. Vamos usar os estimadores dos mínimos quadrados:

MÉTODO DOS MíNIMOS QUADRADOS

- **Suponhamos que temos n pares de obs. $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$.**
- **Temos de estimar α e β.**
- **Vamos usar os valores de α e β que minimizam:**
- $S^2(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$
- **A solução é (porquê ?):**

- $$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x};$$

- $$\hat{\beta} = \frac{\sum x_i y_i - n\overline{x}.\overline{y}}{\sum x_i^2 - n\overline{x}^2} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{SS_{xy}}{SS_{xx}};$$

- e $$S^2(\hat{\alpha}, \hat{\beta}) = \sum (y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \sum (y_i - \overline{y})^2 - \hat{\beta}^2 \sum (x_i - \overline{x})^2;$$

Isto é : $$S^2(\hat{\alpha}, \hat{\beta}) = SS_{yy} - \hat{\beta} SS_{xy}$$

JUSTIFICAÇÃO DO MÉTODO DOS MíNIMOS QUADRADOS

- **Sup. que para x fixo, Y tem dist. normal com variância σ^2 (que não depende de x). O modelo é: $Y_i = \alpha + \beta x_i + \epsilon_i$ com $\epsilon_i \sim N(0, \sigma^2)$**

- **Teorema: De todos os est. centrados, $\hat{\alpha}$ e $\hat{\beta}$ do MMQ são os que têm menor variância. (dem em Dudewicz p. 683)**

- **Teorema: Os EMV de α e β são também o $\hat{\alpha}$ e $\hat{\beta}$ anteriores. O EMV de σ^2 é $S^2(\hat{\alpha}, \hat{\beta})/n$.**

- **Estes estimadores são conjuntamente suficientes.**

- **Exercicio. A altura, x, e o peso, y, de 20 alunos forneceu: $\overline{x} = 1.74m, s_x^2 = 0.0045m^2, \overline{y} = 69.6Kg, s_y^2 = 47.31Kg^2$ e $\sum x_i y_i = 2429.585$.**

- **Determine a recta de regressão do peso na altura. Estime o peso de um aluno com 1.78 m de altura.**

- **Seria talvez mais útil obter um intervalo de valores para o peso do aluno.**

INFERÊNCIA EM REGRESSÃO

- $\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$ e $\hat{\beta} = \frac{SS_{xy}}{SS_{xx}}$ são combinações lineares dos $y_i's$; logo:

- $$\hat{\alpha} \sim N(\alpha, \frac{\sum x_i^2}{n\sum(x_i - \overline{x})^2}\sigma^2)$$

- $$\hat{\beta} \sim N(\beta, \frac{\sigma^2}{\sum(x_i - \overline{x})^2})$$

- Mas σ^2 é desconhecido; no entanto,

$$\frac{S^2(\hat{\alpha}, \hat{\beta})}{\sigma^2} \sim \chi^2_{n-2}, \quad \text{e ind. de } \hat{\alpha}, \hat{\beta}$$

- **logo,** $\hat{\sigma}^2 = \frac{S^2(\hat{\alpha},\hat{\beta})}{n-2}$ **é um estimador centrado de** σ^2**.**
- **Vem então:**
-

$$\frac{(\hat{\alpha} - \alpha)}{\hat{\sigma}} \sqrt{\frac{n \sum (x_i - \overline{x})^2}{\sum x_i^2}} \sim t_{n-2} \text{ (porquê?)}$$

- **Análogamente**

$$\frac{(\hat{\beta}-\beta)}{\hat{\sigma}}\sqrt{\sum(x_i-\overline{x})^2} \sim t_{n-2} \text{ (porquê?)}$$

- **Podemos assim fazer inferência sobre estes parâmetros.**
- **Se os erros não tiverem dist. normal, as dist. acima já não são válidas, MAS continuam estimadores centrados com a mesma variância.**
- **Se a variância de Y depender de x, $\hat{\alpha}$ e $\hat{\beta}$ continuam a ser centrados e com dist. normal (se Y a tiver), mas as variâncias mudam**

- **Seja agora $H_0 : \beta = 0$. O que significa não rejeitar H_0?**

  1. **Não existe rel. entre X e Y (são ind.)**
  2. **A relação existe e é linear mas o declive é muito próximo de 0.**
  3. **A relação não é linear, mas curvílinea.**
  4. **Cometeu-se um erro tipo II (aceitar H_0 quando H_0 é falsa)**

- **O que significa rejeitar H_0?**

  1. **A relação é linear o suficiente.**
  2. **Um modelo linear aproxima bem os dados embora um de grau superior fosse melhor.**

3. **Cometeu-se um erro tipo I (rejeitar H_0 quando H_0 é verdadeira)**

- **Exercicio: teste a hipótese de $\beta = 0$ com os dados do exemplo anterior.**

