Examen Final – Session 1

Durée 2h. Les documents, la calculatrice, les téléphones portables, tablettes, ordinateurs ne sont pas autorisés. Les exercices sont indépendants. La qualité de la rédaction sera prise en compte.

On rappelle que la densité d'un vecteur aléatoire gaussien $\mathbf{Z} \sim \mathcal{N}(\mathbf{m}, \mathbf{\Sigma})$ *de dimension* k*, avec* $\mathbf{\Sigma}$ *supposée définie positive, est donnée par :* $f : \mathbf{z} \mapsto (2\pi)^{-k/2} \det(\mathbf{\Sigma})^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{z}-\mathbf{m})^T \mathbf{\Sigma}^{-1}(\mathbf{z}-\mathbf{m})\right)$*, où* $\det(\mathbf{\Sigma})$ *est le déterminant de la matrice* $\mathbf{\Sigma}$*.*

Exercice 1. Estimation de la variance

1. On se place dans le cadre du modèle de régression multiple gaussien $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ où $\mathbf{Y}$ est un vecteur de $\mathbb{R}^n$, $\mathbf{X}$ est une matrice de taille $n \times p$ de plein rang, $\boldsymbol{\beta}$ un vecteur de $\mathbb{R}^p$ et $\boldsymbol{\epsilon}$ un vecteur aléatoire gaussien de $\mathbb{R}^n$ de variables iid, centrées et de variance σ^2.
   (a) Rappelez (sans le démontrer) la définition et l'expression de l'estimateur $\hat{\boldsymbol{\beta}}$ des moindres carrés. Quelle est l'interprétation géométrique de $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$?

   Voir le cour.

   (b) Donnez la log-vraisemblance $\log f_{\mathbf{Y}}(\mathbf{Y}; \boldsymbol{\beta}, \sigma^2)$ du modèle en fonction de $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$. En déduire l'expression de l'estimateur $\hat{\boldsymbol{\beta}}_{ml}$ du maximum de vraisemblance de $\boldsymbol{\beta}$.

   On a $\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2\mathbf{I}_n)$, d'où:

$$\begin{aligned}\log f_{\mathbf{Y}}\left(\mathbf{Y}; \beta, \sigma^2\right) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(\det(\sigma^2\mathbf{I})) - \frac{1}{2}(\mathbf{Y}-\mathbf{X}\beta)^T(\sigma^2\mathbf{I})^{-1}(\mathbf{Y}-\mathbf{X}\beta) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\|\mathbf{Y}-\mathbf{X}\beta\|^2\end{aligned}$$

   On en déduit que $\hat{\beta}_{ml} = \hat{\beta}$, l'estimateur du maximum de vraisemblance est identique à l'estimateur des moindres carrés.

   (c) En déduire l'expression de $\hat{\sigma}^2_{ml}$ l'estimateur du maximum de vraisemblance de la variance σ^2.

   Par dérivation par rapport à σ^2, on trouve (voir le cours):

$$\hat{\sigma}^2_{ml} = \frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}{n}$$

   (d) Montrez que cet estimateur est biaisé. Proposez un estimateur alternatif non biaisé. *[Indication] On pourra utiliser, après l'avoir démontrée, l'égalité suivante : pour un vecteur aléatoire* $\mathbf{Z}$ *ayant des moments d'ordre un et deux, on a* $\mathbb{E}[\|\mathbf{Z}\|^2] = tr(\mathbb{V}ar[\mathbf{Z}] + \mathbb{E}[\mathbf{Z}]\mathbb{E}[\mathbf{Z}]^T)$*.*

   La relation se montre en utilisant la linéarité de la trace et de l'espérance :

$$\begin{aligned}\mathbb{E}[\|\mathbf{Z}\|^2] = \mathbb{E}[\mathbf{Z}^T\mathbf{Z}] = \mathbb{E}[\mathrm{tr}(\mathbf{Z}^T\mathbf{Z})] = \mathbb{E}[\mathrm{tr}(\mathbf{Z}\mathbf{Z}^T)] \\ = \mathrm{tr}(\mathbb{E}[\mathbf{Z}\mathbf{Z}^T]) = \mathrm{tr}(\mathbb{V}ar[\mathbf{Z}] + \mathbb{E}[\mathbf{Z}]\mathbb{E}[\mathbf{Z}]^T).\end{aligned}$$

   On l'applique au vecteur $\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. $\hat{\boldsymbol{\epsilon}} = \mathbf{P}^{\perp}\mathbf{Y}$ est la projection de $\mathbf{Y}$ sur l'orthogonal de l'espace engendré par les colonnes de $\mathbf{X}$, donc $\hat{\boldsymbol{\epsilon}} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{P}^{\perp})$, et :

$$\mathbb{E}[\|\hat{\boldsymbol{\epsilon}}\|^2] = \mathrm{tr}(\mathbb{V}ar[\hat{\boldsymbol{\epsilon}}]) = \sigma^2\mathrm{tr}(\mathbf{P}^{\perp}) = \sigma^2(n-p).$$

   On en déduit que $\mathbb{E}[\hat{\sigma}^2_{ml}] = \sigma^2\frac{n-p}{n}$. Un estimateur non biaisé est donné par : $\hat{\sigma}^2 = \frac{\|\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}\|^2}{n-p}$.

