Validation régression simple

G. Marot-Briend
guillemette.marot@univ-lille.fr

2020-2021

Plan

Rappels

Langage courant :
Corrélation = liaison entre deux variables quelque soit leur nature.

Sens statistique

- Corrélation : évaluation de la liaison entre deux variables quantitatives (le plus souvent, liaisons essentiellement linéaires)
- Régression : méthode permettant de proposer un modèle mathématique pour expliquer les relations entre les observations.

Problèmes ne relevant pas de la corrélation :

- liaison entre deux variable qualitatives $\Rightarrow \chi^2$
- liaison entre une variable qualitative et une variable quantitative $\Rightarrow$ comparaison de plusieurs moyennes, ANOVA

Coefficient de corrélation

Estimation du coefficient de corrélation

Bravais Pearson :

$$r(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$$r(x,y) = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)(\sum y_i^2 - n\bar{y}^2)}}$$

r mesure seulement le caractère linéaire d'une liaison.

Corrélation de Spearman

Corrélation basée sur les rangs

Corrélation de Spearman

Soient $(x_1, \dots, x_n)$ et $(y_1, \dots, y_n)$, $(R_1, \dots, R_n)$ et $(S_1, \dots, S_n)$ les rangs associés. Le coefficient de corrélation de Spearman calculé entre $(x_1, \dots, x_n)$ et $(y_1, \dots, y_n)$ est égal au coefficient de corrélation de Pearson calculé entre $(R_1, \dots, R_n)$ et $(S_1, \dots, S_n)$.

Le test de Spearman est un test non paramétrique, il ne nécessite pas de loi de probabilité particulière pour (X, Y)
$\Rightarrow$ on l'utilise si $n < 30$ ou pour comparer des classements.

Corrélation de Spearman

En l'absence d'ex aequo, on montre que

$$r_s = 1 - \frac{6 \sum (r_i - s_i)^2}{n(n^2 - 1)}$$

- pour les <u>petits effectifs</u>, les valeurs limites de r_s sont tabulées de façon exacte en fonction du risque α de la table du coefficient de Spearman
- pour les <u>grands effectifs</u>,

$$T = \frac{r_s \sqrt{n-2}}{\sqrt{1 - r_s^2}} \sim St_{n-2}$$

Corrélation de Spearman

Individu k	Age x_k	FCM y_k	Rang âge R_k	Rang FCM S_k
1	40	187	5	6
2	36	195	4	11
3	51	180	9.5	1
4	49	190	7.5	9
5	47	185	6	4
6	51	183	9.5	2
7	32	195	2	11
8	55	185	12.5	4
9	55	189	12.5	7.5
10	23	201	1	13
11	49	189	7.5	7.5
12	52	185	11	4
13	35	195	3	11

$$r_S(Age, FCM) = r(R, S) = -0.73$$

Corrélation de Spearman

Le coefficient de corrélation de Spearman mesure à quel point, quand une variable augmente, l'autre augmente ou diminue.

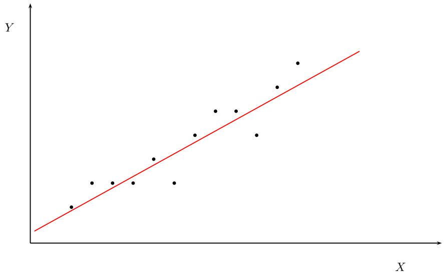
- coefficient compris entre -1 et 1 (0 $\Rightarrow$ pas de liaison)
- alternative à Pearson moins sensible à la non normalité des distributions.
- mesure des liaisons différentes (pas nécessairement linéaires), peut plutôt être vu comme une mesure d'association.

Plan

Régression linéaire simple

La régression linéaire simple consiste à proposer une droite pour expliquer une v.a. quantitative par une autre

