

Séance 3

• Exercice 1

1) $d = 5\%$. $H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$
 $t^* = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{-0.1862118}{0.0041758} = -44.593$

$$t_{T-k}^{*1/2} = t_{50}^{0.025} = 1.645$$

$$|t^*| > t_{T-k}^{*1/2} \Rightarrow \text{rejet de } H_0$$

D'après le test de significativité, la variable est significative au seuil de 5%. Cela signifie que le sexe est une variable explicative du salaire. Plus précisément, le logarithme du salaire baisse de -0.186 par rapport à un homme. Toutes choses égales par ailleurs, en moyenne les femmes gagnent 18% de moins que les hommes.

2) $\hat{\beta}_1 = \hat{\beta}_1 \pm t_{T-k}^{*1/2} \times \hat{\sigma}_{\hat{\beta}_1} = -0.1862118 \pm 1.645 \times 0.0041758 = [-0.193, -0.179]$
L'intervalle de confiance à 95% nous indique qu'avec une certitude de 95%, le logarithme du salaire des femmes est entre -0.193 et -0.179 plus bas que celui des hommes.

3) Test de significativité globale à $d = 5\%$.

Hypothèses: $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ $H_1: \exists \beta_i \neq 0, i=1,2,3,4$

Statistique de test: $F^* = \frac{R^2}{1-R^2} \frac{T-k}{k-1} = \frac{R^2}{1-R^2} \frac{34204}{4} = 7,833.262$

$$R^2 = \frac{SCE}{SCT} = \frac{4231.61527}{8850.96067} = 0.478 = F^*$$

$$F^* \sim F(k-1, T-k) \stackrel{d}{=} F(4, \infty) \quad F(4, \infty, 0.05) = 2.37$$

Règle de décision: rejet de H_0 si $F^* > F(k-1, T-k)$

$F^* > F(4, \infty) \Rightarrow$ rejet de H_0 .

D'après le test de significativité globale, le modèle est globalement significatif.

4) $R^2 = 0.478$

Le modèle permet d'expliquer presque 50% des variations du logarithme du salaire. La qualité d'ajustement du modèle est donc relativement élevée.

5) a) L'intérêt d'intégrer ces deux nouvelles variables est de capturer des relations non linéaires entre le logarithme du salaire et les variables explicatives. En particulier, il s'agirait d'améliorer le modèle en captant des relations quadratiques entre le logarithme du salaire et l'âge de fin d'étude, ainsi que l'expérience.

(L'impact positif de l'âge de fin d'étude diminue au fur et à mesure que l'âge augmente.)

On va regarder le coefficient de détermination ajusté

$$b) \bar{R}_{11}^2 = 1 - \frac{SCR/(T-k)}{SCT/(T-1)} = 1 - \frac{4,619,3454 / 34,204}{8,850,96067 / 34,208} = 0.478$$

$$\bar{R}_{12}^2 = 1 - \frac{4,564,97915 / 34,204}{8,850,96067 / 34,208} = 0.484$$

La qualité d'ajustement du modèle est améliorée par l'ajout des nouvelles variables

La qualité d'ajustement du modèle n'est que très peu améliorée (moins de 1%).

~~Alternativement~~ On doit regarder la statistique de test correspondant au test de significativité conjoint des deux nouvelles variables, contraintes linéaires de Fisher.

contraintes linéaires de Fisher

c) Il s'agit du test de significativité simultané des deux variables

On interprète le test de la manière suivante, les deux nouvelles variables ne sont pas significatives $p\text{-value} = 0 < 0.05 \Rightarrow$ rejet de H_0 , hypothèse selon laquelle $\beta_{11} = \beta_{12} = 0 \Rightarrow$ pertinent d'ajouter ces deux variables là

d) Par conséquent, l'ajout des formes quadratiques de l'âge à la fin des études et de l'expérience n'est pas pertinent puisque ces variables n'ont pas de pouvoir explicatif dans le modèle.