2. L'objectif de cette question est de montrer que l'estimateur non biaisé de la variance rappelé ci-dessus peut être obtenu en maximisant une certaine fonction de vraisemblance dite "restreinte" (procédure REML, pour *restricted maximum likelihood*). On se place dans le cadre de la question précédente.

   (a) Soit $\mathbf{P}$ la matrice de projection orthogonale sur l'espace engendré par les colonnes de $\mathbf{X}$, et $\mathbf{P}^\perp = \mathbf{I}_n - \mathbf{P}$. Quelle est la loi du vecteur $\mathbf{Y}' = \mathbf{P}^\perp \mathbf{Y}$? Montrez que cette loi ne dépend pas de $\boldsymbol{\beta}$, et que la matrice de variance de $\mathbf{Y}'$ n'est pas inversible. Pouvez-vous écrire directement la densité du vecteur $\mathbf{Y}'$?

   $\mathbf{Y}' = \mathbf{P}^\perp \mathbf{Y}$ est un vecteur gaussien comme transformation linéaire d'un vecteur gaussien. De plus,

   $$\mathbb{E}[\mathbf{Y}'] = \mathbf{P}^\perp \mathbb{E}[\mathbf{Y}] = \mathbf{P}^\perp \mathbf{X}\boldsymbol{\beta} = \mathbf{0};$$

   et

   $$\mathbb{V}ar[\mathbf{Y}'] = \mathbf{P}^\perp \mathbb{V}ar[\mathbf{Y}](\mathbf{P}^\perp)^T = \mathbf{P}^\perp(\sigma^2\mathbf{I})(\mathbf{P}^\perp)^T = \sigma^2\mathbf{P}^\perp.$$

   Donc $\mathbf{Y}' \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{P}^\perp)$. La loi de $\mathbf{Y}'$ ne dépend que de σ^2 et pas de $\boldsymbol{\beta}$, sa vraisemblance dépend donc uniquement de σ^2. Comme $\mathbf{P}^\perp$ n'est pas inversible, on ne peut pas écrire directement la densité de $\mathbf{Y}'$.

   (b) La procédure REML se base sur des *contrastes* : l'idée est de trouver une matrice $\mathbf{K}$ de taille $n \times (n-p)$ de plein rang, telle que $\mathbf{K}^T\mathbf{X} = \mathbf{0}$, et de considérer le vecteur $\mathbf{Z} = \mathbf{K}^T\mathbf{Y}$. Montrez qu'au moins une telle matrice $\mathbf{K}$ existe. Quelle est la dimension de $\mathbf{Z}$? La matrice $\mathbf{P}^\perp$ est-elle une matrice de contrastes valide ?
   [Indication] On pourra utiliser le théorème du rang à l'application linéaire associée à $\mathbf{X}^T$.

   On considère la matrice $\mathbf{X}^T$, et $f : \mathbb{R}^n \to \mathbb{R}^p$ l'application linéaire associée, telle que $\forall\ \mathbf{v} \in \mathbb{R}^n, f(\mathbf{v}) = \mathbf{X}^T\mathbf{v}$. Comme $\mathbf{X}$ est de plein rang par hypothèse, on a $\mathrm{rg}(f) = p$. D'après le théorème du rang, la dimension du noyau de f est donc égale à $n-p$. On peut donc trouver $n-p$ vecteurs $\mathbf{k}_1, \ldots, \mathbf{k}_{n-p}$ de $\mathbb{R}^n$ linéairement indépendants tels que pour tout $1 \leq i \leq n-p$, $\mathbf{X}^T\mathbf{k}_i = \mathbf{0}$. En notant $\mathbf{K}$ la matrice de taille $n \times n-p$ dont les vecteurs colonnes sont les $\mathbf{k}_i$. Alors $\mathrm{rg}(\mathbf{K}) = n-p$, et $\mathbf{X}^T\mathbf{K} = \mathbf{0}$, i.e. $\mathbf{K}^T\mathbf{X} = \mathbf{0}$.

   (c) Quelle est la loi de $\mathbf{Z} = \mathbf{K}^T\mathbf{Y}$? De quels paramètres dépend-elle ? Donnez l'expression de $\log f_{\mathbf{Z}}(\mathbf{Z}; \sigma^2)$ la log-vraisemblance de $\mathbf{Z}$, en fonction de $\mathbf{Y}$ et σ^2.