$$Y = f(X) + \epsilon$$



Régression linéaire simple

Modèle de régression

$$Y = \alpha X + \beta + \epsilon$$

Hypothèses : $\forall i, j$

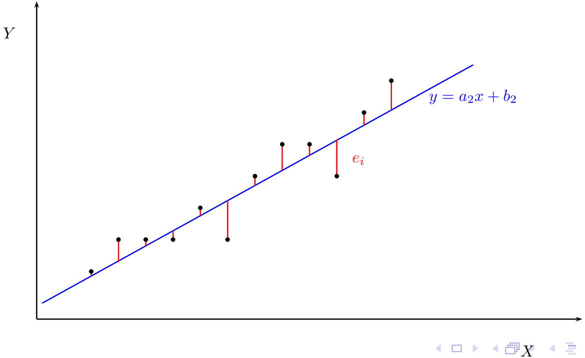
- $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ (normalité des erreurs)
- $E[\epsilon_i] = 0$ (erreurs centrées)
- $V[\epsilon_i] = \sigma^2$ (homoscédasticité des erreurs)
- $E[\epsilon_i \epsilon_j]_{i \neq j} = 0$ (erreurs indépendantes - non corrélées)

Régression linéaire simple

Méthode des Moindres Carrés Ordinaires (MCO)

Minimiser la somme des carrés des écarts

$$\varphi(a,b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$



Régression linéaire simple

Solutions de la minimisation :

$$\hat{a} = \frac{\mathrm{cov}(x, y)}{\mathrm{Var}(x)}$$

$$\hat{b} = \bar{y} - \hat{a}\bar{x}$$

Remarque :
la pente de la droite de régression peut être déduite du coefficient de corrélation r

$$\hat{a} = r\frac{s_y}{s_x}$$

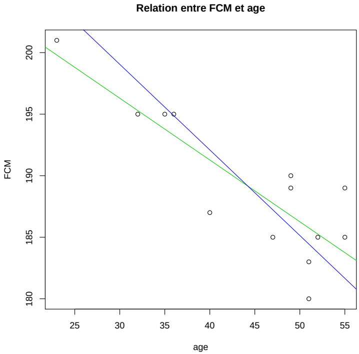
Régression linéaire simple

Exercice :

Individu k	Age x_k	FCM y_k
1	40	187
2	36	195
3	51	180
4	49	190
5	47	185
6	51	183
7	32	195
8	55	185
9	55	189
10	23	201
11	49	189
12	52	185
13	35	195

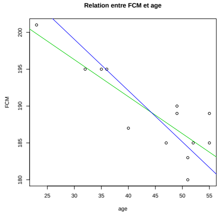
1. expliquer la FCM par l'âge
2. Tracer la droite des MCO sur le nuage de points
3. expliquer l'âge par la FCM
4. Tracer cette deuxième droite des MCO sur le nuage de points

Régression linéaire simple



Les deux droites des MCO sont en général distinctes, elles se coupent toujours au centre de gravité du nuage $(\bar{x}, \bar{y})$.

Régression linéaire simple



L'angle entre ces deux droites donne une mesure de la dépendance entre les variables X et Y : plus cet angle est ouvert, moins la liaison est forte :

- les deux droites de MCO sont confondues $\Longleftrightarrow$ il y a liaison linéaire exacte entre X et Y
- les deux droites de MCO sont perpendiculaires si les deux variables X et Y sont non corrélées.

Régression linéaire simple

Prévision avec la droite des MCO

Si x^* est une nouvelle valeur de X, on prédira la valeur $\hat{y^*}$ de Y donnée par la relation

$$\hat{y^*} = \hat{a}x^* + \hat{b}$$

- s'assurer de la qualité de l'ajustement avant de donner des prévisions
- une prévision d'une valeur de Y n'a de sens que pour des valeurs de X proches de celles utilisées pour déterminer $\hat{a}$ et $\hat{b}$

Prévision avec la droite des MCO

Démarche

1. calcul de la droite des MCO
2. validation du modèle $\Rightarrow$ étude des résidus et détection des valeurs aberrantes et influentes
3. qualité de l'ajustement $\Rightarrow$ décomposition de la variance, coefficient de détermination et test significativité globale
4. qualité de prédiction (PRESS) et prédiction

Prévision avec la droite des MCO

Etude des résidus

On appelle valeur ajustée de la $i^{\text{ème}}$ observation de la variable Y l'approximation

$$\hat{y}_i = \hat{a}x_i + \hat{b}$$

On appelle résidu e_i, l'erreur observée que l'on commet en approchant y_i par $\hat{y}_i$: $e_i = y_i - \hat{y}_i$

Ne pas confondre erreurs non observables et résidus.

