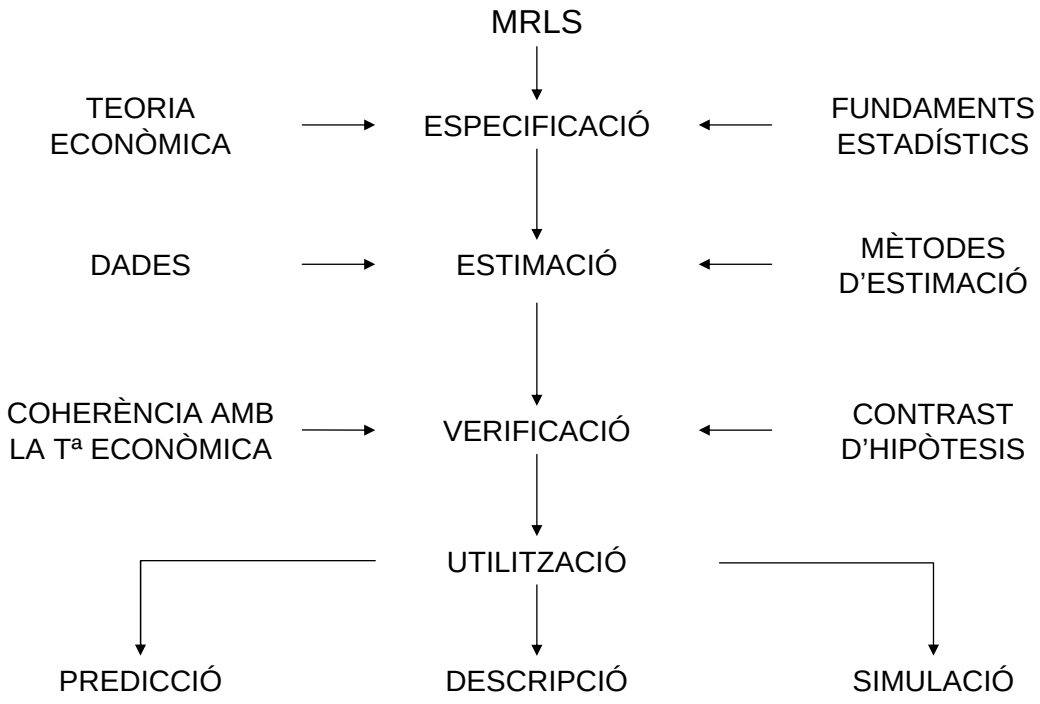
Tema 5. Model de Regressió Lineal Simple

5.1. Objecte i natura del model



Tema 5. Model de Regressió Lineal Simple

5.1. Objecte i natura del model

Exemples:

a) Funció de Consum: C=a+bR

Relació lineal entre el consum i la renda disponible. On 0<b<1 representa la proporció marginal a consumir.

b) Funció de demanda: Q=a+bP

Relació lineal entre les vendes i el preu de venda. On b<0

Quan observem els consums (vendes) de diferents famílies (empreses) amb la mateixa renda (preus) no tos presenten els mateixos consums (vendes).

És a dir, les relacions econòmiques (empresarials) NO SÓN DETERMINISTES. Existeixen altres aspectes que influeixen sobre les variables observades (gustos, mida familiar, mida empresarial, territori, sector productiu, situació econòmica general, etc..)

Tots aquests aspectes els hem de recollir a algun lloc. Per això, incorporem un TERME DE PERTORBACIÓ (u o ε) amb el que procurem recollir tots aquest aspectes i sobre el que farem alguns supòsits. Així:

$$\begin{cases} C = a + bR + U \\ Q = a + bP + U \end{cases}$$

Amb aquesta operació, convertim les relacions teòriques, DETERMINISTES, en relacions ESTOCÀSTIQUES.

