ENSAI - 2A — Année 2024-2025

Chargés de TD/TP :

Julien Jamme — julien.jamme@insee.fr

Théo Leroy — theo.leroy@insee.fr

Yves Ngounou — yves.ngounou@ensai.fr

Clotilde Patarin — clotilde.patarin@square-management.com

Responsable du cours :

Frédéric Lavancier — frederic.lavancier@ensai.fr

Régression linéaire et généralisé - TD 2-5

Exercice 1. *TP avec R sur les données Eucalyptus*

On veut expliquer la hauteur des eucalyptus en fonction de leur circonférence à partir d'une régression linéaire simple. On dispose des mesures des hauteurs (ht) et des circonférences (circ) de 1429 eucalyptus, qui se trouvent dans le fichier "eucalyptus.txt".

1. Extraire et représenter les données dans le plan.
2. Effectuer la régression $y = \beta_1 + \beta_2 x + \epsilon$ où y représente la hauteur et x la circonférence. Commenter les résultats.
3. Quelle est la formule théorique de $\hat{\beta}_1$ et $\hat{\beta}_2$? Retrouver les estimations fournies par R en l'utilisant.
4. Calculer un intervalle de confiance à 95% pour β_1 et β_2, en supposant la normalité des données.
5. Si le bruit ϵ ne suit pas une loi normale, les intervalles de confiance précédents restent-ils valables?
6. Tracer l'estimateur de la droite de régression et un intervalle de confiance à 95% de celle-ci. Que déduisez-vous de la qualité de l'estimation?
7. On veut à présent prédire la hauteur d'une nouvelle série d'eucalyptus de circonférences 50, 100, 150 et 200. Donner les estimateurs de la taille de chacun d'entre eux et les intervalles de prévision à 95% associés, en supposant la normalité des données.

8. Ajouter à la représentation graphique de la question 6 les intervalles de prévision (associés aux mêmes valeurs de la circonférence).

9. Si le bruit ϵ ne suit pas une loi normale, les intervalles de prévision précédents restent-ils valables?

Exercice 2. *Le test de Fisher et le R^2*

On considère un modèle de régression linéaire multiple $y = X\beta + \epsilon$ où $\beta \in \mathbb{R}^p$, X est une matrice de taille (n, p) et ϵ est un vecteur aléatoire de taille n, centré et de matrice de covariance $\sigma^2 I_n$ (I_n est la matrice identité).
On désire tester q contraintes linéaires sur le paramètre β, c'est à dire tester $H_0 : R\beta = 0$ contre $H_1 : R\beta \neq 0$, où R est une matrice de taille (q, p).
On note SCR la somme des carrés résiduelle du modèle initial, et SCR_c la somme des carrés résiduelle du modèle contraint (c'est à dire pour lequel l'hypothèse H_0 est vérifiée).

1. Rappeler la statistique utilisée pour effectuer ce test. On la notera F et on donnera son expression en fonction de SCR et SCR_c.

2. Quelle loi suit cette statistique sous H_0 lorsque ϵ suit une loi normale? Que peut-on dire de sa loi si aucune hypothèse de normalité n'est faite sur ϵ?

3. Montrer que si une constante est présente dans le modèle contraint,

$$F = \frac{R^2 - R_c^2}{1 - R^2} \, \frac{n - p}{q},$$

où R^2 (respectivement R_c^2) désigne le coefficient de détermination du modèle initial (respectivement du modèle contraint).

Exercice 3. *TP avec R sur la consommation de glaces*

On étudie la consommation de glaces aux Etats-Unis sur une période de 30 semaines du 18 Mars 1950 au 11 Juillet 1953. Les variables sont la période (de la semaine 1 à la semaine 30), et en moyenne sur chaque période : la consommation de glaces par personne ("Consumption", en 1/2 litre), le prix des glaces ("Price", en dollars), le salaire hebdomadaire moyen par ménage ("Income", en dollars), et la température ("Temp", en degré Fahrenheit). Les données sont disponibles dans le fichier "icecream-R.dat".

1. Extraire les données et représenter la consommation en fonction des différentes variables.

2. On propose de régresser linéairement la consommation sur les trois variables "Price", "Income" et "Temp", en supposant de plus qu'une constante est présente dans le modèle. On note la constante β_1 et les trois coefficients associés aux variables précédentes respectivement β_2, β_3 et β_4. Réaliser la phase d'estimation de cette régression et commenter le signe des coefficients estimés.

3. Tester la significativité globale du modèle proposé, i.e. $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$, à l'aide du test de Fisher global.