Etude des résidus

Propriété : Un estimateur sans biais de la variance de l'erreur du modèle est

$$s_e^2 = \frac{1}{n-2} \sum e_i^2$$

Validité du modèle

- vérifier la normalité des résidus
- vérifier que les résidus ne contiennent pas d'information structurée
- vérifier que les résidus ne sont pas auto-corrélés entre eux

Etude des résidus

Vérification de la normalité des résidus

- histogramme $\Rightarrow$ la distribution doit être unimodale et symétrique autour de 0.
- tests (Kolmogorov-Smirnov, Shapiro Wilks, ...)
- droite de Henry $\Rightarrow$ confronte les quantiles théoriques de la loi normale et la distribution cumulée estimée sur les données

Exercice :

1. Calculer les résidus du modèle de régression de la FCM par l'âge
2. Tracer la droite de Henry pour ces résidus

Etude des résidus

Droite de Henry :

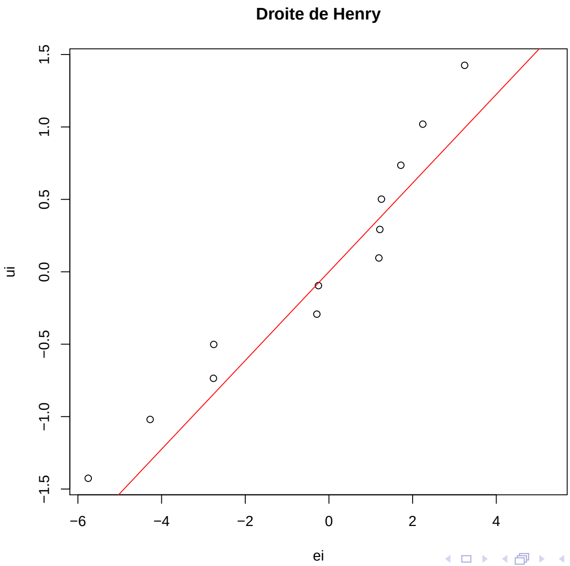
$P(\epsilon \leq e_i)$	0.077	0.154	0.231	0.308	0.385	0.462	0.538
e_i	-5.75	-4.27	-2.76	-2.75	-0.29	-0.25	1.19
u_i	-1.43	-1.02	-0.74	-0.50	-0.29	-0.10	0.10

$P(\epsilon \leq e_i)$	0.615	0.692	0.769	0.846	0.923	1
e_i	1.22	1.25	1.72	2.24	3.24	5.25
u_i	0.29	0.50	0.74	1.02	1.43	∞

Ex : $P(Z \leq u_i) = P(\epsilon \leq e_i) = \frac{7}{13} = 0.538 \Rightarrow u_i = 0.0954$

0,44	0,1510	0,1484	0,1459	0,1434	0,1408	0,1383	0,1358	0,1332	0,1307	0,1282	0,1257	0,50
0,45	0,1257	0,1231	0,1206	0,1181	0,1156	0,1130	0,1105	0,1080	0,1055	0,1030	0,1004	0,54
0,46	0,1004	0,0979	0,0954	0,0929	0,0904	0,0878	0,0853	0,0828	0,0803	0,0778	0,0753	0,53
0,47	0,0753	0,0728	0,0702	0,0677	0,0652	0,0627	0,0602	0,0577	0,0552	0,0527	0,0502	0,52
0,48	0 0502	0 0476	0,0451	0,0426	0,0401	0,0376	0,0351	0,0326	0,0301	0,0276	0,0251	0,51
0,49	0,0251	0,0226	0,0201	0,0175	0,0150	0,0125	0,0100	0,0075	0,0050	0,0025	0,0000	0,50
	0,010	0,009	0,008	0,007	0,006	0,005	0,004	0,003	0,002	0,001	0,000	$F(u)$

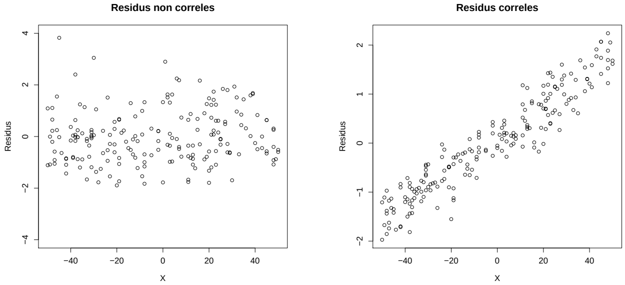
Etude des résidus



Etude des résidus

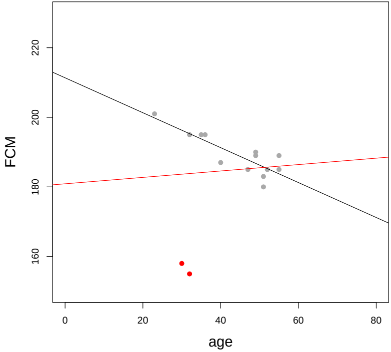
Vérification de l'homoscédasticité des résidus