Tema 5. Model de Regressió Lineal Simple

5.1. Objecte i natura del model

En termes generals:

$$\begin{array}{lll} Y = a + bX & \Rightarrow & Y = a + bX + \varepsilon \\ EEiE\_I & & EEiE\_II \\ Y_i = \alpha + \beta X_i + \varepsilon_i & \Rightarrow & Y_i = \beta_0 + \beta_1 X_{1i} + \cdots \beta_k X_{ki} + \varepsilon_i \\ EEiE\_II & & ECONOMETRIA \end{array}$$

L'equació es composa de:

1. VARIABLE DEPENENT / ENDÒGENA / EXPLICADA (*Y*). Variable que desitgem explicar.
2. VARIABLE INDEPENDENT / EXÒGENA /EXPLICATIVA (*X*). Variable que suposem explica el comportament d'*Y.*
3. PARÀMETRES α i β. Constant i pendent de la RECTA DE REGRESSIÓ LINEAL.
4. TERME DE PERTORBACIÓ (ε). Recull altres aspectes NO PRESENTS en el conjunt de variables explicatives, així com els errors de mesura associats a la(s) variable(s) explicativa(es).

Tema 5. Model de Regressió Lineal Simple

5.1. Objecte i natura del model

LINEALITAT DE LA FUNCIÓ

1. Molts models són lineals
2. Altres són linealitzables:

$$Q = A * K^{\alpha} L^{\beta}$$
$$\ln(Q) = \ln(A) + \alpha \ln(K) + \beta \ln(L)$$

3. Quan s desconeix la forma funcional un primera aproximació es realitza sota aquest supòsit.

5.2. Hipòtesis per a l'especificació del model

S'ha de realitzar una sèrie de hipòtesis a priori sobre l'ESTRUCTURA DEL MODEL. La validesa d'aquestes hipòtesis garantirà la INFERÈNCIA que fem amb posterioritat a partir del model ESTIMAT.

1. No existeixen error sistemàtics. En promig el valor esperat de les pertorbacions és nul.

$$\longrightarrow E(\varepsilon_i) = 0 \quad \forall i$$

2. La dispersió de les pertorbacions és CONSTANT (HOMOCEDASTICITAT)

$$\longrightarrow \text{var}(\varepsilon_i) = E(\varepsilon_i - E(\varepsilon_i))^2 = E(\varepsilon_i)^2 = \sigma^2 \quad \forall i$$

Tema 5. Model de Regressió Lineal Simple

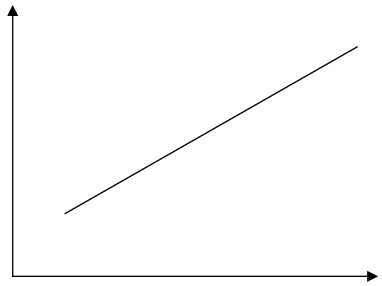
5.1. Hipòtesis per a l'especificació del model

3. No existeixen variables rellevants OMESES i per això les pertorbacions no estan AUTOCORRELACIONADES $\longrightarrow cor(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i * \varepsilon_j) = 0 \quad \forall i \neq j$

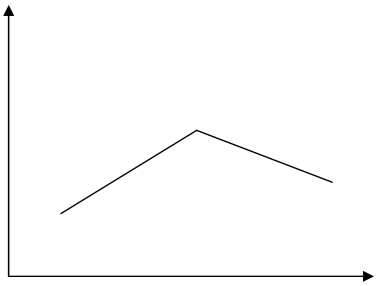
4. Normalitat del terme de pertorbació $\longrightarrow \varepsilon_i \approx N(0, \sigma) \quad \forall i$

5. X és DETERMINISTA (NO ALEATÒRIA) és fixada per l'investigador o de forma EXÒGENA al model.

6. Existeix permanència ESTRUCTURAL, només existeix una recta que representa el COMPORTAMENT DE LA RELACIÓ



SI



NO

Tema 5. Model de Regressió Lineal Simple

5.1. Hipòtesis per a l'especificació del model

Un cop ASUMITS AQUESTS SUPÒSITS, es procedeix a caracteritzar el comportament PROBABILÍSTIC de la VARIABLE ALEATÒRIA ENDÒGENA.

