dmclassif

Antoine Legendre

March 2025

1 Partie 1

1 : Le vecteur associé à la Hongrie est : (55.4, 49.9, 18.1, 32.8, 5.6, 8.5, 10.0, 4.2).
Pour mesurer la distance entre deux individus (pays), on peut utiliser la distance euclidienne.
Pour comparer des classes (groupes de pays), on pourrait utiliser : la distance entre centroïdes (distance entre moyennes des classes).

2 : Inertie totale du nuage de points : 126015.6

3 : Pays Cluster
Allemagne 1
Autriche 1
Belgique 2
Danemark 1
Espagne 2
Finlande 2
France 2
Grece 3
Hongrie 3
Irlande 2
Italie 3
Pays-Bas 1
Portugal 2
Royaume-Uni 1
Suede 1

4 : Les deux premiers éléments réunis sont Allemagne et Royaume-Uni.
La dernière agrégation réunit la classe 1 avec les classes 2 et 3 c'est à dire Allemagne, Royaume-Uni, Suède, Autriche, Danemark et Pays-Bas avec Belgique, France, Irlande, Espagne, Finlande, Portugal, Grèce, Hongrie et Italie.

5 : Au bout de 6 étapes les classes avec stictement plus de deux pays sont la classse composée de Grèce, Hongrie et Italie et la classe composée de Belgique, France et Irlande.
"Inertie intra-classes : 54276.83"

"Inertie inter-classes : 8145.401"

6 : Le coefficient de détermination R^2 mesure la proportion de la variance expliquée par le modèle dans une régression linéaire. Il est défini par :

$$R^2 = 1 - \frac{\text{SCE}_{\text{résiduelle}}}{\text{SCT}}$$

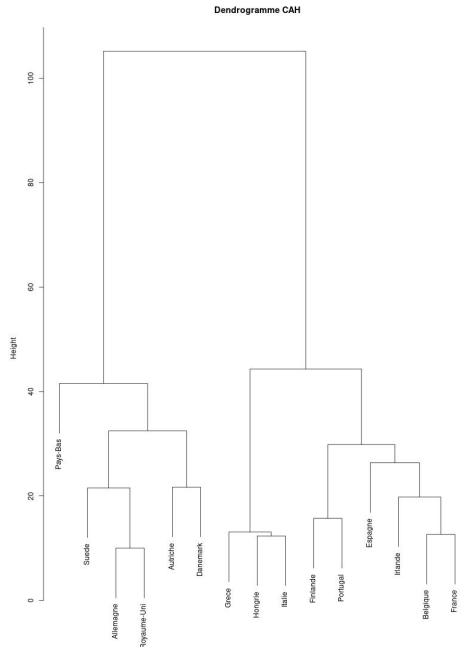
où :
- SCT (somme des carrés totaux) = $\sum(y_i - \bar{y})^2$ → Variabilité totale des valeurs de y.
- $\text{SCE}_{\text{résiduelle}}$ (somme des carrés des erreurs) = $\sum(y_i - \hat{y}_i)^2$ → Variabilité non expliquée par le modèle.

Le R^2 semi-partiel mesure la contribution unique d'une variable explicative supplémentaire dans un modèle de régression multiple. Il est donné par :

$$R^2_{\text{semi-partiel}} = \frac{(R^2_{\text{complet}} - R^2_{\text{réduit}})}{1 - R^2_{\text{réduit}}}$$

où :
- R^2_{complet} est le R^2 du modèle contenant la variable d'intérêt.
- $R^2_{\text{réduit}}$ est le R^2 du modèle sans cette variable.

7 :



Plus la hauteur d'une branche est grande, plus les groupes fusionnés sont éloignés.
Le vecteur des distances entre chaque agrégation est : (2.2046266, 0.2801589, 0.4195891, 2.4734061, 3.8615745, 1.6464685, 0.1221964, 4.3978738, 3.3102337, 2.4908306, 8.5293274, 2.6367516, 57.3909697).

8 : En cherchant :
- Un "coude" dans la courbe R^2,
- Une forte diminution du gain marginal (SPRSQ),
- Un pic dans Pseudo F,
- Un pic dans Pseudo T^2,

on trouve un k optimal à 2, le nombre optimal de classes est donc 2.

9 : La Classe 1 regroupe : Allemagne, Autriche, Danemark, Pays-Bas, Royaume-Uni et Suède.
La Classe 2 regroupe : Belgique, Espagne, Finlande, France, Grece, Hongrie, Irlande, Italie et Portugal.

