UNIVERSITÉ DE NANTES

Université de Nantes – UFR des Sciences et Techniques
Département de Mathématiques

Master 1 Ingénierie Statistique

TD de Régression

Analyse bivariée

Ex 1. *Lien entre 2 variables quantitatives : le coefficient de corrélation linéaire*
On considère deux échantillons de n variables $(X_1, \dots, X_n)$ et $(Y_1, \dots, Y_n)$.

1) Rappeler la définition de la corrélation linéaire empirique $\hat{\rho}$ entre les deux échantillons précédents.

2) Montrer que si les couples (X_i, Y_i), $i = 1, \dots, n$, sont i.i.d et dans L^2 alors $\hat{\rho}$ converge presque surement vers la corrélation théorique ρ entre X_i et Y_i (justifier que ρ ne dépend pas de i), i.e. quel que soit $i \in \{1, \dots, n\}$,

$$\rho = \frac{E[(X_i - E(X_i))(Y_i - E(Y_i))]}{\sqrt{Var(X_1)Var(Y_i)}}.$$

On suppose dans la suite de l'exercice que les couples (X_i, Y_i), $i = 1, \dots, n$, sont i.i.d suivant une loi normale. On admet dans ce cas, et sous l'hypothèse $H_0 : \rho = 0$, que

$$\sqrt{n-2}\,\frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}} \sim T(n-2).$$

3) En déduire une région critique pour tester $H_0 : \rho = 0$ contre $H_1 : \rho \neq 0$ au niveau $\alpha \in]0, 1[$.

4) Montrer que si Z_n suit la loi $T(n-2)$ alors Z_n converge en loi vers une $\mathcal{N}(0, 1)$.

5) En déduire que sous H_0, $\sqrt{n}\hat{\rho}$ converge en loi vers une $\mathcal{N}(0, 1)$.

6) En déduire une région critique très simple pour tester $H_0 : \rho = 0$ contre $H_1 : \rho \neq 0$ au niveau asymptotique $\alpha \in]0, 1[$.

Ex 2. *Lien entre 2 variables quantitatives : le coefficient de corrélation de Spearman*
On considère deux échantillons de n variables $(X_1, \dots, X_n)$ et $(Y_1, \dots, Y_n)$. On note $(r_1, \dots, r_n)$ (resp. $(s_1, \dots, s_n)$) les rangs des variables X_i (resp. Y_i) dans chaque échantillon. On suppose qu'il n'y a pas d'ex-aeaquo, de telle sorte que les rangs vont de 1 à n. On rappelle que la corrélation de Spearman R_S entre les échantillons $(X_1, \dots, X_n)$ et $(Y_1, \dots, Y_n)$ correspond à la corrélation linéaire entre leurs rangs.

1) Donner la formule définissant R_S.

2) Montrer que la moyenne empirique de l'échantillon $(r_1, \dots, r_n)$ vaut $(n+1)/2$ et que sa variance empirique vaut $(n^2-1)/12$.

3) En déduire que $R_S = \frac{n^{-1}\sum_{i=1}^n r_i s_i - (n+1)^2/4}{(n^2-1)/12}$.

4) Soit $d_i = r_i - s_i$. Montrer que $\sum_{i=1}^n r_i s_i = n(n+1)(2n+1)/6 - 1/2 \sum_{i=1}^n d_i^2$.

5) En déduire que

$$R_S = 1 - \frac{6\sum_{i=1}^n d_i^2}{n(n^2-1)}.$$

Ex 3. *Lien variable quantitative/variable qualitative : comparaison de k moyennes*
On considère un échantillon de n variables indépendantes $(X_1, \ldots, X_n)$. On suppose que cet échantillon est composé de k sous-populations (correspondant aux k classes d'un facteur).
Pour tout $i = 1, \ldots, k$, on note n_i le nombre d'individus dans la sous-population i, et $(X_{i,1}, \ldots, X_{i,n_i})$ les éléments de $(X_1, \ldots, X_n)$ associés à la sous-population i. Ainsi $n = \sum_{i=1}^{k} n_i$ et $(X_1, \ldots, X_n) = \cup_{i=1}^{k}(X_{i,1}, \ldots, X_{i,n_i})$.
On suppose que pour $i = 1, \ldots, k$, les variables $X_{i,1}, \ldots, X_{i,n_i}$ sont identiquement distribuées selon une loi normale d'espérance μ_i et de variance σ^2 inconnues (on suppose donc que la variance est la même quelle que soit la sous-population i).