$$E(Y_i) = E(\alpha + \beta X_i + \varepsilon_i) = E(\alpha) + E(\beta X_i) + E(\varepsilon_i) = \underbrace{\alpha + \beta X_i}_{\text{Recta de Regressió}} + \underbrace{0}_{\text{hipòtesi del model}}$$

$$V(Y_i) = E(Y_i - E(Y_i))^2 = E((\alpha + \beta X_i + \varepsilon_i) - (\alpha + \beta X_i))^2 = E(\varepsilon_i)^2 \underbrace{=}_{\text{hipòtesi del model}} \sigma^2$$

Quina forma té la distribució d'Y?

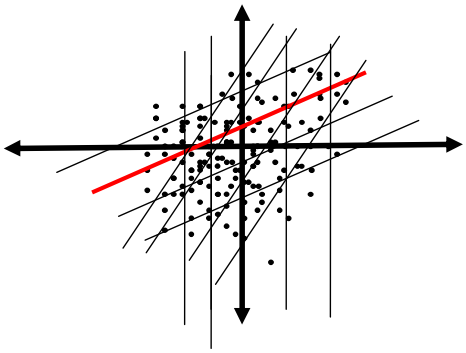
És una combinació lineal on l'única variable que presenta aleatorietat és el terme de pertorbació. Atès que aquest és NORMAL, per les propietats de la normal, la variable Y es distribuirà normalment. Així:

$$Y_i \approx N(\alpha + \beta X_i, \sigma)$$

ECET 5.1

Tema 5. Model de Regressió Lineal Simple

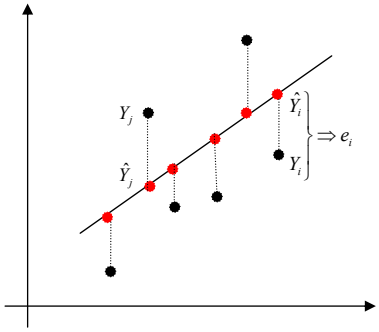
5.3. Estimació dels paràmetres de la recta de regressió per Mínims Quadrats Ordinaris



Per a concretar quina de totes les possibles normals hem d'estimar els paràmetres α, β i σ^2.

De quina informació disposem? → mostra (X i Y) de "n" individus.

Existeixen diferents mètodes d'estimació. Hem vist, durant aquest curs dos: MM, MV. Ara presentem el mètode de MÍNIMS QUADRATS ORDINARIS.



IDEA: La millor recta de regressió serà aquella que minimitzi els errors que cometem.

Tema 5. Model de Regressió Lineal Simple

5.3. Estimació dels paràmetres de la recta de regressió per Mínims Quadrats Ordinaris

CRITERI: Minimització de la suma dels errors al quadrat (no m'importa si m'equivoco per sobre o per baix, m'importa el segment, no el seu signe)

$$Min \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\alpha} - \hat{\beta} X_i\right)^2$$

$$\begin{cases} \dfrac{\partial \sum_{i=1}^{n} e_i^2}{\partial \hat{\alpha}} = -2 \sum_{i=1}^{n} \left(Y_i - \hat{\alpha} - \hat{\beta} X_i\right) = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} Y_i - n\hat{\alpha} - \hat{\beta} \sum_{i=1}^{n} X_i = 0 \\ \dfrac{\partial \sum_{i=1}^{n} e_i^2}{\partial \hat{\beta}} - 2 \sum_{i=1}^{n} X_i \left(Y_i - \hat{\alpha} - \hat{\beta} X_i\right) = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} X_i Y_i - \hat{\alpha} \sum_{i=1}^{n} X_i - \hat{\beta} \sum_{i=1}^{n} X_i^2 = 0 \end{cases}$$