INTERVALOS DE CONFIANÇA E DE PREDIÇÃO

- **Suponhamos agora que queremos fazer inferências não sobre os parâmetros mas sobre o valor de Y a prever no futuro. Sup. que queremos prever o valor de Y em x^***

- **$\hat{\alpha} + \hat{\beta}x^*$ é um estimador de $\alpha + \beta x^*$, resposta média em x^* ou $E(Y|x^*)$.**

- **E se quisermos estimar o valor de $Y|x^*$, e não apenas o seu valor médio $E(Y|x^*)$?**

- **O único estimador disponível é $\hat{y} = \hat{\alpha} + \hat{\beta}x^*$**

- **Portanto, $\hat{\alpha} + \hat{\beta}x^*$ é um estimador de duas coisas diferentes:**

1. **de $\alpha + \beta x^*$, resposta média em x^*.**
2. **de Y em x^*.**

- **Suponhamos que queremos um IC para o valor médio, $\mu_{Y|x^*} = \alpha + \beta x^*$.**

- **Vamos usar a quantidade fulcral**

$$\frac{(\hat{\alpha} + \hat{\beta}x^*) - (\alpha + \beta x^*)}{\sqrt{\mathrm{Var}((\hat{\alpha} + \hat{\beta}x^*) - (\alpha + \beta x^*))}}$$

- **Mas** $\mathrm{Var}((\hat{\alpha}+\hat{\beta}x^*)-(\alpha+\beta x^*)) = \mathrm{Var}(\hat{\alpha}+\hat{\beta}x^*) = \mathrm{Var}(\overline{Y}+\hat{\beta}(x^*-\overline{x})) = \mathrm{Var}(\overline{Y})+(x^*-\overline{x})^2\mathrm{Var}(\hat{\beta})+2(x^*-\overline{x})\mathrm{Cov}(\overline{Y},\hat{\beta})$

- **Prove que** $\mathrm{Cov}(\overline{Y},\hat{\beta}) = 0$**; logo** $\overline{Y}$ **e** $\hat{\beta}$ **são estatísticamente independentes. Segue que** $\hat{\alpha}+\hat{\beta}x^*$ **tem uma dist. normal (porquê?)**

- **Vem então que**

$$\mathrm{Var}(\hat{\alpha}+\hat{\beta}x^*) = \frac{\sigma^2}{n} + (x^*-\overline{x})^2\frac{\sigma^2}{\sum(x_i-\overline{x})^2}$$

- **Logo,**

$$\frac{(\hat{\alpha}+\hat{\beta}x^*)-(\alpha+\beta x^*)}{\sigma\sqrt{\frac{1}{n}+\frac{(x^*-\overline{x})^2}{\sum(x_i-\overline{x})^2}}}\sim N(0,1)$$

- **Mas σ é desconhecido; no entanto sabemos que $\hat{\sigma}^2=\frac{S^2(\hat{\alpha},\hat{\beta})}{n-2}$ e que $\frac{S^2(\hat{\alpha},\hat{\beta})}{\sigma^2}\sim\chi^2_{n-2}$. Logo vem que**

$$\frac{(\hat{\alpha}+\hat{\beta}x^*)-(\alpha+\beta x^*)}{\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x^*-\overline{x})^2}{\sum(x_i-\overline{x})^2}}}\sim t_{n-2}$$

- **Finalmente o IC a $100(1-\alpha)\%$ para o valor médio**

de Y em x^* é:

$$(\hat{\alpha} + \hat{\beta}x^*) \pm t_{n-2,1-\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum (x_i - \overline{x})^2} \right)}$$

- **Fazendo variar x^* obtemos uma banda de confiança para a recta de regressão.**
- **Suponhamos agora que queremos um intervalo de predição para Y em x^*. Usamos a quantidade fulcral**

$$\frac{(\hat{\alpha} + \hat{\beta}x^*) - Y}{\sqrt{\mathrm{Var}((\hat{\alpha} + \hat{\beta}x^*) - Y)}}$$

- **Mas** $\mathrm{Var}((\hat{\alpha}+\hat{\beta}x^*)-Y) = \mathrm{Var}(\overline{Y}+\hat{\beta}(x^*-\overline{x})-Y) = \mathrm{Var}(\overline{Y}+\hat{\beta}(x^*-\overline{x})) + \mathrm{Var}(Y)$ **(porquê?)** $= \frac{\sigma^2}{n} + (x^*-\overline{x})^2 \frac{\sigma^2}{\sum(x_i-\overline{x})^2} + \sigma^2$

- **Procedendo de forma análoga obtém-se um IP para** Y **em** x^***:**

$$(\hat{\alpha}+\hat{\beta}x^*) \pm t_{n-2,1-\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x^*-\overline{x})^2}{\sum(x_i-\overline{x})^2}\right)}$$

- **Fazendo variar** x^* **obtemos uma banda de predição para** Y**.**

- **Exercicio:**
  a)Encontre um ICa 95% para o peso médio de um estudante cuja altura é 1.78m.
  b) Encontre um IP para o peso de um estudante cuja altura é 1.78m.