4. Tester la significativité de la variable "Price" dans ce modèle au seuil de 5%. Tester de même la significativité de "Income", puis de "Temp".

5. Comparer le modèle complet précédent et le modèle sans la variable "Price" à l'aide d'un test de Fisher :
   1. En basant le calcul sur la somme des carrés résiduelle de chaque modèle;
   2. En basant le calcul sur le coefficient de détermination de chaque modèle;
   3. En utilisant la fonction `linearHypothesis` de la librairie `car`.
   4. En utilisant la fonction `anova`.

   Quel est la différence entre ce test et le test de Student de significativité de la variable "Price" ?

6. Comparer le modèle complet et le modèle sans la variable "Price" et sans la constante à l'aide d'un test de Fisher. Procéder selon les 4 manières décrites ci-dessus. Commenter.

7. On désire à présent prédire la consommation de glaces pour les données suivantes : $Price = 0.3$, $Income = 85$ et $Temp = 65$. Proposer la prévision qui vous semble la meilleure au vu des modèles étudiés précédemment. Donner un intervalle de prévision au niveau 95% autour de cette prévision.

8. Sous quelle hypothèse l'intervalle de prévision précédent est-il valable? Vérifier-la en observant le QQ-plot des résidus de la régression et en effectuant un test statistique.

9. Vérifier les autres hypothèses en rappelant la définition et en calculant les VIF ("Variance Inflation Factor") de chaque variable explicative et en effectuant une analyse graphique des résidus.

10. Observer le nuage de points en 3D des variables "Consumption", "Income" et "Temp", et l'ajustement par le modèle linéaire, à l'aide de `scatter3d` de la librairie `car`.

Exercice 4. *Le coefficient de corrélation multiple avec ou sans constante*
On considère le modèle de régression

$$y_i = \beta_0 + \beta_1 x_i^{(1)} + \cdots + \beta_p x_i^{(p)} + \epsilon_i,\ i = 1, ..., n, \quad (*)$$

où les variables ϵ_i sont centrées, de variance σ^2 et non-corrélées. On pose $Y = (y_1, \ldots, y_n)^T$, $X^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)})^T$ et $\mathbf{1} = (1, \ldots, 1)^T$. On suppose que les variables $X^{(k)}$ ne sont pas linéairement liées à $\mathbf{1}$. On note $\overline{y}$ la moyenne empirique de Y et $\hat{Y} = \hat{\beta}_0\mathbf{1} + \hat{\beta}_1 X^{(1)} + \cdots + \hat{\beta}_p X^{(p)}$ où les estimateurs sont ceux obtenus par les moindres carrés ordinaires.

1. Que représente géométriquement $\hat{Y}$? Représenter sur un schéma les vecteurs Y, $\hat{Y}$, $\overline{y}\mathbf{1}$, $Y - \overline{y}\mathbf{1}$, $\hat{Y} - \overline{y}\mathbf{1}$ et $\hat{\epsilon}$.

2. En déduire les égalités suivantes :

   (a) $\sum_{i=1}^n y_i^2 = \sum_{i=1}^n \hat{\epsilon}_i^2 + \sum_{i=1}^n \hat{y}_i^2$
   (b) $\sum_{i=1}^n (y_i - \overline{y})^2 = \sum_{i=1}^n \hat{\epsilon}_i^2 + \sum_{i=1}^n (\hat{y}_i - \overline{y})^2$

3. On considère les ratios :

$$R_1^2 = \frac{\sum_{i=1}^n \hat{y}_i^2}{\sum_{i=1}^n y_i^2} \qquad R_2^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^n (y_i - \overline{y})^2}$$

   Justifier que $R_1^2 \geq R_2^2$. Dans quel cas a-t-on égalité?

4. Quelle est la définition du coefficient de corrélation multiple pour le modèle $(*)$?

5. On considère à présent un modèle de régression sans constante, c'est à dire que l'on fixe $\beta_0 = 0$ dans $(*)$. Les égalités montrées en 2) restent-elles valables? Quelle est dans ce cas la définition du coefficient de corrélation multiple?

6. Après estimation du modèle $(*)$ avec ou sans la constante, on obtient $R^2 = 0.72$ avec la constante et $R^2 = 0.96$ sans la constante. L'introduction de la constante est-elle pertinente?