$$\begin{cases} \overline{Y} = \hat{\alpha} + \hat{\beta} \overline{X} \\ \sum_{i=1}^{n} X_i Y_i = \left(\overline{Y} - \hat{\beta} \overline{X}\right) \sum_{i=1}^{n} X_i + \hat{\beta} \sum_{i=1}^{n} X_i^2 \end{cases} \Rightarrow \hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i - n \overline{X} \overline{Y}}{\sum_{i=1}^{n} X_i^2 - n \overline{X}^2} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} = \frac{S_{XY}}{S_X^2}$$

Tema 5. Model de Regressió Lineal Simple

5.3. Estimació dels paràmetres de la recta de regressió per Mínims Quadrats Ordinaris

CARACTERÍSTIQUES DE LA REGRESSIÓ:

1. Sempre passa pel CENTRE del núvol de punts.

2. $\sum e_i = 0 \quad i \quad E(X_i e_i) = 0$

3. Atès que l'estimador de β, depèn de la covariància, el seu signe depèn d'aquesta. També pot expressar-se com: $\longrightarrow \quad \hat{\beta} = r_{XY} \frac{S_Y}{S_X}$

4. L'ERROR D'ESTIMACIÓ no és el mateix que el TERME DE PERTORBACIÓ

NIVELL POBLACIONAL $\longrightarrow \quad Y_i = \alpha + \beta X_i + \varepsilon_i = E(Y_i) + \varepsilon_i$

NIVELL MOSTRAL $\longrightarrow \quad Y_i = \hat{\alpha} + \hat{\beta} X + e_i = \hat{Y}_i + e_i$

ECET 5.2

Tema 5. Model de Regressió Lineal Simple

5.3. Estimació dels paràmetres de la recta de regressió per Mínims Quadrats Ordinaris

PROPIETATS DELS ESTIMADORS MQO:

Si es compleixen les hipòtesis bàsiques del model

1. LINEALITAT: Són funcions lineals de les variables.
2. NO ESBIAIXATS $E(\hat{\alpha}) = \alpha \quad i \quad E(\hat{\beta}) = \beta$
3. Variància mínima:

$V(\hat{\alpha}_*) \geq V(\hat{\alpha}_{MQO})$

$V(\hat{\beta}_*) \geq V(\hat{\beta}_{MQO})$

$$V(\hat{\beta}) = V\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sum_{i=1}^{n}\left((X_i - \bar{X})^2\right)^2} V\left((Y_i - \bar{Y})\right)$$

$$= \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$V(\hat{\alpha}) = V\left(\bar{Y} - \hat{\beta}\bar{X}\right) =$$

$$= V(\bar{Y}) + \bar{X}^2 V(\hat{\beta}) - 2\bar{X}\,\mathrm{cov}(\hat{\beta}, \bar{X}) =$$

$$= \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right]$$

Tema 5. Model de Regressió Lineal Simple

5.3. Estimació dels paràmetres de la recta de regressió per Mínims Quadrats Ordinaris

4. Distribució Normal.

$$\hat{\alpha} \approx N\left(\alpha, \sqrt{\sigma^2\left(\frac{1}{n}+\frac{\overline{X}^2}{\sum\left(X_i-\overline{X}\right)^2}\right)}\right) \quad i \quad \hat{\beta} \approx N\left(\beta, \sqrt{\frac{\sigma^2}{\sum\left(X_i-\overline{X}\right)^2}}\right)$$

Ara bé, per a poder fer inferència sobre els dos paràmetres poblacionals α i β, necessitem proposar una estimació de σ^2:

$$\sigma^2 = E\left(\varepsilon_i\right)^2 \quad \Rightarrow \quad \text{es proposa com estimador} \quad \hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$$

A nivell empíric, el sumatori dels errors es pot descomposar de la següent manera:

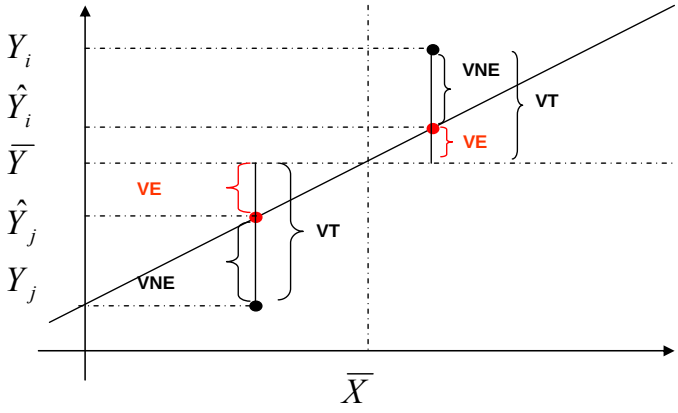
$$\begin{aligned}\sum e_i^2 &= \sum\left(Y_i-\hat{Y}_i\right)^2 = \sum\left(Y_i-\hat{\alpha}-\hat{\beta}X_i\right)^2 = \sum\left(Y_i-\left(\overline{Y}-\hat{\beta}\overline{X}\right)-\hat{\beta}X_i\right)^2 = \\ &= \sum\left(\left(Y_i-\overline{Y}\right)-\hat{\beta}\left(X_i-\overline{X}\right)\right)^2 = \sum\left(Y_i-\overline{Y}\right)^2+\hat{\beta}^2\left(X_i-\overline{X}\right)^2-2\hat{\beta}\sum\left(Y_i-\overline{Y}\right)\left(X_i-\overline{X}\right) = \\ &= \sum\left(Y_i-\overline{Y}\right)^2+\hat{\beta}^2\left(X_i-\overline{X}\right)^2-2\hat{\beta}\hat{\beta}\sum\left(X_i-\overline{X}\right)^2 = \\ &= \sum\left(Y_i-\overline{Y}\right)^2-\hat{\beta}^2\left(X_i-\overline{X}\right)^2 = nS_Y^2-\hat{\beta}^2 nS_X^2 = n\left(S_Y^2-\hat{\beta}^2 S_X^2\right)\end{aligned}$$

ECET 5.3

Tema 5. Model de Regressió Lineal Simple

5.4/5.5 Contrastació del model i bondat de l'ajust

La BONDAT DE L'AJUST valora fins a quin punt La recta estimada s'ajusta al núvol de punts. Per Això, valora quina part de la VARIABILITAT TOTAL de la VARIABLE ENDÒGENA és explicada per la RECTA "AJUSTADA" DE REGRESSIÓ



Tenint present el centre del núvol de punts, La distància entre la observació real de la variable endògena i la seva mitjana (Variació Total- VT), es pot descomposar en dos segments:

- Variació Explicada (**VE**). La distància que existeix entre l'ajust obtingut i la mitjana de la variable endògena.
- Variació NO Explicada (**VNE**). La distància entre l'observació real i la mitjana de la variable endògena.

$$Y_i-\overline{Y} = \left(Y_i-\hat{Y}_i\right)+\left(\hat{Y}_i-\overline{Y}\right) = \quad \text{si elevem al quadrat i sumem tots els segments}$$

$$\underbrace{\sum\left(Y_i-\overline{Y}\right)^2}_{SQT} = \underbrace{\sum\left(Y_i-\hat{Y}_i\right)^2}_{SQE} + \underbrace{\sum\left(\hat{Y}_i-\overline{Y}\right)^2}_{SQR} + \underbrace{2\sum\left(Y_i-\hat{Y}_i\right)\sum\left(\hat{Y}_i-\overline{Y}\right)}_{0}$$

SQT: Suma Quadrats total
SQE: Suma Quadrats Errors
SQR: Suma Quadrats Regressió
Atès que $\sum e_i = 0$,