6) a) Pour prendre en compte l'effet du type de ménage, on devrait introduire 4 variables indicatrices correspondant à 4 modalités, la 5^e modalité étant le groupe de référence représenté par une égalisation des variables indicatrices à 0. En s'assurant qu'il y a suffisamment d'effectif dans chacune des modalités (>5%).

$$b) \bar{R}_{13}^2 = 1 - \frac{4,558,7585 / 34,197}{8,850,96067 / 34,208} = 0.485$$

La qualité d'ajustement du modèle est bonne mais l'incrémentation suite à l'ajout de la variable ménage est faible.

c) Sous réserve que les variables soient significative, on peut dire qu'un couple sans enfant gagne :

- moins qu'une personne seule → voir correction
- plus qu'un ménage à plusieurs
- plus qu'une famille monoparentale
- moins qu'un couple avec enfant.

| | t^* | $t_{0.025}^{0.025} = 1.960$ | pas nécessaire |
|------------------|-------|-----------------------------|---|
| couple_ac_enfant | 2.06 | H_1 | On peut l'affirmer car les signes des coefficients associés à ces deux variables sont significatifs et de signes opposés. |
| pls | -5.01 | H_1 | |

Les deux variables étant significative, on peut affirmer que les personnes en situation de couple avec ou moins un enfant gagne en moyenne plus que les personnes qui vivent à plusieurs mais sans enfants.

Si signes égaux : test décourt entre les deux (test de contraintes linéaires de Fisher)

Test de contraintes linéaire de Fisher

- e) On doit utiliser un test de conformité comparant les coefficients estimés associés aux variable du ménage pour différencier l'effet de chaque type de ménage sur le salaire
- P-value = 0 < 0.05 \Rightarrow H_0 rejeté
- f) D'après le test, l'hypothèse n'est pas rejetée et donc une situation de ménage à plusieurs a le même effet sur le salaire qu'une situation de couple avec enfant.

6)a) Groupe de référence : couple sans enfant

Variables non significatives : personne seule et famille monoparentale
 \Rightarrow TCEEPA, il n'y a pas de différence de salaire entre les catégories personne seule, famille monoparentale et couple sans enfant

TCEEPA, les personnes qui sont en couple avec enfant, gagnent en moyenne 1,5% de plus que les personnes qui sont en couple sans enfant.

A l'inverse, les personnes qui vivent à plusieurs gagnent en moyenne 5,8% de moins que les personnes en couple sans enfants.

Exercice 2

$$\sum_{i=1}^{19} A_{ii} = 362$$

$$\sum_{i=1}^{19} A_{ii}^2 = 12490$$

$$\sum_{i=1}^{19} A_{ii} y_i = 27437$$

$$\sum_{i=1}^{19} y_i = 938$$

$$\sum_{i=1}^{19} y_i^2 = 64926$$

$$N = 19$$

Partie 1

$$y_i = a_0 + a_1 A_{ii} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2 I_N)$$

1) On cherche à minimiser la valeur des résidus :

$$\min Q \quad \text{où } Q = \sum_{i=1}^{19} \epsilon_i^2$$

$$\epsilon_i = y_i - \hat{y}_i = y_i - [a_0 + a_1 A_{ii}]$$

$$\begin{aligned} \epsilon_i^2 &= \{y_i - [a_0 + a_1 A_{ii}]\}^2 = y_i^2 - 2y_i[a_0 + a_1 A_{ii}] + [a_0 + a_1 A_{ii}]^2 \\ &= y_i^2 - 2y_i a_0 - 2y_i a_1 A_{ii} + a_0^2 + 2a_0 a_1 A_{ii} + a_1^2 A_{ii}^2 \end{aligned}$$

$$Q = \sum_{i=1}^{19} y_i^2 - 2a_0 \sum_{i=1}^{19} y_i - 2a_1 \sum_{i=1}^{19} y_i A_{ii} + N \cdot a_0^2 + 2a_0 a_1 \sum_{i=1}^{19} A_{ii} + a_1^2 \sum_{i=1}^{19} A_{ii}^2$$

