11.1. Les modèles de régression à variables qualitatives indépendantes ("dummy variable reg. models")

A) Régression sur une seule variable qualitative : ANOVA

Exemple :

$$Y_i = \alpha + \beta D_i + u_i$$

où
- Y = salaire annuel d'un prof. d'université.
- D_i = 1 si professeur est un homme.
  = 0 sinon
- u_i = terme d'erreur.

Hypothèse : u_i satisfait aux hyp. du modèle de rég. linéaire classique $\Rightarrow$ $E(Y_i / D_i = 0) = \alpha$ $\equiv$ salaire moyen ♀

$E(Y_i / D_i = 1) = \alpha + \beta$ $\equiv$ salaire moyen ♂

$$\hat{Y}_i = 18{,}00 + 3{,}28\, D_i$$

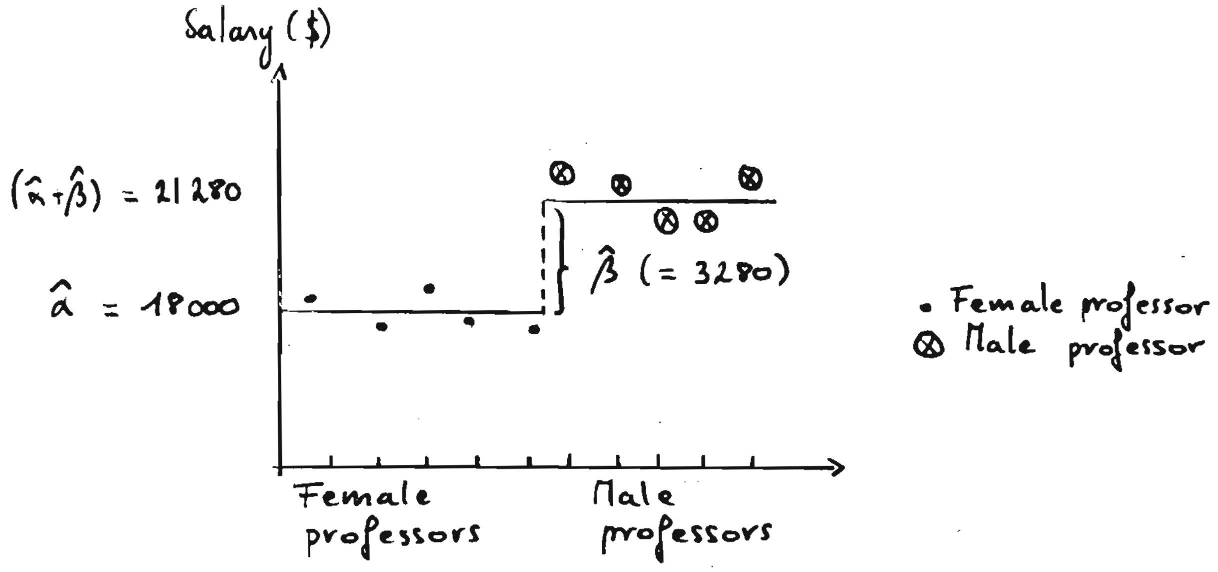
s.e. = (0,32) (0,44)

t = (57,74) (7,439)

$R^2 = 0{,}8737$, $n = 10$, Y en milliers de \$

$\hat{\alpha} = 18000$ \$

$\hat{\alpha} + \hat{\beta} = \pm 21000$ \$ (21 280 \$)



B) Régression sur une variable qualitative et une variable quantitative : ANCOVA

Variables quantitatives ds modèles ANCOVA = „covariates"

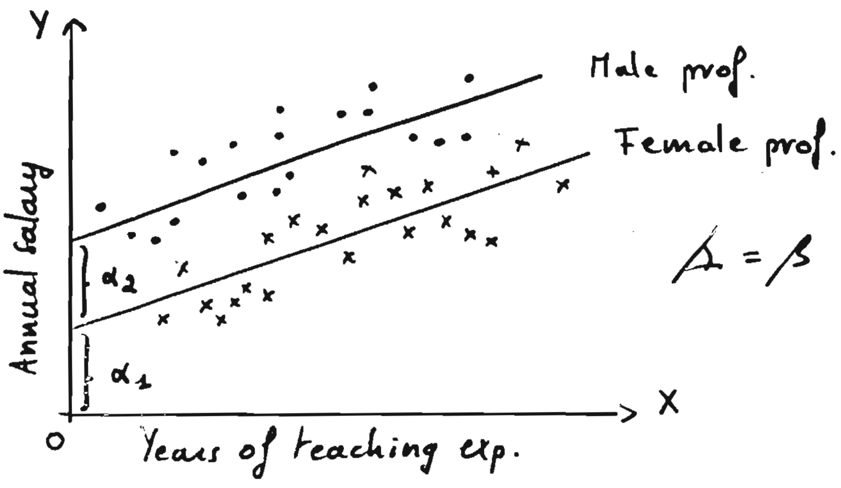
$$Y_i = \alpha_1 + \alpha_2 D_i + \beta X_i + u_i$$

où
- Y = salaire annuel d'un prof.
- X_i = nbre d'années d'expérience.
- $D_i = 1$, si le prof. est un homme
  $= 0$ sinon
- u_i = terme d'erreur