Tema 5. Model de Regressió Lineal Simple

5.4/5.5 Contrastació del model i bondat de l'ajust

El COEFICIENT DE DETERMINACIÓ és la mesura que ens ajudarà a valorar la BONDAT DE L'AJUST

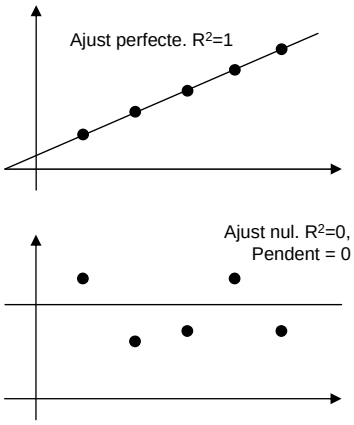
$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT} = \frac{\sum\left(\hat{Y}_i - \overline{Y}\right)^2}{\sum\left(Y_i - \overline{Y}\right)^2} = 1 - \frac{\sum\left(Y_i - \hat{Y}_i\right)^2}{\sum\left(Y_i - \overline{Y}\right)^2}$$

$$R^2 = r_{XY}^2$$

•Si R^2=1, els errors són nuls i totes les observacions mostrals es troben a sobre de la recta de regressió. SQE=0.

• Si R^2=0, tota la variació de la variable endògena se explica pels errors i, aleshores, la recta de regressió no explicar res de la variable endògena. SQR=0. (El coeficient de correlació entre les variables és 0)

• Per tant $0<R^2<1$. Quan més a prop estigui d'1 millor eerà l'ajust i quan més a prop estigui de 0, serà pitjor.



Tema 5. Model de Regressió Lineal Simple

5.4/5.5 Contrastació del model i bondat de l'ajust

CONTRASTACIÓ DEL MODEL: INTERVAL DE CONFIANÇA

Sota les HIPÒTESIS DEL MODEL. L'interval de confiança per a l'estimador del **paràmetre β** és:

$$\hat{\beta} \approx N\left(\beta, \underbrace{\sqrt{\frac{\sigma^2}{\sum\left(X_i - \overline{X}\right)^2}}}_{\sqrt{\sigma_{\hat{\beta}}^2}}\right).$$ Atès que $\Rightarrow$ σ^2 es desconeguda $\Rightarrow$ $\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$

per tant, $\sigma_{\hat{\beta}} = \sqrt{\frac{\sum e_i^2 / n-2}{\sum\left(X_i - \overline{X}\right)^2}}$ i en aquest cas sabem que la distribució segueix una t-student.

Així, l'interval de confiança es construeix com: $P\left[\hat{\beta} - t_{\alpha/2,n-2}\hat{\sigma}_{\hat{\beta}} \le \beta \le \hat{\beta} + t_{\alpha/2,n-2}\hat{\sigma}_{\hat{\beta}}\right] = 1-\alpha$

I en el cas del **paràmetre α** és:

$$\hat{\alpha} \approx N\left(\alpha, \underbrace{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum\left(X_i - \overline{X}\right)^2}\right)}}_{\hat{\sigma}_{\hat{\alpha}}}\right)$$ Per tant, $P\left[\hat{\alpha} - t_{\alpha/2,n-2}\hat{\sigma}_{\hat{\alpha}} \le \alpha \le \hat{\alpha} + t_{\alpha/2,n-2}\hat{\sigma}_{\hat{\alpha}}\right] = 1-\alpha$

Tema 5. Model de Regressió Lineal Simple

5.4/5.5 Contrastació del model i bondat de l'ajust

Finalment, l'interval de confiança per a l'estimador del **paràmetre σ^2** és:

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \approx \chi^2_{n-2} \quad \text{per tant, } P\left[\frac{(n-2)\hat{\sigma}^2}{\chi^2_{1-\alpha/2,n-2}} \le \sigma^2 \le \frac{(n-2)\hat{\sigma}^2}{\chi^2_{\alpha/2,n-2}}\right] = 1-\alpha$$