   On a $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{K}^T\mathbf{K})$ est un vecteur gaussien de dimension $n-p$, avec $\mathbf{K}^T\mathbf{K}$ inversible, car $\mathbf{K}$ est de plein rang. La loi de $\mathbf{Z}$ ne dépend que du paramètre de variance σ^2. Donc :

   $$\begin{aligned}\log f_{\mathbf{Z}}\left(\mathbf{Z}; \sigma^2\right) &= -\frac{n-p}{2}\log(2\pi) - \frac{1}{2}\log(\det(\sigma^2\mathbf{K}^T\mathbf{K})) - \frac{1}{2}\mathbf{Z}^T(\sigma^2\mathbf{K}^T\mathbf{K})^{-1}\mathbf{Z} \\ &= -\frac{n-p}{2}\log(2\pi\sigma^2) - \frac{1}{2}\log(\det(\mathbf{K}^T\mathbf{K})) - \frac{1}{2\sigma^2}\mathbf{Y}^T\mathbf{K}(\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T\mathbf{Y}\end{aligned}$$

   (d) Montrez que $\mathbf{Q} = \mathbf{K}(\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T$ est une matrice de projection orthogonale (i.e., une matrice symétrique idempotente). Quelle est la valeur de la trace de $\mathbf{Q}$?

   On vérifie facilement que $\mathbf{Q}^2 = \mathbf{Q} = \mathbf{Q}^T$, donc $\mathbf{Q}$ est une matrice de projection orthogonale. On en déduit que sa trace est égale à son rang : $\mathrm{tr}(\mathbf{Q}) = n-p$.

   (e) Soit $\mathbf{T} = \mathbf{P}^\perp - \mathbf{Q}$. Montrez que $\mathbf{T}$ est une matrice de projection orthogonale, et en déduire la valeur de $\mathrm{tr}(\mathbf{T}\mathbf{T}^T)$. En déduire que $\mathbf{Q} = \mathbf{P}^\perp$.

$\mathbf{T}$ est symétrique, et on vérifie que $\mathbf{T}^2 = \mathbf{T}$, en exploitant le fait que $\mathbf{K}^T\mathbf{X} = \mathbf{0}$. On en déduit

$$\mathrm{tr}(\mathbf{T}\mathbf{T}^T) = \mathrm{tr}(\mathbf{T}^2) = \mathrm{tr}(\mathbf{T}) = \mathrm{tr}(\mathbf{I}) - \mathrm{tr}(\mathbf{P}) - \mathrm{tr}(\mathbf{Q}) = n - p - (n-p) = 0.$$

Or $\mathrm{tr}(\mathbf{T}\mathbf{T}^T) = \sum_{1\leq i,k\leq n} T_{ik}^2$, donc $\mathrm{tr}(\mathbf{T}\mathbf{T}^T) = 0$ implique que $\mathbf{T} = \mathbf{0}$, i.e. $\mathbf{Q} = \mathbf{I} - \mathbf{P} = \mathbf{P}^\perp$.

(f) En déduire que : $\log f_{\mathbf{Z}}(\mathbf{Z}; \sigma^2) = -\frac{n-p}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|\mathbf{P}^\perp\mathbf{Y}\|^2$.

On déduit des questions précédentes :

$$\begin{aligned}
\log f_{\mathbf{Z}}\left(\mathbf{Z}; \sigma^2\right) &= -\frac{n-p}{2}\log(2\pi\sigma^2) - \frac{1}{2}\log(\det(\mathbf{K}^T\mathbf{K})) - \frac{1}{2\sigma^2}\mathbf{Y}^T\mathbf{K}(\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T\mathbf{Y} \\
&= -\frac{n-p}{2}\log(2\pi\sigma^2) - 0 - \frac{1}{2\sigma^2}\mathbf{Y}^T\mathbf{Q}\mathbf{Y} \\
&= -\frac{n-p}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\mathbf{Y}^T\mathbf{P}^\perp\mathbf{Y} \\
&= -\frac{n-p}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\mathbf{Y}^T(\mathbf{P}^\perp)^T\mathbf{P}^\perp\mathbf{Y} \\
&= -\frac{n-p}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|\mathbf{P}^\perp\mathbf{Y}\|^2.
\end{aligned}$$

(g) L'estimateur du maximum restreint $\hat{\sigma}^2_{reml}$ de la variance peut se définir comme l'estimateur maximisant la vraisemblance $\log f_{\mathbf{Z}}(\mathbf{Z}; \sigma^2)$ de $\mathbf{Z}$. Donnez-en l'expression. Cet estimateur est-il biaisé ? Cet estimateur dépend-il du choix de $\mathbf{K}$?

D'après la formule précédente, on trouve directement :

$$\hat{\sigma}^2_{reml} = \frac{\|\mathbf{P}^\perp\mathbf{Y}\|^2}{n-p} = \frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}{n-p} = \hat{\sigma}^2.$$

Cet estimateur est le même qu'à la question précédente, et est donc non biaisé. L'estimateur REML de la variance ne dépend pas du choix des contrastes, tant que $\mathbf{K}$ est de plein rang et $\mathbf{K}^T\mathbf{X} = \mathbf{0}$.