Exercice 5. *L'interprétation du R^2 comme coefficient de corrélation multiple*
On se place dans un modèle de régression contenant une constante. On définit

$$\rho(Y, X) = \sup_{\beta} cor(Y, X\beta)$$

où *cor* désigne la corrélation empirique. Cette quantité est donc la corrélation maximale que l'on peut obtenir entre Y et une combinaison linéaire des variables explicatives.

1. Montrer que
$$cor(Y, X\beta) = \frac{(X\hat{\beta} - \bar{Y}\mathbf{1})'(X\beta - \bar{X}\beta\mathbf{1})}{\|Y - \bar{Y}\mathbf{1}\| \, \|X\beta - \bar{X}\beta\mathbf{1}\|},$$
où $\hat{\beta}$ est l'estimateur par MCO de la régression de Y sur X, et $\bar{X}$ désigne le vecteur des p moyennes empiriques de chaque variable explicative.

2. En déduire que pour tout β, $cor(Y, X\beta)^2 \leq R^2$, où R^2 est le coefficient de corrélation multiple de la régression de Y sur X.

3. Montrer que la borne précédente est atteinte lorsque $\beta = \hat{\beta}$.

4. Conclure que $\rho(Y, X)^2 = R^2$ justifiant la terminologie "coefficient de corrélation multiple".

Exercice 6. *Effet de la multicolinéarité*

On considère un modèle à deux variables explicatives, supposées centrées. De l'estimation sur n individus, on a obtenu les matrices $X'X$ et $X'Y$ suivantes :
$$X'X = \begin{pmatrix} 200 & 150 \\ 150 & 113 \end{pmatrix} \quad X'Y = \begin{pmatrix} 350 \\ 263 \end{pmatrix}.$$
La suppression d'une observation a modifié ces matrices de la façon suivante :
$$X'X = \begin{pmatrix} 199 & 149 \\ 149 & 112 \end{pmatrix} \quad X'Y = \begin{pmatrix} 347.5 \\ 261.5 \end{pmatrix}.$$

1. Calculer les coefficients estimés de la régression dans les deux cas.

2. Calculer le coefficient de corrélation linéaire entre les deux variables explicatives.

3. Commenter.

Exercice 7. *Modélisation de la concentration maximale journalière en ozone*

Le jeu de données "ozone.txt" contient la concentration maximale d'ozone (maxO3) mesurée chaque jour de l'été 2001 à Rennes. Il contient également les températures, la nébulosité et la vitesse du vent mesurés à 9h, 12h et 15h (respectivement T9, T12, T15, Ne9, Ne12, Ne15 et Vx9, Vx12, Vx15), ainsi que la direction principale du vent et la présence ou non de pluie. On désire expliquer au mieux la concentration d'ozone à l'aide des variables disponibles dans le jeu de données.

1. Analyser le nuage de points et la corrélation linéaire entre maxO3 et chacune des variables quantitatives disponibles (c'est à dire T9, T12, T15, Ne9, Ne12, Ne15, Vx9, Vx12 et Vx15). Est-il raisonnable de supposer qu'il existe un lien linéaire entre maxO3 et ces variables?

2. Ajuster le modèle de régression linéaire expliquant maxO3 en fonction de toutes les variables quantitatives précédentes. Tester la significativité de chacune des variables explicatives dans ce modèle. Le résultat est-il en accord avec les observations de la question précédente ? Tester également la significativité globale du modèle. Commenter.

3. Calculer les VIF (Variance Inflation Factor) pour chacune des variables explicatives du modèle précédent. En quoi ces valeurs expliquent les résultats des tests de Student effectués ci-dessus?

4. On décide d'enlever des variables au modèle précédent. Quelles variables semble-t-il naturel d'enlever au vu de la question précédente? Ajuster le nouveau modèle proposé et répéter les analyses effectuées dans les deux questions précédentes.

5. Mettre en oeuvre une sélection automatique du meilleur sous-modèle possible du "gros" modèle ajusté dans la question 2, selon le critère BIC. On pourra utiliser la fonction `regsubsets` de la libraire `leaps` (puis `plot.regsubsets`) ou `step`. Comparer le modèle retenu avec le modèle choisi à la question précédente.

6. Appliquer la sélection automatique précédente en vous basant sur d'autres critères que BIC. Les modèles retenus sont-ils les mêmes? Si non, lequel semble préférable?

7. Analyser les résidus du modèle sélectionné à la question précédente par des représentations graphiques et en effectuant des tests d'homoscédasticité et de non-corrélation des résidus. Toutes les hypothèses d'un modèle linéaire semblent-elles vérifiées?