ECET 5.4

CONTRAST D'HIPÒTESI sobre els **paràmetres β i α**

$$\begin{cases} H_0: & \beta = \beta_0 \\ H_1: & \beta \neq \beta_0 \end{cases} \Rightarrow \quad Si \quad \left|\frac{\hat{\beta}-\beta_0}{\hat{\sigma}_{\hat{\beta}}}\right| \ge t_{n-2,\alpha/2} \quad \Rightarrow \quad RH_0$$

$$\begin{cases} H_0: & \alpha = \alpha_0 \\ H_1: & \alpha \neq \alpha_0 \end{cases} \Rightarrow \quad Si \quad \left|\frac{\hat{\alpha}-\alpha_0}{\hat{\sigma}_{\hat{\alpha}}}\right| \ge t_{n-2,\alpha/2} \quad \Rightarrow \quad RH_0$$

I, finalment, sobre la BONDAT DE L'AJUST, és a dir, la significació global del model:

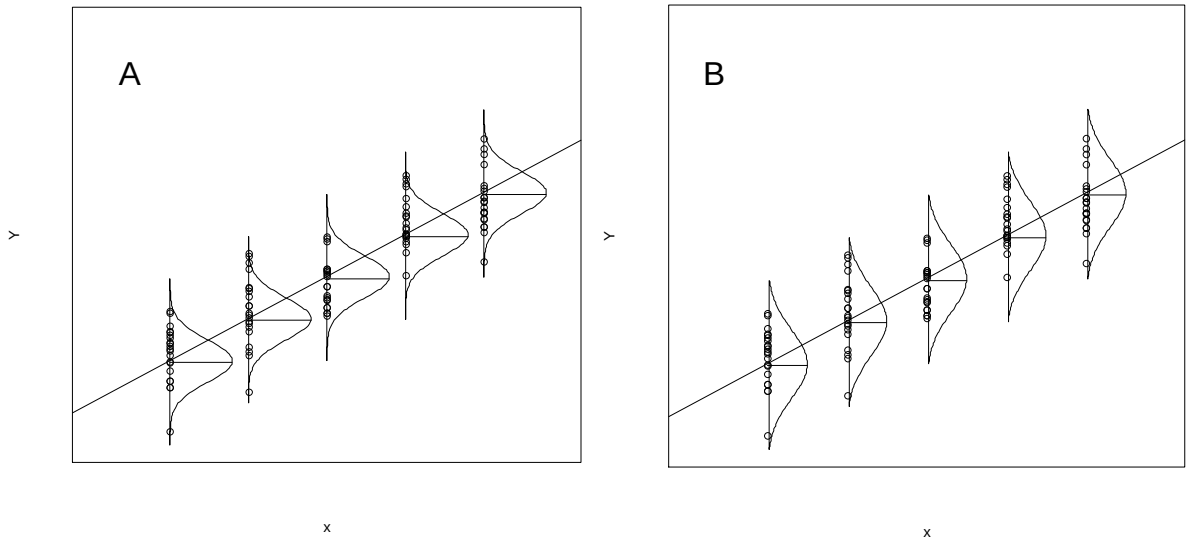
$$\begin{cases} H_0: & R^2 = 0 \\ H_1: & R^2 > 0 \end{cases} \Rightarrow \text{ amb l'estadístic } F^* = \frac{(n-1)R^2}{1-R^2} \Rightarrow \text{ Si } F^* > F_{1,n-2,1-\alpha} \Rightarrow RH_0$$

que en un model de REGRESSIÓ LINEAL SIMPLE és el mateix que fer el contrast sobre el pendent de la recta.

ECET 5.5

Tema 5. Model de Regressió Lineal Simple

ANNEX: Incompliment de les hipòtesis del model: HETEROSCEDASTICITAT



Tant A com B presenten un terme de pertorbació HOMOCEDÀSTIC, malgrat però, a B la variabilitat és major que a A.