On pourra utiliser le résultat suivant :
- *Si $Y_1 \sim \chi^2(p)$, $Y_2 \sim \chi^2(q)$, et Y_1 et Y_2 sont indépendantes, alors $Y_1 + Y_2 \sim \chi^2(p+q)$*
- *Réciproquement, si $Y = Y_1 + Y_2$ et $Y \sim \chi^2(r)$, $Y_1 \sim \chi^2(p)$, $p < r$, alors Y_2 est indépendante de Y_1 et $Y_2 \sim \chi^2(r-p)$.*

1) Montrer que
$$S_T^2 = S_{inter}^2 + S_{intra}^2,$$
où $S_T^2 = \sum_{l=1}^{n}(X_l - \overline{X})^2$ représente la somme des carrés totale de l'échantillon ;
$S_{inter}^2 = \sum_{i=1}^{k} n_i(\overline{X_i} - \overline{X})^2$ la somme interclasses (entre les sous-populations) ;
$S_{intra}^2 = \sum_{i=1}^{k} n_i S_i^2$, où $S_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i}(X_{i,j} - \overline{X_i})^2$, la somme intraclasses (moyenne des variances dans chaque sous-population).

2) Quelle est la loi de $n_i S_i^2/\sigma^2$? En déduire que $S_{intra}^2/\sigma^2 \sim \chi^2(n-k)$.

3) Sous l'hypothèse $H_0 : \mu_1 = \cdots = \mu_k$, quelle est la loi de S_T^2/σ^2 ? En déduire la loi de S_{inter}^2/σ^2 sous H_0.

4) En déduire la loi de $F = \frac{S_{inter}^2/(k-1)}{S_{intra}^2/(n-k)}$ sous H_0 et une région critique pour tester H_0 au niveau $\alpha \in]0, 1[$.

Pour quantifier le lien entre une variable quantitative X et une variable qualitative Q à k modalités, on calcule parfois $\hat{\eta}^2 = \frac{S_{inter}^2}{S_T^2}$, qui estime $\eta^2 = V(E(X|Q))/V(X)$.

5) Montrer que $0 \leq \eta^2 \leq 1$ et que $0 \leq \hat{\eta}^2 \leq 1$. A quoi correspond les cas $\eta^2 = 0$ et $\eta^2 = 1$? Formuler l'hypothèse H_0 ci-dessus en fonction de η^2 et exprimer la région critique du test en fonction de $\hat{\eta}^2$.

Ex 4. *Lien entre 2 variables qualitatives*
On considère deux variables qualitatives Q_1 et Q_2 ayant respectivement I et J modalités. On relève sur un échantillon de n individus le nombre d'individus appartenant à chacune des modalités croisées de (Q_1, Q_2), ce que l'on résume dans un tableau de contingence.
Comment tester l'indépendance de Q_1 et Q_2 à l'aide de ce tableau au niveau $\alpha \in]0, 1[$?

Ex 5. *Moyenne empirique*
Soit $z_1, ..., z_n$ des observations d'une variable Z.

1) Déterminer la valeur de $\hat{m}$ qui minimise la distance quadratique $S(m) = \sum_{i=1}^{n}(z_i - m)^2$.

2) La quantité $\hat{m}$ correspond à l'estimation par moindres carrés ordinaires dans un modèle de régression linéaire : $Y = X\beta + \epsilon$. Préciser ce que valent Y, X, β et ϵ.

3) Retrouver le résultat de la première question à partir de la formule générale de l'estimateur des moindres carrés : $\hat{\beta} = [X'X]^{-1}X'Y$.

Ex 6. *Reconnaître un modèle de régression linéaire*
Les modèles suivants sont-ils des modèles de régression linéaire ? Si non, peut-on appliquer une transformation pour s'y ramener ? Pour chaque modèle de régression linéaire du type $Y = X\beta + \epsilon$, on précisera ce que valent Y, X, β et ϵ.

1) On observe $(x_i, y_i), i = 1, ..., n$ liés théoriquement par la relation $y_i = a_0 + a_1 x_i + \epsilon_i,\ i = 1, ..., n.$ où les variables ϵ_i sont centrées, de variance σ^2 et non-corrélées. On désire estimer a_0 et a_1.