Si $E(u_i) = 0$:

$E(Y_i \mid X_i, D_i = 0) = \alpha_1 + \beta X_i$ ≡ sal. moyen ♀

$E(Y_i \mid X_i, D_i = 1) = (\alpha_1 + \alpha_2) + \beta X_i$ ≡ sal. moyen ♂



$\Delta = \beta$

- Caractéristiques :

1) Une seule dummy (D_i) pour distinguer les ♀ des ♂.
Quid si une dummy pour chaque sexe ?

$$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta X_i + u_i$$

où $\begin{cases} Y = \text{salaire annuel d'un prof.} \\ X_i = \text{nbre d'années d'exp.} \\ D_{2i} = 1 \text{ si prof. est un homme} \\ \quad = 0 \text{ sinon} \\ D_{3i} = 1 \text{ si prof. est une femme} \\ \quad = 0 \text{ sinon} \\ u_i = \text{terme d'erreur} \end{cases}$

Colinéarité pfte ent D_{2i} et D_{3i}.

Illustration : éch. de 3 prof ♂ et 2 prof. ♀

$\exists\, \lambda_1, \lambda_2, \lambda_3$ non tous nuls tq: $\lambda_1 X_1 + \lambda_2 X_2 + \lambda_3 X_3 = 0$

Pq ?

soit $\lambda_1 = 1$, $\lambda_2 = -1$, $\lambda_3 = -1$

		X_1 Intercepte	X_2 D_2	X_3 D_3	X
Homme	Y_1	1	1	0	X_1
Homme	Y_2	1	1	0	X_2
Femme	Y_3	1	0	1	X_3
Homme	Y_4	1	1	0	X_4
Femme	Y_5	1	0	1	X_5

$\Rightarrow X_1 - X_2 - X_3 = 0$ $\quad D_2 = 1 - D_3$ ou $D_3 = 1 - D_2$ $\quad$ ($1 \equiv$ intercepte)

$\Rightarrow$ multicolin. Estimation par MCO n'est pas possible.

Règle générale : si une variable qualitative comporte m catégories, $(m-1)$ variables dummy doivent être incluses dans la régression

(si règle non respectée : „ dummy variable trap")

2) L'attribution des chiffres 0 et 1 est plutôt arbitraire.

$$\left.\begin{array}{l} \text{Si } D = 1 \text{ si prof } ♀ \\ \quad\;\; = 0 \text{ sinon} \end{array}\right\} \Rightarrow \begin{array}{l} E(Y_i / X_i, D_i = 1) = (\alpha_1 + \alpha_2) + \beta X_i \\ E(Y_i / X_i, D_i = 0) = \alpha_1 + \beta X_i \end{array}$$

Catégorie de référence / de base !

3) Intercepte mesure valeur moyenne de la catégorie de référence.

4) Coeff. régress. associés aux variables dummy
≡ différentiels d'intercepte
(„ differential intercept coefficients")

c) Régression sur une variable quantitative et une variable qualitative à plus de 2 catégories.

$$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta X_i + u_i$$

où
- Y_i = dépenses annuelles de santé de l'individu i
- X_i = revenu annuel de l'individu i
- D_{2i} = 1 si l'individu i a un diplôme du second
  = 0 sinon
- D_{3i} = 1 si l'individu i a un diplôme du supérieur ou un diplôme universitaire
  = 0 sinon
- u_i = terme d'erreur stochastique

Catégorie de référence = niveau d'éducation primaire ou moins.

α_1 = intercepte catég. de base
α_2 et α_3 indiquent de combien les intercepts des deux autres catég. varient // intercepte catég. de base.