$$\frac{d}{da_0} Q = -2 \sum_{i=1}^{19} y_i + 2N a_0 + 2a_1 \sum_{i=1}^{19} A_{ii}$$

$$\frac{d}{da_1} Q = -2 \sum_{i=1}^{19} y_i A_{ii} + 2a_0 \sum_{i=1}^{19} A_{ii} + 2a_1 \sum_{i=1}^{19} A_{ii}^2$$

$$\min Q \Leftrightarrow \begin{cases} d/da_0 Q = 0 \\ d/da_1 Q = 0 \end{cases} \quad (\Rightarrow) \begin{cases} -\sum_{i=1}^{19} y_i + N \cdot a_0 + a_1 \sum_{i=1}^{19} A_{ii} = 0 \\ -\sum_{i=1}^{19} y_i A_{ii} + a_0 \sum_{i=1}^{19} A_{ii} + a_1 \sum_{i=1}^{19} A_{ii}^2 = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} a_0 = \frac{\sum_{i=1}^{19} y_i}{N} - a_1 \frac{\sum_{i=1}^{19} A_{ii}}{N} = \bar{y} - a_1 \bar{A}_{ii} \\ (\bar{y} - a_1 \bar{A}_{ii}) \sum_{i=1}^{19} A_{ii} + a_1 \sum_{i=1}^{19} A_{ii}^2 = \sum_{i=1}^{19} y_i A_{ii} \end{cases}$$

$$\Leftrightarrow a_1 \left[\sum_{i=1}^{19} A_{ii}^2 - \frac{1}{N} (\sum_{i=1}^{19} A_{ii})^2 \right] = (\sum_{i=1}^{19} y_i A_{ii}) - \frac{1}{N} (\sum_{i=1}^{19} y_i) (\sum_{i=1}^{19} A_{ii})$$

$$\begin{aligned} \sum_{i=1}^{19} (A_{ii} - \bar{A}_{ii})^2 &= \sum_{i=1}^{19} \{A_{ii}^2 - 2A_{ii}\bar{A}_{ii} + \bar{A}_{ii}^2\} = \sum_{i=1}^{19} A_{ii}^2 - 2N\bar{A}_{ii}^2 + N\bar{A}_{ii}^2 = \sum_{i=1}^{19} A_{ii}^2 + N\bar{A}_{ii}^2 \\ &= \sum_{i=1}^{19} A_{ii}^2 + \frac{1}{N} (\sum_{i=1}^{19} A_{ii})^2 \end{aligned}$$

$$\begin{aligned} \sum_{i=1}^{19} (y_i - \bar{y})(A_{ii} - \bar{A}_{ii}) &= \sum_{i=1}^{19} \{y_i A_{ii} - y_i \bar{A}_{ii} - \bar{y} A_{ii} + \bar{y} \bar{A}_{ii}\} \\ &= \sum_{i=1}^{19} y_i A_{ii} - \bar{y} \cdot N \bar{A}_{ii} - \bar{y} N \bar{A}_{ii} + N \bar{y} \bar{A}_{ii} = \sum_{i=1}^{19} y_i A_{ii} - \frac{1}{N} (\sum_{i=1}^{19} y_i) (\sum_{i=1}^{19} A_{ii}) \end{aligned}$$

$$\Rightarrow \begin{cases} \bar{y} - a_1 \bar{A}_{ii} = a_0 \\ a_1 \sum_{i=1}^{19} (A_{ii} - \bar{A}_{ii})^2 = \sum_{i=1}^{19} (y_i - \bar{y})(A_{ii} - \bar{A}_{ii}) \end{cases} \Rightarrow a_1 = \frac{\sum_{i=1}^{19} (y_i - \bar{y})(A_{ii} - \bar{A}_{ii})}{\sum_{i=1}^{19} (A_{ii} - \bar{A}_{ii})^2}$$