TABLE 1 – Moyennes par classe pour la partition en 2 classes

Classe	TxEmp	TxEmpFem	TxEmpJeun	TxEmpSen	Tpart	CDD	TxCho	TxChoLD
1	72.88	68.73	53.17	56.23	29.68	11.95	6.32	1.47
2	61.83	55.67	28.19	42.80	14.02	13.92	10.27	3.53

TABLE 2 – Valeurs tests pour la comparaison entre classes (k=2)

Variable	F-value	p-value	t-value	df	Diff. moyennes
TxEmp	32.232	7.575e-05	6.142	12.964	-11.050
TxEmpJeun	29.042	1.234e-04	4.891	7.575	-24.978
TxEmpFem	18.351	8.896e-04	4.900	12.154	-13.067
TxChoLD	16.779	1.262e-03	-3.955	9.577	2.067
Tpart	16.126	1.468e-03	3.688	7.918	-15.661
TxEmpSen	9.149	9.764e-03	2.928	9.672	-13.433
TxCho	7.579	1.644e-02	-3.072	12.821	3.950
CDD	0.446	5.157e-01	-0.703	12.519	1.972

Les variables sont classées par ordre de significativité (p-value croissante). Une p-value < 0.05 indique une différence significative entre les classes.

Comparaison générale entre les deux classes :
La Classe 1 présente des valeurs plus élevées pour : Taux d'emploi général (TxEmp), féminin (TxEmpFem), jeune (TxEmpJeun) et senior (TxEmpSen) et Taux de temps partiel (Tpart).
La Classe 2 présente des valeurs plus élevées pour : Taux de chômage (TxCho), de chômage de longue durée (TxChoLD) et Contrats à durée déterminée (CDD) mais le second tableau indique que cette dernière différence n'est pas significative.
La Classe 1 semble correspondre à des pays avec un marché du travail plus dynamique (emploi élevé, chômage faible), tandis que Classe 2 regroupe des pays avec plus de précarité (emploi faible, chômage plus élevé).

2 Partie 2

1 : Le tableau présenté est un tableau de contingence.

Les variables associées sont :

Les filières universitaires :
- Droit, sciences politiques
- Économie, AES (Administration économique et sociale)
- Arts, lettres, langues, SHS (Sciences humaines et sociales)
- Sciences
- STAPS (Sciences et techniques des activités physiques et sportives)
- Santé

L'origine sociale des étudiants :
- AgArtCommChEnt : Agriculteurs, artisans, commerçants et chefs d'entreprise
- CadPrIntel : Cadres et professions intellectuelles supérieures
- Pinter : Professions intermédiaires
- EmplOuv : Employés et ouvriers

Le nombre d'inscrits : effectif d'étudiants dans chaque filière universitaire selon leur origine sociale.

Ce tableau permet d'analyser la répartition des étudiants en fonction de leur origine sociale et des choix d'orientation universitaire.

2 : La distance du χ^2 car elle pondère les différences en fonction des effectifs globaux.

3 : Nombre d'étudiants d'origine sociale "Cadres et professions intellectuelles supérieures" étudiant une autre matière que les sciences : $68475 + 54497 + 112705 + 18114 + 96241 = 360032$.
Nombre d'étudiants d'origine sociale "Ouvriers" étudiant une autre matière que les sciences : $17771 + 25970 + 53805 + 7126 + 12470 = 117142$.

4 : Cote de choisir la filière "Sciences" contre une autre filière pour un.e étudiant.e du groupe "Employés" :
- Nombre d'étudiants "Employés" en Sciences : 50 914
- Nombre d'étudiants "Employés" dans les autres filières : $32148 + 37731 + 91446 + 12674 + 21964 = 195963$
- Cote (odds) :

$$\text{Odds}_{\text{Employés}} = \frac{\text{Nombre en Sciences}}{\text{Nombre dans les autres filières}} = \frac{50914}{195963} \approx 0.2598$$

La probabilité qu'un étudiant du groupe "Employés" choisisse la filière "Sciences" plutôt qu'une autre filière est d'environ 0.2598, ce qui signifie qu'il y a environ 1 étudiant en Sciences pour 3.85 étudiants dans d'autres filières.

Cote de choisir la filière "Sciences" contre une autre filière pour un.e étudiant.e du groupe "Ouvriers" :
- Nombre d'étudiants "Ouvriers" en Sciences : 33 089
- Nombre d'étudiants "Ouvriers" dans les autres filières : $17771 + 25970 + 53805 + 7126 + 12470 = 117142$
- Cote (odds) :

$$\text{Odds}_{\text{Ouvriers}} = \frac{33089}{117142} \approx 0.2825$$

La probabilité qu'un étudiant du groupe "Ouvriers" choisisse la filière "Sciences" plutôt qu'une autre filière est d'environ 0.2825, soit environ 1 étudiant en Sciences pour 3.54 étudiants dans d'autres filières.