3. Dans la question précédente, la vraisemblance restreinte était définie comme la vraisemblance d'un vecteur de contrastes, qui ne dépend pas du vecteur des coefficients $\boldsymbol{\beta}$. Dans cette question, on montre une définition alternative de la vraisemblance restreinte, basée sur une vraisemblance intégrée en $\boldsymbol{\beta}$. On définit ici la vraisemblance intégrée par :

$$L\left(\mathbf{Y}; \sigma^2\right) = \int_{\boldsymbol{\beta}\in\mathbb{R}^p} f_{\mathbf{Y}}\left(\mathbf{Y}; \boldsymbol{\beta}, \sigma^2\right) d\boldsymbol{\beta}.$$

(a) La vraisemblance intégrée $L(\mathbf{Y}; \sigma^2)$ dépend-elle de $\boldsymbol{\beta}$?

Non, puisque que l'on intègre sur ce paramètre.

(b) Montrez que l'on peut écrire :

$$L\left(\mathbf{Y}; \sigma^2\right) = (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{P}^\perp\mathbf{Y}\|^2\right)\int_{\boldsymbol{\beta}\in\mathbb{R}^p}\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2\right)d\boldsymbol{\beta}.$$

On a, en utilisant le théorème de Pythagore :

$$\begin{aligned}
\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 &= \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}\|^2 = \|\mathbf{P}^\perp\mathbf{Y} + \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2 \\
&= \|\mathbf{P}^\perp\mathbf{Y}\|^2 + \|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2.
\end{aligned}$$

La formule demandée se déduit directement de cette décomposition.

(c) En déduire :

$$L\left(\mathbf{Y};\sigma^2\right) = (2\pi\sigma^2)^{-(n-p)/2}\det(\mathbf{X}^T\mathbf{X})^{-1/2}\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{P}^{\perp}\mathbf{Y}\|^2\right).$$

On a : $\|\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|^2 = (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T(\mathbf{X}^T\mathbf{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$, donc, en notant $\boldsymbol{\Sigma} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, on obtient :

$$\begin{aligned}
I(\sigma^2) &:= \int_{\boldsymbol{\beta}\in\mathbb{R}^p} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|^2\right) d\boldsymbol{\beta} \\
&= \int_{\boldsymbol{\beta}\in\mathbb{R}^p} \exp\left(-\frac{1}{2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\right) d\boldsymbol{\beta} \\
&= (2\pi)^{p/2}\det(\boldsymbol{\Sigma})^{1/2}\int_{\boldsymbol{\beta}\in\mathbb{R}^p}(2\pi)^{-p/2}\det(\boldsymbol{\Sigma})^{-1/2}\exp\left(-\frac{1}{2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\right) d\boldsymbol{\beta} \\
&= (2\pi)^{p/2}(\sigma^2)^{p/2}\det(\mathbf{X}^T\mathbf{X})^{-1/2}\times 1
\end{aligned}$$

Donc :

$$\begin{aligned}
L\left(\mathbf{Y};\sigma^2\right) &= (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{P}^{\perp}\mathbf{Y}\|^2\right)\int_{\boldsymbol{\beta}\in\mathbb{R}^p}\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|^2\right) d\boldsymbol{\beta} \\
&= (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{P}^{\perp}\mathbf{Y}\|^2\right)(2\pi)^{p/2}(\sigma^2)^{p/2}\det(\mathbf{X}^T\mathbf{X})^{-1/2} \\
&= (2\pi\sigma^2)^{-(n-p)/2}\det(\mathbf{X}^T\mathbf{X})^{-1/2}\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{P}^{\perp}\mathbf{Y}\|^2\right).
\end{aligned}$$

(d) Conclure. Quel estimateur obtenez vous en maximisant cette quantité?

À une constante près ($\det(\mathbf{X}^T\mathbf{X})^{-1/2}$), cette expression est la même qu'à la question précédente. On obtient donc le même estimateur de la variance, non biaisé.

Exercice 2. Histoire de l'Art

On s'intéresse ici à la représentation des artistes dans les manuels d'histoire de l'art, telle qu'étudiée par Holland Stam, 2022, *Quantifying Art Historical Narratives*. Plus précisément, on se concentre sur l'édition de 2001 du manuel *Gardner's Art Through the Ages*, qui est une référence du domaine, et dont la première édition date de 1927.

Table 1: Extrait de quelques lignes du jeu de données.

	space_ratio	moma_count	nationality	gender
Édouard Manet	1.6282475	0	French	Male
Eugène Delacroix	1.7315482	1	French	Male
Francisco Goya	1.3889939	3	Spanish	Male
Frida Kahlo	0.4654877	0	Other	Female
Jacques-Louis David	2.0529051	0	French	Male
Joan Miró	0.6513492	17	Spanish	Male
Pablo Picasso	2.7728655	30	Spanish	Male
Walker Evans	0.2215694	12	American	Male

Pour les 165 artistes cités dans l'ouvrage, on considère les variables suivantes :

- `space_ratio`: aire relative, figures comprises, accordée à l'artiste, relativement à la surface totale d'une page. Par exemple, un peu moins d'une demi page est consacrée à Kahlo, alors que Manet bénéficie de plus d'une page et demie (voir la Table 1 ci-dessus).