$$\mathbb{E}(\hat{\alpha}_1) = \mathbb{E} \left(\frac{\sum_{i=1}^N (y_i - \bar{y})(\alpha_i - \bar{\alpha})}{\sum_{i=1}^N (\alpha_i - \bar{\alpha})^2} \right) = \frac{\sum_{i=1}^N (\mathbb{E}(y_i) - \mathbb{E}(\bar{y}))(\alpha_i - \bar{\alpha})}{\sum_{i=1}^N (\alpha_i - \bar{\alpha})^2}$$

$$\mathbb{E}(y_i) = \mathbb{E}(a_0 + a_1 \alpha_i + \epsilon_i) = a_0 + a_1 \alpha_i + \mathbb{E}(\epsilon_i)$$

$$\mathbb{E}(\bar{y}) = \mathbb{E}\left(\frac{1}{N} \sum_{i=1}^N y_i\right) = \frac{1}{N} \sum_{i=1}^N \mathbb{E}(y_i) = \frac{1}{N} (N \cdot a_0 + a_1 \sum_{i=1}^N \alpha_i) = a_0 + a_1 \bar{\alpha}$$

$$\Rightarrow \mathbb{E}(\hat{\alpha}_1) = \frac{\sum_{i=1}^N \{[a_0 + a_1 \alpha_i - (a_0 + a_1 \bar{\alpha})](\alpha_i - \bar{\alpha})\}}{\sum_{i=1}^N (\alpha_i - \bar{\alpha})^2} = \frac{a_1 \sum_{i=1}^N (\alpha_i - \bar{\alpha})^2}{\sum_{i=1}^N (\alpha_i - \bar{\alpha})^2} = a_1$$

$$\mathbb{E}(\hat{\alpha}_0) = \mathbb{E}(\bar{y} - \hat{\alpha}_1 \bar{\alpha}) = \mathbb{E}(\bar{y}) - \bar{\alpha} \mathbb{E}(\hat{\alpha}_1) = a_0 + a_1 \bar{\alpha} - \bar{\alpha} a_1 = a_0$$

$$\hat{\alpha}_1 = \frac{\sum_{i=1}^N (y_i - \bar{y})(\alpha_i - \bar{\alpha})}{\sum_{i=1}^N (\alpha_i - \bar{\alpha})^2} = \frac{\sum_{i=1}^N \{y_i \alpha_i - \bar{y} \bar{\alpha} + \bar{y} \alpha_i + \bar{y} \bar{\alpha}\}}{\sum_{i=1}^N \{\alpha_i^2 - 2\bar{\alpha} \alpha_i + \bar{\alpha}^2\}} = \frac{\sum_{i=1}^N y_i \alpha_i + N \bar{y} \bar{\alpha}}{\sum_{i=1}^N \alpha_i^2 - N \bar{\alpha}^2}$$

$$\bar{\alpha} = \frac{1}{N} \sum_{i=1}^N \alpha_i = \frac{362}{15} \quad \bar{y} = \frac{1}{N} \sum_{i=1}^N y_i = \frac{938}{15}$$

$$\hat{\alpha}_1 = \frac{27437 + 15 \cdot (362/15) \cdot (938/15)}{12490 - 15 \cdot (362^2/15)^2} = 13.340 \text{ ou } 1,28$$

$$\hat{\alpha}_0 = \frac{938}{15} - \hat{\alpha}_1 \frac{362}{15} = -259.405 \text{ ou } 31,67$$

1) SCT = SCE + SCR ($\Rightarrow \sum_{i=1}^N (y_i - \bar{y})^2 = \sum_{i=1}^N (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^N (y_i - \hat{y}_i)^2$)

$$SCE = \sum_{i=1}^{15} (\hat{y}_i - \bar{y})^2 = 6137.71889$$

$$SCR = \sum_{i=1}^{15} e_i^2 = SCT - SCE$$

$$SCT = \sum_{i=1}^{15} (y_i - \bar{y})^2 = 6269.733$$

$$\Rightarrow SCR = 132.014 \quad /$$

3) $R^2 = \frac{SCE}{SCT} = \frac{6137.71889}{6269.733} = 0.979 \quad /$

L'âge du véhicule du coût d'entretien annuel
Le modèle permet d'expliquer environ 98% des variations de y. du véhicule
La qualité d'ajustement est extrêmement élevée.