Rapport de cotes (odds ratio) pour "Ouvriers" par rapport aux "Employés" :**

$$\text{Odds Ratio} = \frac{\text{Odds}_{\text{Ouvriers}}}{\text{Odds}_{\text{Employés}}} = \frac{0.2825}{0.2598} \approx 1.087$$

L'odds ratio de 1.087 indique que les étudiant.e.s du groupe "Ouvriers" ont une probabilité légèrement plus élevée (environ 8.7% plus élevée) de choisir la filière "Sciences" par rapport aux étudiants du groupe "Employés". Cependant, cette différence est assez faible, ce qui suggère que l'appartenance à l'un ou l'autre de ces groupes professionnels n'a qu'un effet limité sur le choix de la filière "Sciences".

5 : On souhaite modéliser la probabilité $P(Y = 1|X)$ qu'un étudiant choisisse la filière "Sciences" ($Y = 1$) plutôt qu'une autre filière ($Y = 0$), en fonction de son origine sociale.
Le groupe de référence est fixé comme étant "Employés". Les autres groupes à encoder sous forme de variables indicatrices sont :
- Cadres et professions intellectuelles supérieures (CadPrIntel)
- Professions intermédiaires (Pinter)
- Ouvriers (Ouvriers)
- Retraités et inactifs (Retraités-inactifs)
- Agriculteurs, artisans, commerçants, chefs d'entreprise (AgArtCommChEnt)

Le modèle s'écrit :

$$\log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \beta_0 + \beta_1 \cdot \text{CadPrIntel} + \beta_2 \cdot \text{Pinter} + \beta_3 \cdot \text{Ouvriers} + \beta_4 \cdot \text{Retraités-inactifs} + \beta_5 \cdot \text{AgArtCommChEnt}$$

où :
- β_0 est l'intercept, correspondant au log-odds de choisir "Sciences" pour un étudiant du groupe de référence ("Employés").
- $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ sont les coefficients associés à chaque groupe social, mesurant la différence de log-odds par rapport au groupe de référence.

Interprétation des coefficients :
- Pour le groupe de référence ("Employés") :

$$\text{log-odds} = \beta_0 \quad \Rightarrow \quad \text{Odds} = e^{\beta_0}$$

- Pour un autre groupe (ex. "Ouvriers") :

$$\text{log-odds} = \beta_0 + \beta_3 \quad \Rightarrow \quad \text{Odds} = e^{\beta_0+\beta_3}$$

- Si $\beta_3 > 0$, les "Ouvriers" ont une probabilité plus élevée de choisir "Sciences" que les "Employés".
- Si $\beta_3 < 0$, ils ont une probabilité plus faible.
- L'odds ratio (OR) entre un groupe et le groupe de référence est donné par :

$$\text{OR} = e^{\beta_j} \quad \text{(pour le groupe } j\text{)}$$

Par exemple, pour "Ouvriers" vs "Employés" :

$$\text{OR} = e^{\beta_3}$$

Ce modèle permet de quantifier l'effet de l'origine sociale sur le choix de la filière "Sciences", en comparant chaque groupe au groupe de référence ("Employés"). Les coefficients β_j indiquent si un groupe a une probabilité significativement plus forte ou plus faible de choisir "Sciences".

6 :

TABLE 3 – Résultats de la régression logistique pour le choix de la filière Sciences

Groupe	Estimate (log-odds)	Std. Error	p-value	Interprétation (OR = exp(Estimate))
(Intercept)	-1.3478	0.00497	< 2e-16 ***	Odds de base : exp(−1.3478) ≈ 0.259 (1 Science pour 3.86 autres)
AgArtCommChEnt	+0.0330	0.00849	0.000102 ***	OR ≈ 1.034 (+**3.4%** vs Employés)
CadPrIntel	+0.0964	0.00610	< 2e-16 ***	OR ≈ 1.101 (+**10.1%** vs Employés)
Ouvriers	+0.0836	0.00797	< 2e-16 ***	OR ≈ 1.087 (+**8.7%** vs Employés)
Pinter	+0.1601	0.00714	< 2e-16 ***	OR ≈ 1.174 (+**17.4%** vs Employés)
Retraités-inactifs	-0.1896	0.00759	< 2e-16 ***	OR ≈ 0.827 (**-17.3%** vs Employés)

Tous les coefficients sont significatifs (p-value < 0.001), indiquant des différences robustes entre les groupes.