- `moma_count` représente le nombre d'expositions consacrées à l'artiste en question au Museum of Modern Art (MoMA) de New-York, avant 2001.
- `nationality` et `gender` donnent la nationalité et le sexe de chaque artiste. Les artistes dont la nationalité n'est ni *French*, *Spanish*, *British*, *American* ou *German*, qui ne représentent que 10% des artistes dans le jeu de données, sont marqués comme *Other*.

1. On commence par se poser la question d'un éventuel biais de genre dans l'importance accordée aux artistes dans le manuel. On execute les commandes suivantes dans `R`:

```
fit1 <- lm(space_ratio ~ gender, data = artists)
summary(fit1)
```

```
##
## Call:
## lm(formula = space_ratio ~ gender, data = artists)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31487 -0.18724 -0.10208  0.02214  2.27953
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.43502    0.06865   6.337  2.2e-09 ***
## genderMale   0.05831    0.07453   0.782    0.435
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3433 on 163 degrees of freedom
## Multiple R-squared:  0.003741, Adjusted R-squared:  -0.002371
## F-statistic: 0.6121 on 1 and 163 DF,  p-value: 0.4351
```

(a) Explicitez le modèle utilisé. Quelle est la forme de la matrice $\mathbf{X}$ des prédicteurs si on la met sous la forme standard utilisée dans `R` ?

On utilise le modèle suivant, pour $1 \leq i \leq n$, avec $n = 165$:

$$\begin{cases} Y_i = \beta_0 + \epsilon_i & \text{si l'individu i est une femme} \\ Y_i = \beta_0 + \beta_1 + \epsilon_i & \text{sinon,} \end{cases}$$

avec les ϵ_i iid centrés gaussiens de variance σ^2. La matrice $\mathbf{X}$ est de taille $n \times 2$, avec la première colonne ne comportant que des 1 (intercept), et la seconde des 1 uniquement pour les hommes, et des zéros sinon.

(b) Quelle est la moyenne de la surface relative accordée aux femmes ? Aux hommes ?

La surface relative moyenne accordées aux femmes est de 0.43502 (coefficient de l'intercept). La surface relative moyenne accordées aux hommes est de $0.43502 + 0.05831 = 0.49333$.

(c) Explicitez les hypothèses des tests de Student sur les coefficients, et du test de Fisher global. Interprétez les valeurs de la sortie ci-dessus.

Le premier test de Student correspond à $\mathcal{H}_0$: La moyenne des surfaces pour les femmes est nulle vs $\mathcal{H}_1$: différente de zéro. On constate que la p-valeur associées est très petite, on rejette donc l'hypothése nulle.
Le second test de Student et le test de Ficher global testent : $\mathcal{H}_0$: Les moyennes des surfaces pour les femmes et les hommes est la même vs $\mathcal{H}_1$: les moyennes sont différentes. Pour ces deux tests, on a la même p-valeur de 0.435, on ne peut donc pas rejetter l'hypothèse nulle.

(d) A partir de cette analyse, peut-on conclure que les femmes et les hommes sont aussi bien représentés dans le manuel ? On pourra s'appuyer sur la table de contingence suivante pour préciser la réponse (de manière qualitative).

```
table(artists$gender, artists$nationality)
```

```
##
##          Other American British French German Spanish
##   Female     5       14       0      3      3       0
##   Male      35       38      13     39     11       4
```

Parmis les artistes représentés, il ne semble pas y avoir de différence de moyenne dans la surface allouée à un homme ou une femme. Cependant, ce test ne prend pas en compte le biais de sélection : il y a plus d'hommes représentés dans le manuel que de femmes (140 hommes pour 25 femmes). La surface totale dévolue à des artistes hommes reste donc bien supérieure. De plus, le R^2 de la régression est très faible, si bien que le modèle n'est pas très explicatif des données, et l'inclusion d'autres covariables serait nécessaire.

