Régression Simple.

→ On suppose qu'il existe une relation linéaire entre Y et x.

Modèle : $Y_i = \mu + \alpha x_i + E_i$

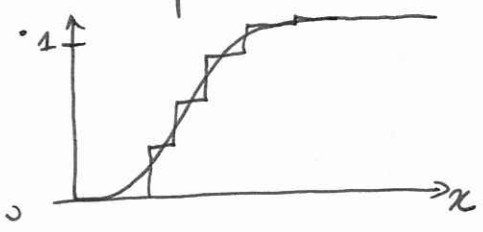
- E_i iid
- $V(E_i) = \sigma^2$: Variance résiduelle. (à minimiser).

→ 1ère chose : Regarder si les hypothèses sont vérifiées

⇒ Graphe des résidus : pour { . homoscedasticité ? . tendance ? . indép ? . }

→ Test d'ajustement des résidus à une Gaussienne $N(0, \sigma^2)$.

Exemple : . Kolmogoroff. Smirnoff : fondé sur les fonctions de répartition



$D_x = \sup_x |F_n(x) - F(x)|$.

= Statistique de test dont la loi ne dépend pas de F sous H_0.

La tendance à rejeter l'hypothèse de normalité.

. Q-Q plot : visualisat°

Conseil : Modèle linéaire robuste à l'hypothèse de normalité.

→ Graphique des résidus

Lecture de la table d'ANOVA.

Source :		dF	SS	MS
Model :	Variabilité expliquée par x.	1	$\sum(\hat{Y}_i - \bar{Y})^2$	$\frac{SCM}{1}$
Error :	Variabilité de Y autour de la droite de reg.	$n-2$	$\sum(Y_i - \hat{Y}_i)^2$	$\frac{SCE}{n-2}$
Total :	Variabilité intrinsèque de Y (⊥ du modèle)	$n-1$	$\sum(Y_i - \bar{Y})^2$	$\frac{SCT}{n-1}$

- **1er test de Fisher :** test du modèle nul contre complet.

$H_0 = \{ Y_i = \mu^0 + \varepsilon_i^0 \}$ ↓ SCR_0, $n-k_0$ dF

$H_1 = \{ Y_i = \mu^1 + \alpha^1 x_i + \varepsilon_i^1 \}$ ↓ SCR_1, $n-k_1$ dF.

$$F = \frac{(SCR_0 - SCR_1)/(k_1 - k_0)}{SCR_1/(n-k_1)} \sim \mathcal{F}(k_1-k_0, n-k_1).$$

↳ revient à faire l'hypothèse sur le paramètre α : $H_0 = \{\alpha = 0\}$.

- **Indicateurs :** . $R^2 = \frac{SCM}{SCT}$ = interprétat° explicative (non pas prédictif)

. $MSE = \hat{\sigma}^2$ On cherche σ petit

Lois des estimateurs et tests sur les paramètres

Cours : $\begin{pmatrix} \hat{\mu} \\ \hat{\alpha} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu \\ \alpha \end{pmatrix}, \begin{pmatrix} \mathbb{V}(\hat{\mu}) & \text{cov}(\hat{\mu}, \hat{\alpha}) \\ & \mathbb{V}(\hat{\alpha}) \end{pmatrix} \right)$.

Gauss Markov : estimateurs sans biais de α, μ fonction linéaire des Y_i, ceux de variance minimale

$$\begin{cases} \mathbb{V}(\hat{\alpha}) = \sigma^2 / \sum (x_i - \bar{x})^2 . \\ \mathbb{V}(\hat{\mu}) = \sigma^2 \left(1/n + (\bar{x})^2 / \sum (x_i - \bar{x})^2 \right) . \end{cases}$$

Test : $H_0 : \{ \alpha = 0 \} \rightarrow T = \dfrac{\hat{\alpha}}{S_{\hat{\alpha}}}$ avec $S_{\hat{\alpha}} = \dfrac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}$

$T \sim \mathcal{T}(n-2)$.

or $\mathcal{H}_p$, si $T \sim \mathcal{T}(p)$ alors $T^2 \sim \mathcal{F}(1, p)$

↳ On retrouve les $\hat{m}$ valeurs des statistiques dans la table d'ANOVA.

Intervalles de Confiance, Intervalle de Prédiction

IC de la droite de regression :

$$IC_{1-\alpha}(\hat{\alpha}) = \left[\hat{\alpha} \pm t_{n-2;\, 1-\alpha/2} \times S_A \right]$$

$$IC_{1-\alpha}(\hat{\mu}) = \left[\hat{\mu} \pm t_{n-2;\, 1-\alpha/2} \times S_\mu \right].$$

$$\hat{Y}_i \sim \mathcal{N}\left(\mu + \alpha x_i, \sigma \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \right)$$

$$\Rightarrow IC_{1-\alpha}(\hat{Y}_i) = \left[\hat{\mu} + \hat{\alpha} x_i \pm t_{n-2,\, 1-\alpha/2} \times \hat{\sigma} \times \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_j - \bar{x})^2}} \right.$$

Intervalle de prediction :

→ On veut predire Y_0 pour un x_0 non observé.

$$Y_0 = \hat{\mu} + \hat{\alpha} x_0 + E_0 .$$

$$\mathbb{V}(Y_0) = \mathbb{V}(\hat{\mu} + \hat{\alpha} x_0) + \mathbb{V}(E_0) . \rightarrow Y_0 \text{ ne depend pas de } x_0$$

étant donné qu'on n'a pas observé E_0, on a bien $\perp\!\!\!\perp$.

$$\mathbb{V}(Y_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) + \sigma^2 .$$

$$IP_{1-\alpha}(Y) = \left[\hat{\mu} + \hat{\alpha} x_0 \pm t_{n-2,\,1-\alpha/2} \times \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \right] .$$