8. Afin de résoudre le problème d'auto-corrélation des résidus, on propose d'ajouter la maximum d'ozone de la veille dans le modèle. Créer cette variable, que l'on nommera maxO3v et ajouter-la au jeu de données. Observe-t-on un lien linéaire entre maxO3 et maxO3v ?

9. Ajuster le modèle de régression contenant maxO3v comme variable explicative supplémentaire. Analyser les résultats de l'ajustement : les hypothèses d'un modèle linéaire sont-elles vérifiées?

10. Comparer ce dernier modèle au modèle sans maxO3v à l'aide d'un test de Fisher et en comparant les différents critères de sélection (AIC, BIC, Cp de Mallows, R^2 ajusté).

Exercice 8. *Comparaison des critères de sélection d'un modèle*
On considère un modèle de régression linéaire visant à expliquer Y en fonction des variables $X^{(1)}, \ldots, X^{(p)}$. On désire choisir entre le modèle avec $X^{(p)}$ et le modèle sans $X^{(p)}$ (les autres variables étant incluses dans les deux cas), sur la base d'un échantillon de n individus.

On note F la statistique :

$$F = (n-p)\frac{SCR_c - SCR}{SCR},$$

où SCR désigne la somme des carrés résiduelle dans le modèle avec $X^{(p)}$, et SCR_c désigne la somme des carrés résiduelle dans le modèle sans $X^{(p)}$.

1. En appliquant un test de Fisher de modèles emboités, selon quelle règle de décision, basée sur F, choisira-t-on d'inclure la variable $X^{(p)}$ dans le modèle?

2. On rappelle que le R^2 ajusté dans un modèle à k variables et n individus est défini par
$$R^2_a = 1 - \frac{n-1}{n-k}\,\frac{SCR_k}{SCT},$$
où SCR_k désigne la somme des carrés résiduelles dans le modèle, et SCT la somme des carrés totaux.
Montrer qu'on décidera d'inclure $X^{(p)}$ selon le critère du R^2 ajusté si $F > 1$.

3. On rappelle que le C_p de Mallows dans un modèle à k variables et n individus est défini par
$$C_p = \frac{SCR_k}{\hat{\sigma}^2} - n + 2k,$$
où SCR_k désigne la somme des carrés résiduelles dans le modèle, et $\hat{\sigma}^2$ est un estimateur de σ^2 basé sur le plus gros modèle possible. On prendra ici $\hat{\sigma}^2 = SCR/(n-p)$, où SCR désigne la somme des carrés résiduelle dans le modèle avec $X^{(p)}$.
Montrer qu'on décidera d'inclure $X^{(p)}$ selon le critère du C_p de Mallows si $F > 2$.

4. On rappelle que le critère AIC dans un modèle à k variables, à n individus, avec des résidus gaussiens, est défini par
$$AIC = n(1 + \log(2\pi)) + n\log\frac{SCR_k}{n} + 2(k+1),$$

où SCR_k désigne la somme des carrés résiduelles dans le modèle.
Montrer qu'on décidera d'inclure $X^{(p)}$ selon le critère AIC si $F > (n-p)(e^{2/n}-1)$.

5. On rappelle que le critère BIC (aussi parfois appelé SBC) dans un modèle à k variables, à n individus, avec des résidus gaussiens, est défini par

$$BIC = n(1+\log(2\pi)) + n\log\frac{SCR_k}{n} + \log(n)\,(k+1),$$

où SCR_k désigne la somme des carrés résiduelles dans le modèle.
Montrer qu'on décidera d'inclure $X^{(p)}$ selon le critère BIC si $F > (n-p)(e^{\log(n)/n}-1)$.

6. En admettant que les quantiles à 95% d'une loi de Fisher de degré de liberté $(1,\nu)$ prennent leurs valeurs dans l'intervalle $[3.8, 5]$ dès que $\nu > 10$, classer les critères précédents du plus conservatif (i.e. ayant tendance à refuser plus facilement l'introduction de $X^{(p)}$) au moins conservatif (i.e. ayant tendance à accepter plus facilement l'introduction de $X^{(p)}$). On pourra utiliser un développement limité pour l'étude des critères AIC et BIC, en supposant que n est suffisamment grand.

Exercice 9. *Probabilité de sur-ajustement des critères de sélection*
On se place dans le cadre de l'exercice précédent, mais on suppose de plus que la variable $X^{(p)}$ n'est pas significative dans le modèle (i.e. son coefficient est nul dans la régression) et que les résidus sont i.i.d. gaussiens. On admet les résultats énoncés dans les questions de l'exercice précédent.