2. On ajoute maintenant la représentation de chaque nationalité.

```
fit2 <- lm(space_ratio ~ nationality * gender, data = artists)
fit2
```

```
##
## Call:
## lm(formula = space_ratio ~ nationality * gender, data = artists)
##
## Coefficients:
##                       (Intercept)              nationalityAmerican
##                          0.408609                        -0.038051
##               nationalityBritish                nationalityFrench
##                          0.019505                         0.160263
##                nationalityGerman                nationalitySpanish
##                          0.237426                         0.897856
##                        genderMale   nationalityAmerican:genderMale
##                         -0.005382                         0.023480
##    nationalityBritish:genderMale     nationalityFrench:genderMale
##                                NA                         0.069608
##     nationalityGerman:genderMale    nationalitySpanish:genderMale
##                         -0.204821                                NA
```

```
anova(fit2)
```

```
## Analysis of Variance Table
##
## Response: space_ratio
##                      Df  Sum Sq Mean Sq F value   Pr(>F)
## nationality           5  4.3754 0.87508  9.1747 1.15e-07 ***
## gender                1  0.0012 0.00122  0.0127   0.9103
## nationality:gender    3  0.1179 0.03930  0.4121   0.7446
## Residuals           155 14.7839 0.09538
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) Explicitez le modèle utilisé, et donnez la matrice $\mathbf{X}$ des prédicteurs correspondante, avec la paramétrisation utilisée dans `R`. Combien y-a-t-il de paramètres à estimer ?

Soit $Y_{i,k,l}$ la variable représentant la surface relative pour l'artiste i, de nationalité $k \in \{Other, American, British, French, German, Spanish\}$ et de genre $l \in \{Female, Male\}$.

On écrit :

$$Y_{i,k,l} = \mu + \alpha_k + \beta_l + \gamma_{kl} + \epsilon_i,$$

avec ϵ_i iid gaussiens centrés de variance σ^2, et en imposant les contraintes suivantes :

$\alpha_1 = 0$	le coefficient associé à "Other" est nul
$\beta_1 = 0$	le coefficient associé à "Female" est nul
$\gamma_{1,l} = 0$	les coefficients d'intéraction entre "Other" et le sexe sont nuls
$\gamma_{k,1} = 0$	les coefficients d'intéraction entre "Female" et la nationalité sont nuls

Ce qui conduit $6 \times 2 = 12$ paramètres à estimer.
La matrice des prédicteurs s'écrit, en notant $\mathbf{1}_{S_k}$ (respectivement, $\mathbf{1}_{T_l}$) le vecteur de taille n qui vaut 1 si l'individu i est de nationalité k (respectivement, de sexe l), et 0 sinon, et $\odot$ le produit de Hadamard (terme à terme) :

$$\mathbf{X} = (\mathbf{1}, \mathbf{1}_{S_2}, \mathbf{1}_{S_3}, \mathbf{1}_{S_4}, \mathbf{1}_{S_5}, \mathbf{1}_{S_6}, \mathbf{1}_{T_2}, \mathbf{1}_{S_2} \odot \mathbf{1}_{T_2}, \mathbf{1}_{S_3} \odot \mathbf{1}_{T_2}, \mathbf{1}_{S_4} \odot \mathbf{1}_{T_2}, \mathbf{1}_{S_5} \odot \mathbf{1}_{T_2}, \mathbf{1}_{S_6} \odot \mathbf{1}_{T_2}).$$

(b) Certains paramètres ne sont pas estimés par `R`, qui retourne des `NA`. Quels sont ces paramètres ? Pourquoi ne peut-on pas les estimer ?

Les paramètres associés à l'intéraction entre le sexe et les nationalités *British* et *Spanish* ne sont pas estimés. On constate qu'il n'y a aucune femme artiste de ces nationalités présentes dans le jeu de données. On a donc $\mathbf{1}_{S_{British}} \odot \mathbf{1}_{T_{Male}} = \mathbf{1}_{S_{British}}$ et $\mathbf{1}_{S_{Spanish}} \odot \mathbf{1}_{T_{Male}} = \mathbf{1}_{S_{Spanish}}$, si bien que la matrice $\mathbf{X}$ n'est pas de plein rang. On ne peut donc pas estimer ces effets d'interaction.

(c) Quels sont les tests effectués dans la table d'anova ? Précisez les statistiques de test associés à chacun, et leur loi sous $\mathcal{H}_0$. Concluez sur les paramètres du modèle.

Il s'agit d'une table d'anova de type I, où les facteurs sont ajoutés un par un. Les tests réalisés sont :

- $\mathcal{H}_0$: seul l'intercept est non nul vs $\mathcal{H}_1$: au moins un des α_k est non nul.
  Statistique de test : $F_0 = \frac{R(\alpha|\mu)/5}{\hat{\sigma}^2} \underset{\mathcal{H}_0}{\sim} \mathcal{F}^5_{155}$, où $\hat{\sigma}^2 = \frac{RSS}{155}$ est l'estimateur global de la variance, et $R(\alpha|\mu) = RSS_{\alpha,\mu} - RSS_\mu$ est la différence des carrés résiduels du modèle avec uniquement l'intercept et du modèle avec l'intercept et le facteur nationalité.
- $\mathcal{H}_0$: seuls l'intercept et les coefficients α_k sont non nuls vs $\mathcal{H}_1$: au moins un des β_l est non nul.
  Statistique de test : $F_1 = \frac{R(\beta|\mu,\alpha)/1}{\hat{\sigma}^2} \underset{\mathcal{H}_0}{\sim} \mathcal{F}^1_{155}$.
- $\mathcal{H}_0$: seuls l'intercept et les coefficients α_k et β_l sont non nuls vs $\mathcal{H}_1$: au moins un des γ_{kl} est non nul.
  Statistique de test : $F_2 = \frac{R(\gamma|\mu,\alpha,\beta)/3}{\hat{\sigma}^2} \underset{\mathcal{H}_0}{\sim} \mathcal{F}^3_{155}$.

Pour les carrés totaux, comme on n'a estimé que 10 paramètres (et non 12, voir la question précédente), on a bien $165 - 10 = 155$ degrés de libertés. On rejette $\mathcal{H}_0$ pour le premier test, mais on ne peut pas le rejeter pour les deux suivants. Le facteur nationalité semble donc important, mais pas le facteur sexe, ni leurs intéractions.

3. On cherche maintenant à expliquer la variable de surface relative sur la page du manuel par le nombre d'expositions au MoMA consacré à l'artiste. On considère ici que la variable `moma_count` est une variable continue.