7 : L'estimation du maximum de vraisemblance du coefficient du modèle associé à une personne appartenant au groupe professionnel ouvriers correspond à la valeur $\hat{\beta}_3$ qui estime β_3.
Dans les résultats de la régression logistique, le coefficient estimé pour le groupe "Ouvriers" est :

$$\hat{\beta}_3 = 0.0836$$

Dans la Q4, nous avions calculé :
- Cote (odds) pour "Employés" (groupe de référence) :

$$\text{Odds}_{\text{Employés}} = \frac{50\,914}{195\,963} \approx 0.2598$$

Log-odds théorique : $\log(0.2598) \approx -1.3478$ (cohérent avec l'intercept du modèle).
- Cote pour "Ouvriers" :

$$\text{Odds}_{\text{Ouvriers}} = \frac{33\,089}{117\,142} \approx 0.2825$$

Log-odds : $\log(0.2825) \approx -1.2641$.

Différence de log-odds ("Ouvriers" vs "Employés") :

$$\hat{\beta}_{\text{Ouvriers}} = \log(\text{Odds}_{\text{Ouvriers}}) - \log(\text{Odds}_{\text{Employés}}) \approx -1.2641 - (-1.3478) = 0.0837$$

Le résultat est très proche de l'estimation R (0.0836), la différence est sûrement due aux arrondis faits dans le calcul manuel.

8 : Estimation de la Variance de l'Estimateur $\hat{\beta}_3$ (Ouvriers) :
- Sortie R :

$$\widehat{\text{Var}}(\hat{\beta}_3) = (\text{Std. Error})^2 = (0.007969)^2 \approx 6.35 \times 10^{-5}$$

Statistique du Test de Wald pour β_3 :
La statistique W est calculée comme :

$$W = \frac{\text{Estimation du coefficient}}{\text{Erreur standard de l'estimation}}$$

Donc :

$$W = \frac{\hat{\beta}_3}{\text{Std. Error}(\hat{\beta}_3)} = \frac{0.083602}{0.007969} \approx 10.491$$

- Valeur R : z value = 10.491‘ (identique au calcul).

Hypothèses Testées :
- H_0 (hypothèse nulle) : $\beta_3 = 0$ (l'appartenance aux "Ouvriers" n'a pas d'effet sur le choix de "Sciences" vs "Employés").
- H_1 (hypothèse alternative) : $\beta_3 \neq 0$ (il existe un effet significatif).

On rejette donc H_0, les "Ouvriers" ont donc une probabilité de choisir "Sciences" significativement différente de celle des "Employés".

9 : La sortie R du modèle indique une déviance de 1 457 676 ce qui est élevée mais pouvait être attendu pour un échantillon de telle taille.
La déviance mesure l'adéquation du modèle aux données. Dans le cas particulier de la régression logistique, elle se réduit à :

$$D = -2 \log \mathcal{L}(\text{modèle courant})$$

c'est à dire :

$$D = -2 \sum_{i=1}^{n} (y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i))$$

où $\hat{p}_i$ est la probabilité prédite par le modèle pour l'observation i.

10 : Interprétation des Sorties du Modèle Logistique
Coefficients et Odds Ratios (OR) :

TABLE 4 – Coefficients du modèle et interprétation

Groupe	$\hat{\beta}$ (log-odds)	Erreur std.	OR	Interprétation
(Intercept)	-1.3478^{***}	0.00497	0.26	Probabilité de base (Employés) : 25.9%
AgArtCommChEnt	0.0330^{***}	0.00849	1.03	+3.4% vs Employés
CadPrIntel	0.0964^{***}	0.00610	1.10	+**10.1**% vs Employés
Ouvriers	0.0836^{***}	0.00797	1.09	+8.7% vs Employés
Pinter	0.1601^{***}	0.00714	1.17	+**17.4**% vs Employés
Retraités-inactifs	-0.1896^{***}	0.00759	0.83	−**17.3**% vs Employés

Intervalles de Confiance :
Le fait que tous les intervalles de confiance des OR sont étroits et ne contiennent pas 1 (limite à laquelle un effet est positif ou négatif), que la valeur de la variance est faible et le résultat du test de Wald nous confirme la significativité des résultats.

On obtient donc que :
- les enfants de cadres et professions intermédiaires ont un avantage net de choisir "Sciences".
- les enfants d'ouvriers sont légèrement favorisés par rapport aux "Employés", mais bien moins que les cadres.
- les enfants de retraités-inactifs sont les plus désavantagés.