4) L'estimateur sans biais de σ^2 est

$$S^2 = \frac{\sum_{i=1}^N \alpha_i^2}{N - k} = \frac{SCR}{N - k} = \frac{132.014}{15 - 2} = 10.155 \quad /$$

$$\hat{\sigma}_{\hat{\alpha}_0}^2 = S^2 \left(\frac{1}{N} + \frac{\bar{\alpha}^2}{\sum_{i=1}^N (\alpha_i - \bar{\alpha})^2} \right) = 10.155 \left(\frac{1}{15} + \frac{(362/15)^2}{12490 + 362^2/15} \right) =$$

$$\hat{\sigma}_{\hat{\alpha}_1}^2 = \frac{S^2}{\sum_{i=1}^N (\alpha_i - \bar{\alpha})^2} = \frac{10.155}{12490 + 362^2/15} =$$

$$\hat{\sigma}_{\hat{\alpha}_0}^2 = 2,253$$

$$\hat{\sigma}_{\hat{\alpha}_1}^2 = 0,003$$

5) D'après la formule, a_1 est une somme de variables aléatoires suivant une loi normale et suit donc également une loi student d'espérance \bar{a}_1 et de variance $\sigma_{\bar{a}_1}$ que l'on peut approximer par $\hat{\sigma}_{\bar{a}_1}$, et de degrés de liberté ($N-k$)

Un intervalle de confiance au seuil α se construit de la manière suivante :

$$\Pr(\bar{a}_1 \leq a_1 \leq u) = 1 - \alpha \Leftrightarrow \Pr(a_1 \geq u) = \alpha/2 \text{ par symétrie de la loi student}$$

$$\Pr\left(\frac{a_1 - \bar{a}_1}{\hat{\sigma}_{\bar{a}_1}} \geq \frac{u - \bar{a}_1}{\hat{\sigma}_{\bar{a}_1}}\right) = \alpha/2 \Leftrightarrow \frac{u - \bar{a}_1}{\hat{\sigma}_{\bar{a}_1}} = t_{(N-k)}^{(1-\alpha)/2} \Leftrightarrow a_1 = \bar{a}_1 \pm t_{(N-k)}^{(1-\alpha)/2} \hat{\sigma}_{\bar{a}_1}$$

$$t_{(15-2)}^{0.05/2} = t_{13}^{0.025} = 2.160 \Rightarrow a_1 = 13.340 \pm 2.160 \times \frac{1.279}{0.00271} = [13.228; 13.452]$$

$1.161 \quad 1.397$

0.003

6) $a_1 \quad t^* = \hat{a}_1 / \hat{\sigma}_{\bar{a}_1} \quad t_{(N-k)}^{(1-\alpha)/2} = t_{13}^{0.025} = 2.160 /$

~~0~~ $-259.405 / 1.279 = -52.611 \quad H_1$

~~1~~ $13.340 / 0.00271 = 256.254 \quad H_1$

Au seuil de 5%, les coefficients a_0 et a_1 sont significatifs.