```r
fit3 <- lm(space_ratio ~ moma_count * nationality, data = artists)
anova(fit3)
```

```
## Analysis of Variance Table
##
## Response: space_ratio
##                          Df  Sum Sq Mean Sq F value    Pr(>F)
## moma_count                1  0.5336 0.53361  6.2276  0.013639 *
## nationality               5  3.9404 0.78808  9.1973 1.134e-07 ***
## moma_count:nationality    5  1.6945 0.33891  3.9552  0.002114 **
## Residuals               153 13.1098 0.08569
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) Explicitez le modèle utilisé, et donnez la matrice $\mathbf{X}$ des prédicteurs correspondante, avec la paramétrisation utilisée dans R. Combien y-a-t-il de paramètres à estimer ?

Soit $Y_{i,k}$ la variable représentant la surface relative pour l'artiste i, de nationalité $k \in \{Other, American, British, French, German, Spanish\}$. On écrit :

$$Y_{i,k} = \mu + \alpha_k + (\beta + \gamma_k) \times x_i + \epsilon_i,$$

avec ϵ_i iid gaussiens centrés de variance σ^2, $x_i = \texttt{moma\_count}_i$ la covariable pour l'individu i, et en imposant les contraintes suivantes :

$$\begin{aligned} \alpha_1 = 0 &\qquad \text{le coefficient associé à "Other" est nul} \\ \gamma_1 = 0 &\qquad \text{les coefficients de pente associé à "Other" est nul} \end{aligned}$$

Ce qui conduit $6 + 6 = 12$ paramètres à estimer.
La matrice des prédicteurs s'écrit, en notant $\mathbf{1}_{S_k}$ le vecteur de taille n qui vaut 1 si l'individu i est de nationalité k, et 0 sinon, $\mathbf{x}$ le vecteur des n covariables, et $\odot$ le produit de Hadamard (terme à terme) :

$$\mathbf{X} = (\mathbf{1}, \mathbf{1}_{S_2}, \mathbf{1}_{S_3}, \mathbf{1}_{S_4}, \mathbf{1}_{S_5}, \mathbf{1}_{S_6}, \mathbf{1}_{S_2} \odot \mathbf{x}, \mathbf{1}_{S_3} \odot \mathbf{x}, \mathbf{1}_{S_4} \odot \mathbf{x}, \mathbf{1}_{S_5} \odot \mathbf{x}, \mathbf{1}_{S_6} \odot \mathbf{x}).$$

(b) Est-ce que tous les paramètres sont bien estimés par R dans ce modèle ?

Oui, les 12 paramètres sont bien estimés. On peut le voir en regardant le nombre de degrés de libertés associés aux RSS globaux, qui est bien de $153 = 165 - 12$ (et non 155 comme précédement).

(c) Quels sont les tests effectués dans la table d'anova ? Précisez les statistiques de test associés à chacun, et leur loi sous $\mathcal{H}_0$. Concluez sur les paramètres du modèle.

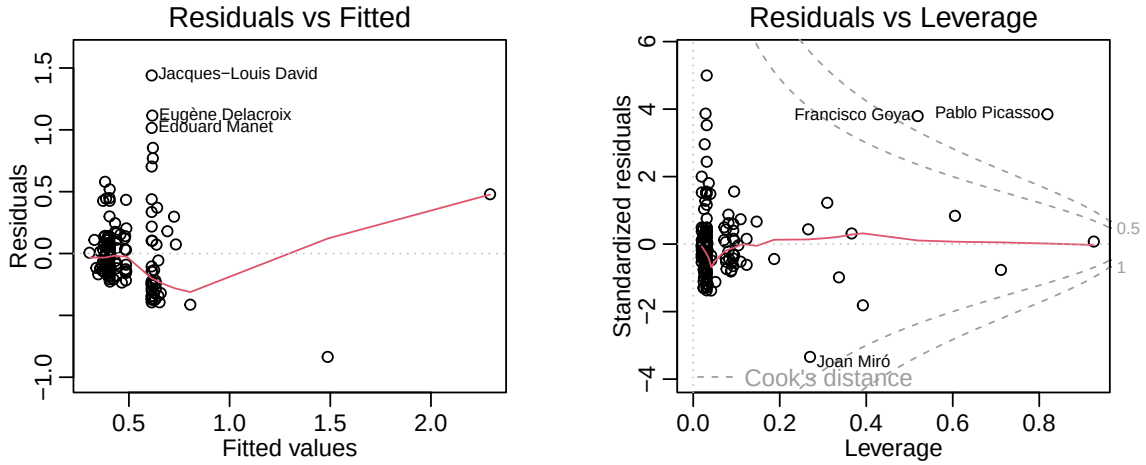
Il s'agit d'une table d'anova de type I, où les facteurs sont ajoutés un par un. Les tests réalisés sont :