Les résidus sont homoscédastiques si leur répartition est homogène et ne dépend pas des valeurs de la variable explicative (et donc pas non plus des valeurs prédites).
On vérifie que les résidus n'ont pas de structure particulière en traçant un graphe des résidus :



Observations aberrantes / influentes

Exemple d'observations aberrantes



- Effet important sur l'estimation de la droite de régression
- Mauvais ajustement aux données

Observations aberrantes / influentes

Etude des observations aberrantes / influentes

Effet levier de l'observation i

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

Impact de y_i sur $\hat{y}_i$, lié à l'éloignement de l'observation x_i à la moyenne $\bar{x}$.
Levier grand $\Rightarrow$ observation atypique.

Remarque : Même si l'hypothèse d'homoscédasticité est vérifiée, les résidus n'ont pas la même variance.
$E(e_i) = 0$ et $Var(e_i) = \sigma^2(1 - h_i)$

Etude des résidus

Résidus standardisés "internes" (avec i)

$$r_i = \frac{e_i}{s_e\sqrt{1-h_i}}$$

La standardisation interne dépend de e_i dans le calcul de l'estimation de $Var(e_i)$. Une estimation non biaisée de cette variance est basée sur

$$s^2(-i) = \left[(n-2)s_e^2 - \frac{e_i^2}{1-h_i}\right]/(n-3)$$

qui ne tient pas compte de l'observation i.

Etude des résidus

Résidus studentisés externes (sans i)

$$t_i = \frac{y_i - \hat{y}_i(-i)}{s(-i)\sqrt{1 - h_i(-i)}}$$

- $h_i(-i), \hat{y}_i(-i)$: levier et prédiction de i à partir du modèle estimé sans observation i.
- Sous hypothèse de normalité, on montre que ces résidus studentisés suivent une loi de Student à (n-3) degrés de liberté.

$\Rightarrow$ graphe des résidus : les résidus studentisés sont comparés aux bornes -2 et 2.

Observations aberrantes / influentes

La distance de Cook permet d'évaluer l'influence d'une observation i sur l'estimation des coefficients. Elle prend en compte à la fois l'effet levier (éloignement par rapport à la moyenne) et la taille des résidus.

Distance de Cook pour une observation i

$$D_i = \frac{\sum_{i=1}^{n}(\hat{y}_i - \hat{y}_i(-i))^2}{2s_e^2} = \frac{h_i}{2(1-h_i)} r_i^2$$

Règle de décision (cas régression simple) : $D_i > 1$

Si la différence entre les prédictions est élevée, l'observation i joue un rôle sur l'estimation des coefficients.

Etude des résidus

Graphe des résidus

- on peut considérer les points mal expliqués, si ils ne sont pas trop nombreux, comme des points exceptionnels, les éliminer et recalculer $\hat{a}$ et $\hat{b}$.
- on peut aussi attribuer un poids moindre aux points aberrants $\Rightarrow$ moindres carrés pondérés (fonction de l'écart $|y - \hat{y}|/2s_e$). Méthode plus robuste
- si il y a beaucoup de points mal expliqués (en dehors de la bande), c'est que le modèle est mal choisi ou mal spécifié.