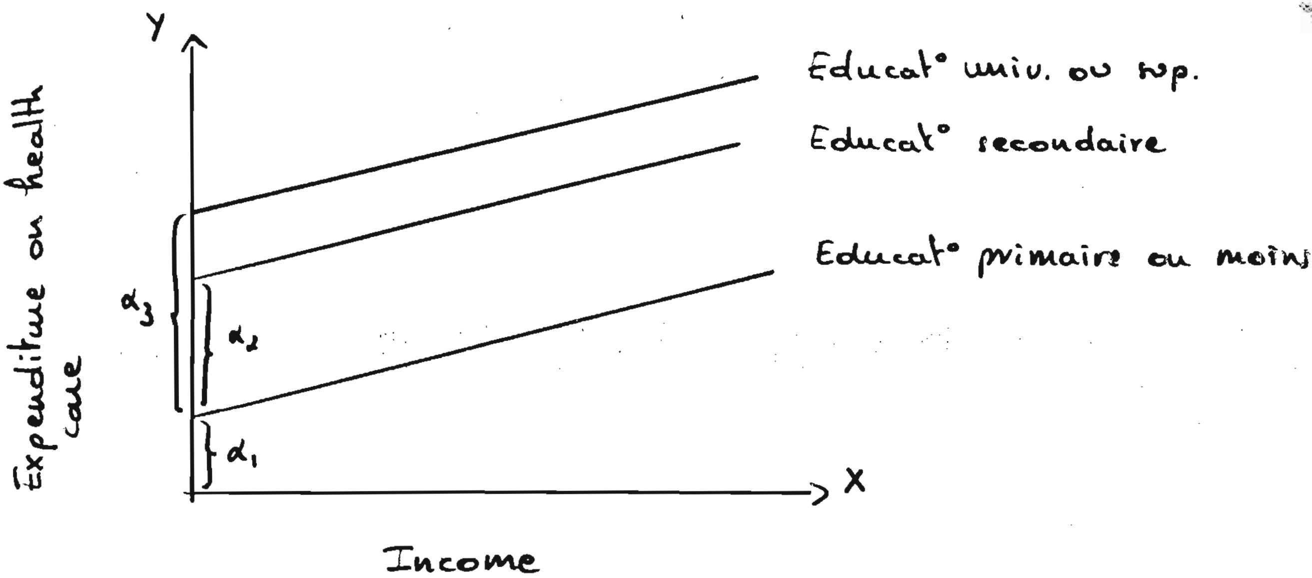
Si $E(u_i) = 0$

$E(Y_i \mid D_2 = 0, D_3 = 0, X_i) = \alpha_1 + \beta X_i$

$E(Y_i \mid D_2 = 1, D_3 = 0, X_i) = (\alpha_1 + \alpha_2) + \beta X_i$

$E(Y_i \mid D_2 = 0, D_3 = 1, X_i) = (\alpha_1 + \alpha_3) + \beta X_i$

si $\alpha_3 > \alpha_2 > 0$



D) <u>Régression sur 2 variables qualitatives et 1 variables quantitative</u>

$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta X_i + u_i$

où
- Y_i = salaire annuel de l'individu i
- X_i = nombre d'années d'expérience de l'individu i
- D_{2i} = 1 si l'individu i est un homme
  = 0 sinon
- D_{3i} = 1 si l'individu i est blanc
  = 0 sinon
- u_i = terme d'erreur

individu = professeur.

Catégorie de référence : professeurs féminins noirs

Si $E(u_i) = 0$:

$E(Y_i / D_2 = 0, D_3 = 0, X_i) = \alpha_1 + \beta X_i$ $\equiv$ sal. moyen prof ♀ noirs

$E(Y_i / D_2 = 1, D_3 = 0, X_i) = (\alpha_1 + \alpha_2) + \beta X_i$ $\equiv$ sal. moyen prof. ♂ noirs

$E(Y_i / D_2 = 0, D_3 = 1, X_i) = (\alpha_1 + \alpha_3) + \beta X_i$ $\equiv$ sal. moyen prof ♀ blancs

$E(Y_i / D_2 = 1, D_3 = 1, X_i) = (\alpha_1 + \alpha_2 + \alpha_3) + \beta X_i$ $\equiv$ sal. moyen prof ♂ blancs