- **Exercicio:**
  Considere a amostra aleatória $Y_1, \ldots, Y_{12}$ para valores fixos de x:

x_i	8	6	11	22	14	17	18	24	19	23	26	40
Y_i	59	58	56	53	50	45	43	42	39	38	30	27

- $\sum x_i = 228, \quad \sum x_i^2 = 5256, \quad \sum x_i Y_i =$

$\mathbf{9324}$, $\sum Y_i = 540$, $\sum Y_i^2 = 25522$.

- **Encontre $\hat{\alpha}$ e $\hat{\beta}$. Se $x^* = 20$ encontre o valor previsto para Y e um IP a 95% para Y. Desenhe a recta de regressão com as bandas de confiança e de predição (a 95%).**

Em R:

```
> x=c(8,6,11,12,14,17,18,24,19,23,26,40)
> y=c(59,58,56,53,50,45,43,42,39,38,30,27)
> plot(x,y)
> modelo=lm(y ~ x)
> modelo
> confint(modelo)
> confint(modelo,level=0.99)
> predict(modelo,data.frame(x=20))
> predict(modelo,data.frame(x=20),int="c")
```

ou então:

```
> predict(modelo,data.frame(x=20),int="confidence")
> predict(modelo,data.frame(x=20),int="p")
```

ou então

```
> predict(modelo,data.frame(x=20),int="prediction")
> novo=data.frame(x=seq(7,40,1))
> bandas.prev= predict(modelo, novo, interval="prediction")
> bandas.conf= predict(modelo, novo, interval="c")
> matplot(novo$x,cbind(bandas.conf, bandas.prev[,-1]),
+ lty=c(1,2,2,3,3), type="l",xlab="x", ylab="y")
> points(x,y)
```

COEFICIENTE DE DETERMINAÇÃO

- **Fornece um critério de avaliação da qualidade do ajustamento. Considere a soma dos quadrados dos resíduos** $S^2(\hat{\alpha}, \hat{\beta}) = \sum (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$.

- $0 \leq S^2(\hat{\alpha}, \hat{\beta}) \leq \; ?$

- **Uma forma de avaliar a utilidade da recta é comparando a dispersão dos pontos em torno dela com a dispersão dos pontos em torno do modelo mais simples** $Y = \overline{Y}$.

- **Parece óbvio que a dispersão em torno de $\overline{Y}$ é muito maior do que em torno da recta de regressão.**

- $SS_{yy} = \sum(y_i - \overline{y})^2 = \sum y_i^2 - (\sum y_i)^2/n$: **soma de quadrados total**

- $SSR = \sum(\hat{y}_i - \overline{y})^2 = \hat{\beta}^2[\sum x_i^2 - (\sum x_i)^2/n] = \hat{\beta}^2 SS_{xx}$: **soma de quad. explicada pela regressão**

- $SSE = \sum(y_i - \hat{y}_i)^2 = SS_{yy} - \hat{\beta} SS_{xy} = SS_{yy} - SSR$: **soma de quad. não explicada pela regressão.**

- $SS_{yy} = SSR + SSE$.

- **Se a recta de regressão for adequada, então é de esperar que SSR seja bem superior a SSE.**

- **Uma forma de o determinar é através do <u>coeficiente de determinação</u>:**

- $$\frac{SSR}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}} = \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2} = r^2$$

- **r^2 (quadrado do coef. de correlação) é um estimador de ρ^2, coef. de determinação na população.**

REGRESSÃO POLINOMIAL E REGRESSÃO MÚLTIPLA

- **Generalização:**
- **Regressão Quadrática:** $Y_i = \alpha + \beta x_i + \gamma x_i^2 + \epsilon_i$
- **Podemos assim testar se o modelo linear é adequado testando** $H_0 : \gamma = 0$
- **Regressão Polinomial:**
  $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_p x_i^p + \epsilon_i$
- **São tudo casos particulares da**
- **Regressão Linear Múltipla:**
  $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \epsilon_i$

- **As variáveis $x_1, x_2, \ldots, x_p$ podem ser quantitativas; codificação binária de var. qualitativas; potências de outras variáveis; produto de 2 ou mais variáveis....**

- **É por isso um modelo muito genérico que permite construir funções lineares e não lineares com uma ou mais var. independentes.**