2) On observe $(x_i, y_i), i = 1, ..., n$ liés théoriquement par la relation $y_i = a_1 x_i + a_2 x_i^2 + \epsilon_i,\ i = 1, ..., n.$ où les variables ϵ_i sont centrées, de variance σ^2 et non-corrélées. On désire estimer a_1 et a_2.

3) On relève pour différents pays $(i = 1, ..., n)$ leur production P_i, leur capital K_i, leur facteur travail T_i qui sont théoriquement liées par la relation de Cobb-Douglas $P = \alpha_1 K^{\alpha_2} T^{\alpha_3}$. On désire vérifier cette relation et estimer α_1, α_2 et α_3.

4) Le taux de produit actif y dans un médicament est supposé évoluer au cours du temps t selon la relation $y = \beta_1 e^{-\beta_2 t}$. On dispose des mesures de n taux y_i effectués à n instants t_i. On désire vérifier cette relation et estimer β_1 et β_2.

5) Même problème que précedemment mais le modèle théorique entre les observations s'écrit $y_i = \beta_1 e^{-\beta_2 t_i} + u_i$, $i = 1, \ldots, n$, où les variables u_i sont centrées, de variance σ^2 et non-corrélées.

Ex 7. *Régression simple*
On considère le modèle de régression linéaire simple où l'on observe n réalisations $(x_i, y_i), i = 1, ..., n$ liées par la relation

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, ..., n.$$

On suppose que les x_i sont déterministes et que les variables ϵ_i centrées, de variance σ^2 et non-corrélées.

1) Ecrire le modèle sous forme matricielle.

2) De quel problème de minimisation l'estimateur des moindres carrés $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ est-il la solution ?

3) Calculer $\hat{\beta}$ en résolvant directement le problème de minimisation précédent et vérifier la formule $\hat{\beta} = [X'X]^{-1} X'y$.

4) Calculer $\text{var}(\hat{\beta})$ en utilisant la formule précédente et en déduire $\text{var}(\hat{\beta}_0)$, $\text{var}(\hat{\beta}_1)$ et $\text{cov}(\hat{\beta}_0, \hat{\beta}_1)$.

5) Montrer que la moyenne empirique des résidus $\hat{\epsilon}_i$ est nulle.

6) Calculer la matrice de variance-covariance du vecteur de résidus $\hat{\epsilon}$ et du vecteur des valeurs estimées $\hat{y}$.

Ex 8. *Convergence des estimateurs*

On sait que si $X = \begin{pmatrix} 1, \ldots, 1 \\ x_1, \ldots, x_n \end{pmatrix}'$, alors en notant $\overline{x}_n = n^{-1} \sum_{i=1}^n x_i$,

$$(X'X)^{-1} = \frac{1}{\sum_{i=1}^n (x_i - \overline{x}_n)^2} \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n x_i^2 & -\overline{x}_n \\ -\overline{x}_n & 1 \end{pmatrix}$$

1) On a observé un échantillon de couples (x_i, y_i), $i = 1, \ldots, n$. On suppose le lien suivant : $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, où les variables ϵ_i sont i.i.d, centrées, de variance σ^2. Les

régresseurs x_i sont supposés ici aléatoires, i.i.d, de carré intégrable et de variance non nulle. On note $\epsilon = (\epsilon_1, \ldots, \epsilon_n)'$, $\beta = (\beta_0, \beta_1)'$ et $\hat{\beta}$ l'estimateur de β par MCO. On suppose que X et ϵ sont indépendants.

a) Exprimer $\hat{\beta} - \beta$ en fonction de la matrice X et du vecteur ϵ.

b) En déduire que $\hat{\beta}$ converge presque sûrement vers β lorsque $n \to \infty$.

2) Lors d'une expérience chimique, on observe la teneur d'un certain produit à différents instants réguliers allant de 1 à n. Le résultat à l'instant i est noté y_i. On suppose le lien temporel suivant : $y_i = \beta_0 + \beta_1 i + \epsilon_i$, $i = 1, \ldots, n$, où les variables ϵ_i représentent les erreurs de mesures. Elles sont supposées aléatoires, centrées, de variance σ^2 et non corrélées. Soit $\hat{\beta}$ l'estimateur de β par MCO.

a) Calculer $Var(\hat{\beta})$ et donner sa limite lorsque $n \to \infty$.

b) En déduire le comportement asymptotique en moyenne quadratique de $\hat{\beta}_0$ et $\hat{\beta}_1$.