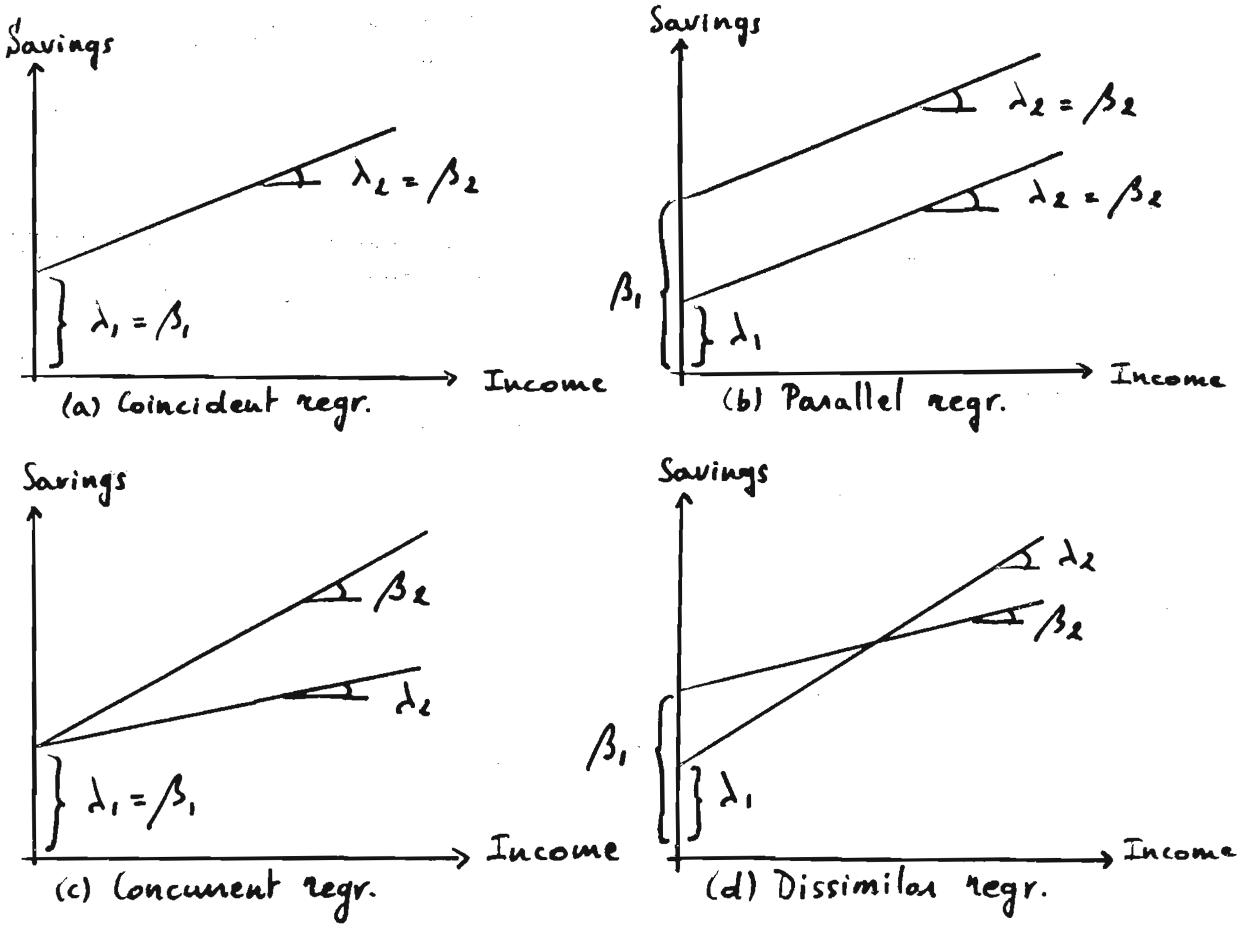
E) Les variables dummy comme alternative au test de Chow

- Rappel : test de Chow pour tester stabilité des paramètres dans relation lel épargne et revenu des ménages sur la période 1970 - 1995 aux USA

Période 1970 - 1981 : $Y_t = \lambda_1 + \lambda_2 X_t + u_{1t}$ $(n_1 = 12)$ (1)

Période 1982 - 1995 : $Y_t = \beta_1 + \beta_2 X_t + u_{2t}$ $(n_2 = 14)$ (2)

Période 1970 - 1995 : $Y_t = \alpha_1 + \alpha_2 X_t + u_{3t}$ $(n = 26)$ (3)



(a) Coincident regr.

(b) Parallel regr.

(c) Concurrent regr.

(d) Dissimilar regr.

- *En pratique*, pour déterminer la source de la $\neq$ lel les 2 régressions, on regroupe ens. des obs. et on estime modèle suivant :

$$\boxed{Y_t = \alpha_1 + \alpha_2 D_t + \delta_1 X_t + \delta_2 (D_t . X_t) + u_t} \qquad (4)$$

où
$$\begin{cases} Y = \text{Épargne} \\ X = \text{revenu} \\ t = \text{tmps} \\ D = 1 \text{ pour obs. de la période } 1982-1995 \\ \quad = 0 \text{ sinon} \\ u = \text{terme d'erreur} \end{cases}$$

▷ Si $E(u_i) = 0$:

$E(Y_t / D_t = 0, X_t) = \alpha_1 + \delta_1 X_t$ ≡ Épargne moyenne période 70-8

$E(Y_t / D_t = 1, X_t) = (\alpha_1 + \alpha_2) + (\delta_1 + \delta_2) X_t$ ≡ Épargne moyenne période 82-95

▷ Si $(\lambda_1 = \alpha_1)$ et $(\lambda_2 = \delta_1)$
$(\beta_1 = \alpha_1 + \alpha_2)$ et $(\beta_2 = \delta_1 + \delta_2)$ } on obtient :

$\Rightarrow Y_t = \lambda_1 + \lambda_2 X_t + u_{1t}$ (période 1970 - 1981) (1)