Etude des résidus

Vérification de l'indépendance entre les résidus

Test de Durbin Watson

H_0 : il n'y a pas de corrélation entre ϵ_i et ϵ_{i-1}

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

La valeur de d est toujours comprise entre 0 et 4, $d = 2$ quand il n'y a pas d'autocorrélation.
La loi de d est tabulée (existence de tables statistiques)

Qualité de l'ajustement

Equation d'analyse de la variance

$$y_i - \bar{y} = (\widehat{y}_i - \bar{y}) + (y_i - \widehat{y}_i)$$

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\substack{\text{Somme des carrés}\\ \text{totale}\\ \text{SCT}}} = \underbrace{\sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2}_{\substack{\text{Somme des carrés}\\ \text{expliquée}\\ \text{SCE}}} + \underbrace{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}_{\substack{\text{Somme des carrés}\\ \text{résiduelle}\\ \text{SCR}}}$$

Qualité de l'ajustement

Equation d'analyse de la variance

$$y_i - \bar{y} = (\widehat{y}_i - \bar{y}) + (y_i - \widehat{y}_i)$$

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\substack{\text{Somme des carrés} \\ \text{totale} \\ \text{SCT}}} = \underbrace{\sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2}_{\substack{\text{Somme des carrés} \\ \text{expliquée} \\ \text{SCE}}} + \underbrace{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}_{\substack{\text{Somme des carrés} \\ \text{résiduelle} \\ \text{SCR}}}$$

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\substack{\text{Variance} \\ \text{totale}}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2}_{\substack{\text{Variance} \\ \text{expliquée}}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}_{\substack{\text{Variance} \\ \text{résiduelle}}}$$

Qualité de l'ajustement

Coefficient de détermination

Part de la variance de y expliquée par la relation $\hat{y} = \hat{a}x + \hat{b}$

$$R^2 = \frac{\text{Var}(\hat{y})}{\text{Var}(y)}$$

Dans le cas d'un ajustement linéaire, on peut montrer que $R^2 = r^2(x, y)$ (où r est le coefficient de corrélation linéaire)

- $R^2 \in [0, 1]$
- Plus R est proche de 1, plus le modèle explique correctement la variabilité de Y.

Significativité globale

Le **test F** permet d'évaluer la significativité globale de la régression.

H_0 : La variabilité expliquée est identique à la variabilité résiduelle

Sous H_0

$$F = \frac{\text{Variabilité expliquée par } X}{\text{Variabilité non-expliquée}} = \frac{\frac{SCE}{1}}{\frac{SCR}{n-2}} \sim \mathcal{F}_{1,n-2} \text{ ddl}$$

Interprétation :

$$\begin{cases} \mathrm{H}_0 \ : \ \text{"Le modèle est non explicatif"} \\ \mathrm{H}_1 \ : \ \text{"Le modèle est explicatif"} \end{cases}$$

Intervalles de confiance

Intervalles de confiance des coefficients

$$\mathrm{IC}_{1-\alpha_r}(\alpha) = \left[\hat{a} \pm t_{(1-\alpha_r/2;n-2)} \frac{s_e}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right]$$

$$\mathrm{IC}_{1-\alpha_r}(\beta) = \left[\hat{b} \pm t_{(1-\alpha_r/2;n-2)} s_e \sqrt{\frac{1}{n} \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right]$$

Intervalles de confiance

Intervalles de confiance autour d'une prédiction

$$\mathrm{IC}_{1-\alpha_r}(y*) = \left[\widehat{y_*} \pm t_{(1-\alpha_r/2;n-2)} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}\right]$$

- Lorsque $n \to +\infty$ le radicande tend vers 1 $\Rightarrow$ l'IC devient petit (bonne prédiction)
- Si x_* est proche de $\bar{x}$ alors l'IC devient petit
- A l'inverse, si x_* est éloigné de $\bar{x}$, alors l'IC devient grand (mauvaise prédiction)

Qualité de prédiction

PRESS : predicted residual sum of squares

$$PRESS = \frac{1}{n}\sum_{i=1}^{n} e_{(-i)i}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{e_i}{1-h_i}\right)^2$$

- $e_{(-i)i} = y_i - \hat{y}_{(-i)i}$, $\hat{y}_{(-i)i}$: prévision de y sans observation i
- Estimation **sans biais** de la qualité de prévision, car une même observation n'est pas utilisée pour estimer le modèle et l'erreur de prévision
- Utile pour comparer les qualités **prédictives** de plusieurs modèles (vs. explicatives, R^2) : doit être le plus petit possible

Plan

Conclusion

Croisement de deux variables quantitatives

Représentation graphique (nuage de points)

Coefficient de corrélation

- Calcul de l'indicateur statistique
- Test de nullité du coefficient de corrélation

Régression linéaire

- Estimation des coefficients
- Validité du modèle (Etude des résidus et des observations influentes)
- Qualité d'ajustement (R^2, significativité globale)
- Prédiction