3) On se place sous les mêmes hypothèses que la question précédente mais on suppose cette fois-ci que le lien temporel est : $y_i = \beta_0 + \beta_1/i + \epsilon_i$, $i = 1, \ldots, n$.

a) Calculer $Var(\hat{\beta})$ et donner sa limite lorsque $n \to \infty$.

b) En déduire le comportement asymptotique en moyenne quadratique de $\hat{\beta}_0$ et $\hat{\beta}_1$.

Ex 9.
Nous souhaitons exprimer la hauteur y d'un arbre en fonction de son diamètre x à 1m30 du sol. Pour cela, nous avons mesuré 20 couples diamètre-hauteur et les résultats ci-dessous sont disponibles :

$$\overline{x} = 34.9,\ \overline{y} = 18.34,\ \frac{1}{20}\sum_{i=1}^{20}(x_i - \overline{x})^2 = 28.29,$$

$$\frac{1}{20}\sum_{i=1}^{20}(y_i - \overline{y})^2 = 2.85,\ \frac{1}{20}\sum_{i=1}^{20}(x_i - \overline{x})(y_i - \overline{y}) = 6.26$$

1) On note $\hat{y} = \hat{\beta}_0 + x\hat{\beta}_1$ l'estimation de la droite de régression. Donner l'expression de $\hat{\beta}_0$ et $\hat{\beta}_1$ en fonction des statistiques élémentaires ci-dessus. Calculer $\hat{\beta}_0$ et $\hat{\beta}_1$.

2) Donner une mesure de qualité d'ajustement des données au modèle. Exprimer cette mesure à l'aide des statistiques élémentaires. Calculer et commenter.

3) Tester $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$ pour $j = 0, 1$. Commenter.

Ex 10. *Le coefficient de corrélation multiple*
On considère le modèle de régression

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

où les variables ϵ_i sont centrées, de variance σ^2 et non-corrélées. On pose $Y = (y_1, \ldots, y_n)^T$, $X_k = (x_{k,1}, \ldots, x_{k,n})^T$ et $\mathbf{1} = (1, \ldots, 1)^T$. 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 :

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

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 (géométriquement) 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 ?

Ex 11. *L'interprétation du R^2 comme coefficient de corrélation multiple*
On se place dans un modèle de régression contenant une constante. On définit

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

où *corr* 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

$$corr(Y, X\beta) = \frac{(X\hat{\beta} - \bar{Y})'(X\beta - \bar{X}\beta)}{\|Y - \bar{Y}\| \, \|X\beta - \bar{X}\beta\|},$$

où $\hat{\beta}$ est l'estimateur des 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, $corr(Y, X\beta)^2 \leq R^2$, où R^2 est le coefficient de corrélation multiple de la régression de Y sur X, et que cette borne est atteinte lorsque $\beta = \hat{\beta}$.

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

Ex 12. *Le test de Fisher en pratique*
Dans un modèle de régression linéaire multiple comprenant 5 variables explicatives (dont éventuellement la constante), estimé sur n individus, on considère le test de Fisher de $q \leq 5$ contraintes linéaires sur les coefficients : $H_0 : R\beta = 0$ contre $H_1 : R\beta \neq 0$, où R est une matrice de taille $(q, 5)$.

1) Rappeler la statistique utilisée pour le test précédent.

2) Dans les cas suivants, donner l'expression de la matrice R, la loi suivie par la statistique de test sous H_0 et la démarche pratique pour mettre en oeuvre le test :
i) $H_0 : \beta_1 = 0$; ii) $H_0 : \beta_1 = \beta_2 = \beta_4 = 0$; iii) $H_0 : \beta_2 = \beta_3$; iv) $H_0 : \beta_1 = \beta_2$ et $\beta_2 = 2\beta_3$.

Ex 13. *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 $n \times p$ et ϵ est un vecteur aléatoire de taille n, centré et de matrice de covariance $\sigma^2 I$ (I 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).