$Y_t = \beta_1 + \beta_2 X_t + u_{2t}$ (période 1982 - 1995) (2)

▷ Dans équation (4) :

α_2 ≡ différentiel d'intercepte lel 2 périodes

δ_2 ≡ différentiel de pente lel 2 périodes

▷ Inclusion d'une dummy sous la forme :

- interactive / multiplicative permet de tester des $\neq$ de pentes
- additive permet de tester des $\neq$ d'interceptes

- $$\hat{Y}_t = 1,0161 + 152,4786\, D_t + 0,0803\, X_t - 0,0655\, (D_t X_t)$$

$$\begin{array}{lllll} se = & (20,1648) & (33,0824) & (0,0144) & (0,0159) \\ t = & (5,5043) & (4,6090) & (5,5413) & (-4,0963) \end{array}$$

où $\begin{cases} Y = \text{épargne} \\ X = \text{revenu} \\ t = \text{temps} \\ D = 1 \quad \text{pour obs. lel 1982 et 1995} \\ \quad = 0 \quad \text{sinon} \end{cases}$

- Fonctions de régression de chaque période :

$$\hat{Y}_t = 1,0161 + 0,0803\, X_t$$ ≡ regr. épargne - revenu pour 1970 - 1981

$$\begin{aligned} \hat{Y}_t &= (1,0161 + 152,4786) \\ &\quad + (0,0803 - 0,0655)\, X_t \\ &= 153,4947 + 0,0148\, X_t \end{aligned}$$ ≡ régr. épargne - revenu pour 1982 - 1995

- <u>Avantages „dummy" par rapport test de Chow</u> ?

1) On estime 1 seule régression.
2) Estimation + précise car + de df.
3) Permet de tester de nbreuses hyp :
   * tester si $\alpha_2 = \delta_2 = 0$ avec F - test permet de vérifier l'∃ d'un chgmt structurel
   * tester à l'aide d'un test en t si ce sont les intercepts et/ou les pentes qui sont ≠ : ($\alpha_2 = 0$ et/ou $\delta_2 = 0$)

5) Les effets d'interactions en utilisant des dummy

Impact sur la variable dépendante de l'interaction entre plusieurs variables explicatives.

Illustration :

$$\boxed{Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta X_i + u_i} \qquad (1)$$

où
$$\begin{cases} Y = \text{salaire horaire en } \$ \\ X = \text{nbre d'années d'éducation} \\ D_{2i} = 1 \text{ si individu } i \text{ est une femme} \\ \quad\;\; = 0 \text{ sinon} \\ D_{3i} = 1 \text{ si individu } i \text{ est ni blanc, ni hispanique} \\ \quad\;\; = 0 \text{ sinon} \\ u_i = \text{terme d'erreur} \end{cases}$$

Hypothèses implicites :

i) α_2 est identique pour individus ayant couleurs de peau ≠

ii) α_3 est identique pour ♀ et ♂

Pas tjs réalistes*. Il est possible que lorsque les 2 variables indicatrices valent 1 qu'il y ait un effet additionnel sur le salaire moyen. (interaction tel D_2 et D_3)

⇒ effet des variables D_2 et D_3 sur salaire moyen n'est pas nécess. uniq. additif comme dans équation (1) il peut aussi être multiplicatif.

* Il est possible que l'effet du sexe sur la rémunération soit + imp. pour les blancs -1.49- ou les hisp que pour les noirs.

- Prise en compte de l'effet multiplicatif :

$$\boxed{Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 (D_{2i} D_{3i}) + \beta X_i + u_i} \qquad (2)$$

$$\Rightarrow \boxed{E(Y_i / D_{2i} = 1, D_{3i} = 1, X_i) = (\alpha_1 + \alpha_2 + \alpha_3 + \underline{\underline{\alpha_4}}) + \beta X_i}$$

Salaire horaire moyen des ♀ ni blanches, ni hisp. =

$\alpha_1 \equiv$ sal. horaire moyen de la categ. de réf.
(hommes blancs ou hisp.)

$\alpha_2 \equiv$ effet différentiel d'être une ♀

$\alpha_3 \equiv$ effet différentiel d'être ni blanc, ni hisp.

$\alpha_4 \equiv$ effet différentiel d'être une ♀ ni blanche, ni hisp.

- En l'absence de variables d'interaction (uniq. effet additif) :

$$\boxed{E(Y_i / D_{2i} = 1, D_{3i} = 1, X_i) = (\alpha_1 + \alpha_2 + \alpha_3) + \beta X_i}$$

$\Rightarrow$ Introduction des variables dummy sous une forme multiplicative modifie l'effet combiné des variables qualitatives sur la variable dépendante.