1. Quelle loi suit la statistique F ? Montrer que lorsque $n \to \infty$, cette loi est équivalente à une loi $\chi^2(1)$.

2. Lors de la mise en oeuvre du test de Fisher des modèles emboités au niveau $\alpha \in [0,1]$, quelle est la probabilité de décider (à tort) d'inclure la variable $X^{(p)}$ dans le modèle?

3. Vers quoi tend la probabilité précédente si on base la décision sur le R^2_a?

4. Même question si la décision est basée sur le C_p de Mallows.

5. Même question si la décision est basée sur le critère AIC.

6. Même question si la décision est basée sur le critère BIC.

7. Quel critère est-il préférable de choisir si l'on souhaite minimiser le risque d'inclure une variable en trop dans le modèle?

Complément : Dans la situation inverse où $X^{(p)}$ est significative dans le modèle et qu'il est donc préférable de l'inclure, on peut montrer (mais c'est plus difficile) qu'en se fiant à n'importe lequel des critères ci-dessus, la probabilité de décider (à tort) ne pas inclure $X^{(p)}$ tend vers 0 lorsque $n \to \infty$.

Exercice 10. *ANCOVA : Retour à la modélisation de l'ozone*

On considère de nouveau les données "ozone.txt" étudiées dans l'exercice 7. On désire tirer profit des variables qualitatives présentes dans le jeu de données (c'est à dire la présence ou non de pluie, et la direction principale du vent) pour éventuellement améliorer le modèle construit dans l'exercice 7.

1. On reprend le modèle sélectionné dans l'exercice 7, soit "maxO3" en fonction de "T12", "Ne9", "Vx9" et "maxO3v" où "maxO3v" représente la concentration maximale en ozone de la veille (créer cette variable si besoin). Ajuster ce modèle sur les données.

2. Représenter graphiquement "maxO3" en fonction de la présence de pluie. Un lien semble-t-il présent ? Le confirmer par un test statistique.

3. Ajouter au modèle de la première question la variable "pluie" de manière la plus générale possible (i.e. en incluant une interaction avec chaque variable en plus d'un effet sur la constante). Tester la significativité de ces ajouts en effectuant un test de Fisher de modèles emboités entre ce modèle et le modèle initial.

4. Tester de même le modèle plus simple dans lequel seul un effet additif de la variable "pluie" est intégré, et non ses interactions avec les autres variables. Le résultat est-il en désaccord avec l'analyse graphique de la question 2 ? Comment expliquer le résultat ?

5. De même : représenter graphiquement "maxO3" en fonction de la direction du vent et étudier la pertinence d'inclure un effet vent dans le modèle initial.

Exercice 11. *Moindres carrés généralisés*

Soit un modèle de régression linéaire multiple

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

où $\beta \in \mathbb{R}^p$, X est une matrice de taille $n \times p$ et ϵ est un vecteur aléatoire de taille n, centré. On considère ici la situation où les variables ϵ_i ne sont plus homoscédastiques et

non-corrélés, mais de façon générale $\mathbb{V}(\epsilon) = \Sigma$ où Σ est une matrice inversible. On suppose dans cet exercice que Σ est connue (il conviendra dans la pratique de l'estimer).

1. Préciser la matrice Σ lorsque les variables ϵ_i sont non-corrélés mais hétéroscédastiques de variance σ_i^2 $(i = 1, \ldots, n)$.

2. Déterminer l'espérance et la variance de l'estimateur $\hat{\beta}$ des moindres carrés ordinaires (dans le cas général d'une matrice Σ quelconque).

3. Pour $S \in \mathbb{R}^n$ et $T \in \mathbb{R}^n$, on définit le produit scalaire entre S et T associé à la matrice Σ^{-1} par $S'\Sigma^{-1}T$, et donc la norme de T associée à Σ^{-1} est $\|T\|_\Sigma^2 = T'\Sigma^{-1}T$. Montrer que la forme explicite de l'estimateur $\hat{\beta}_G$ des moindres carrés généralisés défini comme le minimiseur de $\|Y - X\beta\|_\Sigma$ est

$$\hat{\beta}_G = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y.$$

   En déduire son espérance et sa variance.

4. Montrer que la matrice de covariance entre $\hat{\beta}$ et $\hat{\beta}_G$ est égale à la matrice de variance-covariance de $\hat{\beta}_G$. En déduire que $\hat{\beta}_G$ est meilleur que $\hat{\beta}$ au sens du coût quadratique.