Ex 14. (issu du livre "Régression, Théorie et Applications")
Nous voulons expliquer la concentration de l'ozone sur Rennes en fonction des variables T9, T12, Ne9, Ne12 et Vx. Suite à l'estimation du modèle de régression linéaire, les sorties données par R sont (aux points d'interrogation près) :

```
Coefficients :
             Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)       62          10        ?         0
 T9                -4           ?       -5         0
 T12                5        0.75        ?         0
 Ne9             -1.5           1        ?      0.14
 Ne12            -0.5         0.5        ?      0.32
 Vx               0.8        0.15      5.3         0

--
Multiple R-Squared: 0.6666, Adjusted R-squared: 0.6532
Residual standard error: 16 on 124 degrees of freedom
F-statistic: ? on ? and ? DF, p-value: 0
```

1) Retrouver les valeurs manquantes dans la sortie ci-dessus.

2) Rappeler la statistique de test et tester la nullité des paramètres séparément au seuil de 5%.

3) Rappeler la statistique de test et tester la nullité simultanée des paramètres autres que la constante au seuil de 5%.

4) Les variables Ne9 et Ne12 ne semblent pas influentes et nous souhaitons tester la nullité simultanée de β_{Ne9} et β_{Ne12}. Proposer un test permettant de tester ces contraintes et l'effectuer en vous aidant de la sortie R du modèle sans Ne9 et Ne12 suivante :

```
Coefficients :
             Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)       66          11        6         0
 T9                -5           1       -5         0
 T12                6        0.75        8         0
 Vx                 1         0.2        5         0

--
Multiple R-Squared: 0.6525, Adjusted R-squared: 0.6442
Residual standard error: 16.2 on 126 degrees of freedom
```

Ex 15. *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.

Ex 16. *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.

Ex 17. *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_a^2 ?

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 $n \to \infty$.

Ex 18. *Moindres carrés généralisés*
Soit un modèle de régression linéaire multiple

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

où $\beta \in \mathbb{R}^k$, X est une matrice de taille $n \times k$ 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 $var(\epsilon) = \Sigma$ où Σ est une matrice de rang n. 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.

Ex 19. *Estimation dans un modèle ANOVA*
On considère une variable quantitative Y et un facteur A ayant I modalités. On suppose disposer d'un échantillon de n individus répartis en n_i individus dans chaque modalité A_i de A, pour $i = 1, \dots I$. On note $y_{i,k}$ la valeur de Y pour l'individu k appartenant à la modalité A_i de A, et on note $\bar{y}_i$ la moyenne de Y dans A_i.
On considère le modèle ANOVA liant Y à A :

$$y_{i,k} = m + \alpha_i + \epsilon_{i,k}, \tag{1}$$

pour tout $i = 1, \dots, I$, $k = 1, \dots, n_i$.

1) Le modèle précédent peut s'écrire sous la forme $y = X\beta + \epsilon$ où $y = (y_1, \dots, y_n)$ et $\epsilon = (\epsilon_1, \dots, \epsilon_n)$. On suppose que dans cette écriture les individus ont été rangés par modalités de A, i.e. $y = (y_{1,1}, \dots, y_{1,n_1}, y_{2,1}, \dots, y_{2,n_2}, \cdots, y_{I,1}, \dots, y_{I,n_I})'$. Donner la forme de la matrice X et du vecteur β. Pourquoi l'estimation de β dans ce modèle n'est pas possible par les MCO ?

2) On considère la contrainte $m = 0$. Ecrire le modèle (1) avec cette contrainte sous la forme $Y = X\beta + \epsilon$ en précisant la nouvelle matrice X et le nouveau vecteur β. Calculer l'estimateur $\hat{\beta}$ issu des MCO.

3) On considère la contrainte $\alpha_1 = 0$. Ecrire le modèle (1) avec cette contrainte sous la forme $Y = X\beta + \epsilon$ en précisant la nouvelle matrice X et le nouveau vecteur β. Calculer l'estimateur $\hat{\beta}$ issu des MCO. On commencera par déterminer les $\hat{\alpha}_i$ en résolvant directement le problème de minimisation, pour en déduire finalement $\hat{m}$.

4) Montrer que quelle que soit la contrainte choisie précédemment, $\hat{y}_{i,k} = \bar{y}_i$, pour tout $i = 1, \dots, I$ et $k = 1, \dots, n_i$.