Exemple :

(1)
$$\hat{Y}_i = -0,2610 - 2,3606\, D_{2i} - 1,7327\, D_{3i} + 0,8028\, X_i$$
$$t = (-5,367) \quad (-5,4873) \quad (-2,1803) \quad (9,9094)$$
$$R^2 = 0,2032 \quad , \quad n = 528$$

(2)
$$\hat{Y}_i = -0,2610 - 2,3606\, D_{2i} - 1,7327\, D_{3i} + 2,1289\, \underline{\underline{D_{2i} D_{3i}}} + 0,8028\, X_i$$
$$t = (-5,367) \quad (-5,4873) \quad (-2,1803) \quad (1,9802) \quad (9,9095)$$
$$R^2 = 0,2532 \quad , \quad n = 528$$

$\Rightarrow \hat{\alpha}_2 + \hat{\alpha}_3 + \hat{\alpha}_4 = -1,964$ compris lel $-2,3606$ & $-1,7327$

Interaction ‌lel dummy et variable quantitative

Rendement d'une année d'éducation identique pour les ♀ et le ♂ ?

$$\boxed{Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta X_i + \delta (D_{2i} X_i) + u_i}$$

Si $E(u_i) = 0$:

$E(Y_i / D_{2i} = 1, D_{3i} = 0, X_i) = \alpha_1 + \alpha_2 + (\beta + \underline{\underline{\delta}}) X_i$ ♀ BH

$E(Y_i / D_{2i} = 0, D_{3i} = 0, X_i) = \alpha_1 + \beta X_i$ ♂ BH

$E(Y_i / D_{2i} = 1, D_{3i} = 1, X_i) = \alpha_1 + \alpha_2 + \alpha_3 + (\beta + \underline{\underline{\delta}}) X_i$ ♀ N

$E(Y_i / D_{2i} = 0, D_{3i} = 1, X_i) = \alpha_1 + \alpha_3 + \beta X_i$ ♂ N

δ = différentiel de pente (= de rdmt) pour les ♀

6) L'utilisation des variables dummy ds l'analyse saisonnière

Nbreuses séries éco. (p.ex. mensuelles, trimestrielles) présentent des mvts oscillatoires réguliers (structure saisonnière).

Ajustement saisonnier = processus qui consiste à enlever comp. saisonnière d'une série temporelle.

Méthode des variables dummy pour désaisonnaliser une série temporelle.

- Exemple : Ventes de réfrigérateurs aux Etats-Unis entre 1978 et 1995

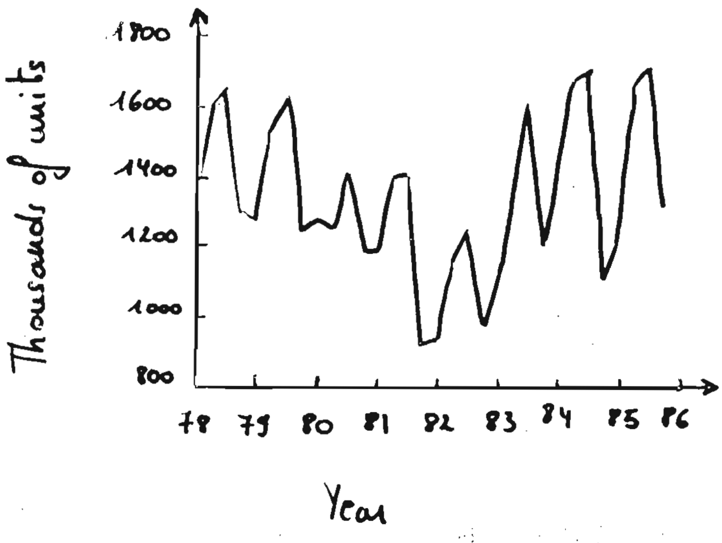


Fig. : Sales of refrigerators 1978 - 1985 (quarterly)

$$\boxed{Y_t = \alpha_1 + \alpha_2 D_{2t} + \alpha_3 D_{3t} + \alpha_4 D_{4t} + u_t}$$

où $\begin{cases} Y = \text{ventes de réfrigérateurs (milliers d'unités)} \\ D_{2,3,4,t} = \text{variables dummy qui prennent valeur 1 (0 sinon)} \\ \quad \text{si donnée relative resp. au 2ème, 3ème ou 4ème trimestre} \end{cases}$

$$\Rightarrow \quad \hat{Y}_t = 1222{,}13 + 245{,}37\, D_{2t} + 347{,}63\, D_{3t} - 62{,}13\, D_{4t}$$

$$t = (20{,}37) \quad (2{,}89) \quad (4{,}10) \quad (-0{,}73)$$

$$R^2 = 0{,}53$$

- Obtention d'une série désaisonnalisée ?

$$\hat{u}_t = Y_t - \hat{Y}_t$$

Série désaisonnalisée = résidus $\hat{u}_t$ de la régression des ventes de réfrigérateurs sur une cst et 3 dummy.