HETEROSCEDASTICITAT: Quan aquesta variabilitat no és igual per a cada error "i"

5.6 Predicció puntual i per interval

PREDICCIÓ PUNTUAL:

$$\hat{Y}_{n+1} = \hat{\alpha} + \hat{\beta} X_{n+1}$$

Igual que existeix un ERROR D'ESTIMACIÓ, existeix un ERROR DE PREDICCIÓ.

$$e_{n+1} = Y_{n+1} - \hat{Y}_{n+1} = \left(\alpha + \beta X_{n+1} + \varepsilon_{n+1}\right) - \left(\hat{\alpha} + \hat{\beta} X_{n+1}\right) =$$

$$= \underbrace{\left(\alpha - \hat{\alpha}\right) + \left(\beta - \hat{\beta}\right) X_{n+1}}_{\text{FONT D'ERROR 1}} + \underbrace{\varepsilon_{n+1}}_{\text{FONT D'ERROR 2}}$$

FONT D'ERROR 1: Per l'error comés a l'hora d'estimar els paràmetres

FONT D'ERROR 2: Associat al terme de pertorbació de l'individu n+1

La pregunta ara és, com es distribueix aquest error?. Haurem de caracteritzar, la seva esperança, la seva variància i la forma de la distribució.

$$E\left(e_{n+1}\right) = E\left(\left(\alpha - \hat{\alpha}\right) + \left(\beta - \hat{\beta}\right) X_{n+1} + \varepsilon_{n+1}\right) = 0 + 0 + \underbrace{0}_{\text{Per hipòtesis del model } E(\varepsilon_{n+1})=0} = 0$$

Tema 5. Model de Regressió Lineal Simple

5.6 Predicció puntual i per interval

$$V\left(e_{n+1}\right) = V\left(\left(\alpha - \hat{\alpha}\right) + \left(\beta - \hat{\beta}\right) X_{n+1} + \varepsilon_{n+1}\right) = V\left(\hat{\alpha}\right) + X_{n+1}^2 V\left(\hat{\beta}\right) + V\left(\varepsilon_{n+1}\right) + 2 X_{n+1} \operatorname{cov}\left(\hat{\alpha}, \hat{\beta}\right)$$

$$\left\{ \begin{array}{l} V\left(\hat{\alpha}\right) = \sigma^2 \left(\dfrac{1}{n} + \dfrac{\overline{X}^2}{\sum \left(X_i - \overline{X}\right)^2} \right) \\ V\left(\hat{\beta}\right) = \dfrac{\sigma^2}{\sum \left(X_i - \overline{X}\right)^2} \\ \operatorname{cov}\left(\hat{\alpha}, \hat{\beta}\right) = \dfrac{\sigma^2 \overline{X}}{\sum \left(X_i - \overline{X}\right)^2} \end{array} \right. \Rightarrow \text{ Al final: } V\left(e_{n+1}\right) = \sigma^2 \left(1 + \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{\sum \left(X_i - \overline{X}\right)^2} \right)$$

Quan més lluny estigui X_{n+1} de $\overline{X}$ més gran serà la variància.

Aquesta és mínima si $X_{n+1} = \overline{X}$.

Si $e_{n+1} \approx N\left(0, \underbrace{\sqrt{\sigma^2 \left(1 + \frac{1}{n} + \frac{\left(X_{n+1} - \overline{X}\right)^2}{\sum \left(X_i - \overline{X}\right)} \right)}}_{\hat{\sigma}_{e_{n+1}}}\right) \Rightarrow P\left[\hat{Y}_{n+1} - t_{\alpha/2, n-2} \hat{\sigma}_{e+1} \le Y_{n+1} \le \hat{Y}_{n+1} + t_{\alpha/2, n-2} \hat{\sigma}_{e+1}\right] = 1 - \alpha$

ECET 5.6