5) On note $\mu_i = E(Y|A_i)$ et on désire tester l'effet du facteur A sur Y, c'est à dire $H_0 : \mu_1 = \cdots = \mu_I$. Comment se traduit cette hypothèse nulle sur les paramètres β du modèle (selon chaque contrainte précédemment choisie) ?

6) Montrer que la statistique de Fisher permettant d'effectuer le test précédent sur β, s'écrit (quelle que soit la contrainte sur β choisie initialement) :

$$F = \frac{S_A^2/(I-1)}{S_R^2/(n-I)},$$

où $S_A^2 = \sum_{i=1}^I n_i(\bar{y}_i - \bar{y})^2$ et $S_R^2 = \sum_{i=1}^I \sum_{k=1}^{n_i}(y_{i,k} - \bar{y}_i)^2$. Quelle est la région critique du test au niveau α ? Sous quelle(s) hypothèse(s) ce test est-il valable ?

Ex 20. *Application de l'ANOVA à 1 facteur*
On veut étudier l'impact d'une ancienne mine d'arsenic sur les composantes hydrochimiques et hydrobiologiques d'un réseau hydrographique de Corse. Les mesures ont été faites sur 3 stations : B2, B3 (sur la Bravona) et P2 (sur un affluent la Presa) où est située la mine d'arsenic. Le tableau suivant résume la bioaccumulation de l'arsenic (en $\mu g/g$) sur les branchies des truites capturées pour chaque station. La variance considérée correspond à la variance empirique corrigée.

Station	effectif	moyenne	variance
P2	22	4.83	1.58
B2	21	0.66	0.07
B3	24	0.24	0.02

1) On désire tester l'effet des stations sur la bioaccumulation de l'arsenic. Quelle statistique peut-on utiliser pour mettre en oeuvre le test ? Les hypothèses d'application du test sont-elles réunies ?

2) On propose de transformer les données d'arsenic à l'aide de la fonction $x \mapsto \sqrt{x}$. Les résultats sont fournis dans le tableau suivant. Pourquoi cette transformation rend plus raisonnable un test d'effet des stations sur la teneur en arsenic ? Mettre en oeuvre le test et conclure.

Station	effectif	moyenne	variance
P2	22	2.18	0.08
B2	21	0.8	0.02
B3	24	0.48	0.01

Ex 21. *Régression logistique*
On suppose avoir collecté auprès d'un groupe d'étudiants ayant suivi le module "Régression linéaire et logistique" les variables : X_1 : nombre d'heures à travailler le module (hors enseignements), X_2 : moyenne obtenue au premier semestre, et Y : module validé (oui/non). L'ajustement d'un modèle logistique permettant d'expliquer Y en fonction de X_1 et X_2 a fourni les estimations : $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$ et $\hat{\beta}_2 = 0.4$.

1) Estimer la probabilité qu'un étudiant ayant obtenu 12 de moyenne générale au premier semestre et travaillé 40h son module puisse le valider.

2) Combien d'heures cet étudiant aurait-il dû travailler pour espérer obtenir son module avec une probabilité (estimée) supérieure à 0.8 ?

Ex 22. *Régression logistique*
On considère une variable binaire $Y \sim B(p)$ où $p \in [0, 1]$ et une variable quantitative réelle X. On suppose que la loi de X sachant que $Y = 0$ est $\mathcal{N}(\mu_0, \sigma^2)$ et que la loi de X sachant que $Y = 1$ est $\mathcal{N}(\mu_1, \sigma^2)$. On note $\pi(x) = P(Y = 1|X = x)$.

1) Montrer que $\pi(x) = e^{a+bx}/(1+e^{a+bx})$ pour des constantes a et b que l'on explicitera en fonction des paramètres de l'énoncé.

2) Etant donné l'observation d'un échantillon (X_i, Y_i), $i = 1, \ldots, n$, de couples indépendants distribués comme (X, Y), quel modèle est-il naturel d'utiliser pour modéliser $\pi(x)$. Comment peut-on estimer les paramètres ?

3) On suppose a présent que la loi de X sachant que $Y = 0$ est $\mathcal{N}(\mu_0, \sigma_0^2)$ et la loi de X sachant que $Y = 1$ est $\mathcal{N}(\mu_1, \sigma_1^2)$, où $\sigma_0 \neq \sigma_1$. Montrer qu'un modèle logistique est toujours approprié pour modéliser $\pi(x)$.