Série de résidus = série initiale − composante saisonnière
(il reste : composante cyclique, tendance et comp. aléatoire)

- <u>Introduction d'une variable expl. quantitative (covariate)</u> :

$$\hat{Y}_t = 456,24 + 242,50\, D_{2t} + 325,26\, D_{3t} - 86,08\, D_{4t} + 2,77\, X_t$$
$$t = (2,56) \quad (3,70) \quad (4,94) \quad (-1,31) \quad (4,45)$$
$$R^2 = 0,729$$

où
$$\begin{cases} Y = \text{ventes de réfrig. (milliers d'unités)} \\ D_t = \text{dummy pour trimestres} \\ X_t = \text{dépenses en biens durables des ménages (milliards de \$)} \\ t = \text{temps} \end{cases}$$

X_t se caractérise églmt par une comp. saisonnière ?

- <u>Théorème de Frish - Waugh</u> montre que l'inclusion des variables dummy D_2, D_3, D_4 (qui contrôlent pour la saisonnalité de la variable Y) suppriment la composante saisonnière de la variable X.

<u>Vérification</u> :

1) $Y \sim cst, D_2, D_3, D_4 \longrightarrow S_1$ (variable Y désais.)
2) $X \sim cst, D_2, D_3, D_4 \longrightarrow S_2$ (var. X désais.)
3) $S_1 \sim S_2$ (+cst) (coeff ass. à S_2 m̃ que coeff. ass. à X de régr. $Y \sim cst, D_2, D_3, D_4, X$)

H) La régression linéaire par morceau („piecewise linear regr.")

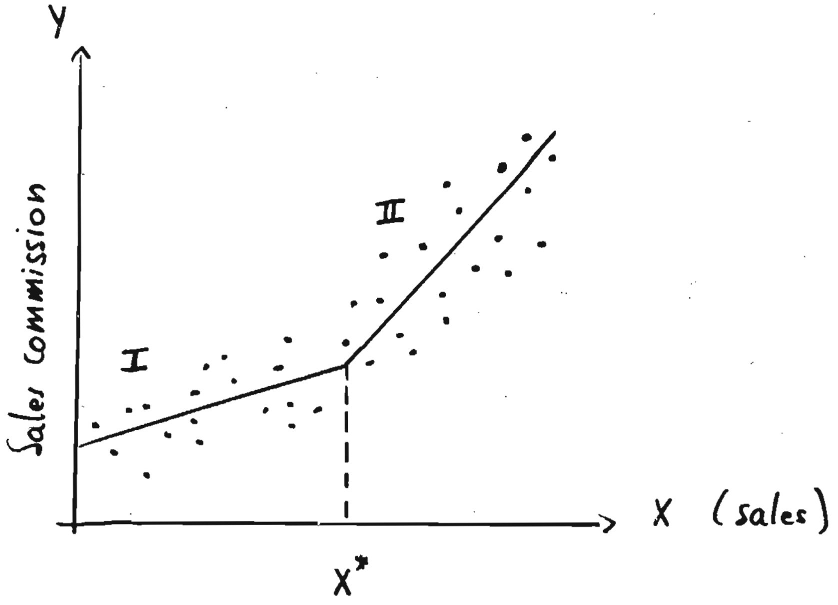


Fig. : Hyp. relationship b/ sales commission and sales volume

Note : The intercept on the Y axis denotes min. guaranteed commission.

- Estimation des segments linéaires (I) et (II) ?

$$\boxed{Y_i = \alpha_1 + \beta_1 X_i + \beta_2 (X_i - X^*) D_i + u_i}$$

où
$$\begin{cases} Y_i = \text{commission de l'individu „i"} \\ X_i = \text{volume de vente de l'individu „i"} \\ X^* = \text{valeur critique pour les ventes (connue a priori)} \\ D = 1 \quad \text{si } X_i > X^* \\ \quad = 0 \quad \text{si } X_i < X^* \end{cases}$$

Remarque : X^* pas tjs connu a priori

i) Analyse graphique

ii) „Switching regression models"

- Si $E(u_i) = 0$:

$$\boxed{E(Y_i \mid D_i = 0, X_i, X^*) = \alpha_1 + \beta_1 X_i} \qquad (X_i < X^*)$$

$$\boxed{E(Y_i \mid D_i = 1, X_i, X^*) = \alpha_1 - \beta_2 X^* + (\beta_1 + \beta_2) X_i} \qquad (X_i > X^*)$$