Interval de prévision:

7) 4 ans = (4×12) mois = 48 mois $\hat{y}_t = \hat{y}_0 + t_{(N-k)}^{(1-\alpha)/2} \times \hat{\sigma}_{\bar{y}_t}$

$$\hat{y}_0 = -259.405 + 13.340 \times 48 = 380.945 \quad 31.67 + 1.28(48) = 93$$

$$\hat{y}_t = -259.405 + 2.160 \times 1.279 = [270.555; 248.755]$$

$$y_t = [364.889; 396.911] = [85,45; 100,65]$$

D'après le modèle, un véhicule de 4 ans a un coût de maintenance annuel de 38,000€. On peut dire avec 95% de certitude que le coût annuel d'un tel véhicule est compris entre 85,45€ et 100,65€.

$$8,545 \quad 10,065 \text{ €}$$

Partie 2

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + u_i$$

| | | | | |
|-----|---|---|---|---|
| i | 0 | 1 | 2 | 3 |
|-----|---|---|---|---|

| | | | | |
|-------|--------|-------|--------|-------|
| b_i | 31.748 | 1.152 | -0.025 | 5.600 |
|-------|--------|-------|--------|-------|

| | | | | |
|----------------------|-------|-------|-------|-------|
| $\hat{\sigma}_{b_i}$ | 1.253 | 0.061 | 1.949 | 2.022 |
|----------------------|-------|-------|-------|-------|

$$SCE = \sum_{i=1}^3 (\hat{y}_i - \bar{y})^2 = 6194.34598$$

1) $i \quad t^* = \hat{b}_i / \hat{\sigma}_{b_i} \quad t_{(N-k)}^{(1-\alpha)/2} = t_{11}^{0.025} = 2.201$

~~1~~ $1.152 / 0.061 = 18.85 / \quad H_1$

~~2~~ $-0.025 / 1.949 = -0.0161 / \quad H_0$

~~3~~ $5.600 / 2.022 = 2.770 / \quad H_1$

D'après les tests, les variables x_{1i} et x_{3i} sont significatives, tandis que les variables x_{2i} ne le sont pas.

Cela signifie que les coefficients estimés \hat{b}_1 et \hat{b}_3 sont juste au seuil de 5% tandis que \hat{b}_2 ne l'est pas et $b_2 = 0$ avec une certitude de 95%.

→ interprétation

$$2) R^2 = \frac{SCE}{SCT} = \frac{6194.3598}{6269.733} = 0.988 /$$

$$F^* = \frac{R^2}{1-R^2} \cdot \frac{T-k}{k-1} = \frac{R^2}{1-R^2} \cdot \frac{11}{3} = 301.336 /$$

$$F(11-1, T-2) = F(3, 11) = 3.59$$

$|F^* > F \Rightarrow$ rejet de H_0 /

D'après le test, le modèle est globalement significatif, c'est qu'il existe au moins un coefficient différent de 0 /

$$P1/1) \hat{a}_1 = \frac{27437 - \frac{1}{15}(362)(938)}{12490 - \frac{1}{15}(362)^2} = 1.279$$

$$\hat{a}_0 = \frac{1}{15}(938) - 0.12(362) = 31.674$$

annuel

TCEEPA, le coût d'un véhicule augmente de 1.279 centaines d'euro par mois d'âge
Le coût d'un véhicule neuf est de 3.167,4 par mois.

annuel

P2/1) TCEEPA, le coût de maintenance d'un véhicule augmente de 115€ par mois d'utilisation.

TCEEPA, le coût de maintenance d'un véhicule à moteur diesel est de 560€ de plus que celui d'un véhicule à moteur essence.

TCEEPA, le coût de maintenance annuel est le même pour un véhicule de couleur clair et un véhicule de couleur foncé

P2/3) Comparer les résultats de l'estimation des modèles

- coefficient de détermination ajusté

$$\bar{R}^{(1)} = 0.97 \quad \bar{R}^{(2)} = 0.98$$

\Rightarrow modèle 2 privilégier car amélioration de la qualité d'ajustement

- test de contraintes linéaires de Fisher

$$H_0: b_2 = b_3 = 0$$

\Rightarrow rejet de $H_0 \Rightarrow$ privilégier modèle 2 car aucun de ces variables a un pouvoir explicatif.