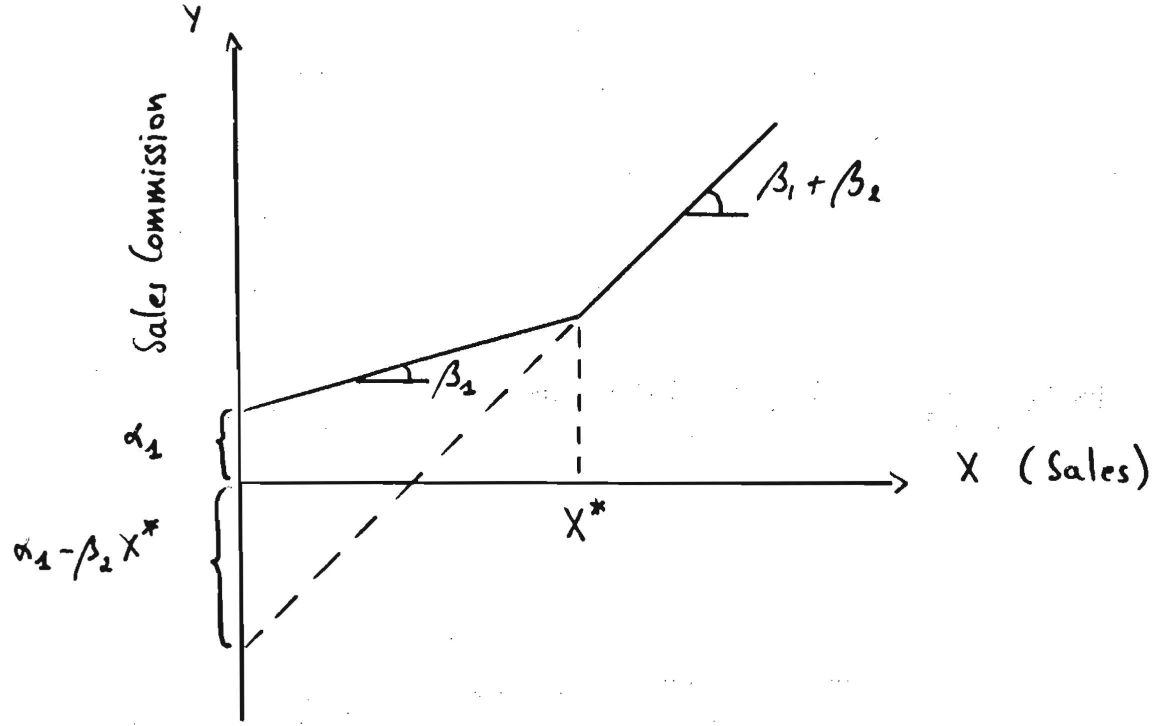


Fig. : Parameters of piecewise linear regression

<u>Exemple</u> : Relation lel coûts totaux & niveau de production

Hyp : coûts marg. de prod. ↘ ou ↗ lorsque prod. > 5500 unités

$$\hat{Y}_i = -145,72 + 0,28\, X_i + 0,09\, (X_i - X^*)\, D_i$$
$$t = \quad (-0,82) \qquad (6,07) \qquad (1,14)$$

où
$$\begin{cases} Y_i = \text{coûts totaux de l'entreprise } i \text{ (en \$)} \\ X_i = \text{niveau de production ds entrep. } i \text{ (en unités)} \\ X^* = 5500 \text{ unités} \\ D_i = 1 \text{ si } X_i > X^* \\ \quad\; = 0 \text{ si } X_i < X^* \end{cases}$$

(1 unité = 1 kg)

I) Interprétation des variables dummy ds rég. semi - logarithmique

Coeff. de régr. ds modèle sem - log. → semi - élasticités

(= mesure variation en % de la variable dép. suite à un chgmt unitaire de la variable explicative).

Interprétation s'applique uniq. aux variables expl. quant.

Quid si var. expl. qualitative ?

$$\boxed{\ln Y_i = \beta_1 + \beta_2 D_i + u_i}$$

où
$$\begin{cases} Y_i = \text{salaire horaire de l'individu } i \text{ (en \$)} \\ D_i = 1 \text{ si individu } i \text{ est une } ♀ \\ \quad = 0 \text{ sinon} \\ u_i = \text{terme d'erreur} \end{cases}$$

Si $E(u_i) = 0$:

Fction de salaire des ♂ : $E(\ln Y_i / D_i = 0) = \beta_1$

Fction de salaire des ♀ : $E(\ln Y_i / D_i = 1) = \beta_1 + \beta_2$

β_1 = moyenne du log. des salaires horaires de ♂

β_2 = différence lel moyenne du log. des salaires horaire ♂ et ♀

$\exp(\beta_1)$ = salaire horaire „médian" des ♂

$\exp(\beta_1 + \beta_2)$ = salaire horaire „médian" des ♀

Rem.: exp (moyenne (ln)) = médiane

exp (ln (moyenne)) = moyenne