- $\mathcal{H}_0$: seul l'intercept est non nul vs $\mathcal{H}_1$: β est aussi non nul.
  Statistique de test : $F_0 = \frac{R(\beta|\mu)/1}{\hat{\sigma}^2} \underset{\mathcal{H}_0}{\sim} \mathcal{F}^1_{153}$, où $\hat{\sigma}^2 = \frac{RSS}{153}$ est l'estimateur global de la variance, et $R(\beta|\mu) = RSS_{\beta,\mu} - RSS_\mu$ est la différence des carrés résiduels du modèle avec uniquement l'intercept et du modèle avec l'intercept et le coefficient de pente global.
- $\mathcal{H}_0$: seuls l'intercept et la pente β sont non nuls vs $\mathcal{H}_1$: au moins un des α_k est non nul.
  Statistique de test : $F_1 = \frac{R(\alpha|\mu,\beta)/1}{\hat{\sigma}^2} \underset{\mathcal{H}_0}{\sim} \mathcal{F}^5_{153}$.

– $\mathcal{H}_0$: seuls l'intercept et les coefficients β et α_k sont non nuls vs $\mathcal{H}_1$: au moins un des γ_k est non nul.
Statistique de test : $F_2 = \frac{R(\gamma|\mu,\beta,\alpha)/5}{\hat{\sigma}^2} \underset{\mathcal{H}_0}{\sim} \mathcal{F}^3_{153}$.

On rejette $\mathcal{H}_0$ au niveau de 5% pour les trois tests. Il y a donc bien une ralation significative entre la surface relative sur la page et le nombre d'expositions au MoMA, avec un effet de nationalité, et une pente différente par nationalité.

4. On trace les graphiques suivants pour la régression de la question précédente.



(a) A quoi correspondent ces deux graphiques ? Explicitez toutes les axes et variables ("Fitted Values", "Residuals", "Standardized Residuals", "Leverage", "Cooks's distance"). Quelle est la loi des résidus et des résidus standardisés représentés ici ?

Les variables sont :
– "Fitted Values" : $\hat{y}_{i,k} = \hat{\mu} + \hat{\alpha}_k + (\hat{\beta} + \hat{\gamma}_k) \times x_i$ sont les valeurs ajustées par le modèle.
– "Residuals" : $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I}_n - \mathbf{P}^{\mathbf{X}})\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \sigma^2(\mathbf{I}_n - \mathbf{P}^{\mathbf{X}}))$, avec $\mathbf{P}^{\mathbf{X}}$ la matrice de projection sur l'espace engendré par les colonnes de $\mathbf{X}$, sont les résidus.
– "Standardized Residuals" : $t_i = \frac{\hat{\epsilon}_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}$ où $h_{ii} = \mathbf{P}^{\mathbf{X}}_{ii}$, sont les résidus standardisés.
En général, on ne connaît pas leur loi, car $\hat{\epsilon}_i$ et $\hat{\sigma}^2$ ne sont pas indépendants.
– "Leverage" : h_{ii} la diagonale de la matrice de projection, quantifie l'effet de l'observation i sur sa propre prédiction.
– "Cooks's distance" : $C_i = \frac{1}{p\hat{\sigma}^2}\left(\hat{\boldsymbol{\theta}}_{(-i)} - \hat{\boldsymbol{\theta}}\right)^T (\mathbf{X}^T\mathbf{X}) \left(\hat{\boldsymbol{\theta}}_{(-i)} - \hat{\boldsymbol{\theta}}\right) = \frac{h_{ii}(y_i - \hat{y}_i^P)^2}{p\hat{\sigma}^2} = \frac{h_{ii}}{p(1-h_{ii})^2}\frac{\hat{\epsilon}_i^2}{\hat{\sigma}^2} = \frac{h_{ii}}{p(1-h_{ii})}t_i^2$, en notant $\boldsymbol{\theta} = (\mu, \alpha_k, \beta, \gamma_k)$ le vecteur des 12 paramètres du modèle, quantifie l'effet de l'observation i sur les coefficients de la régression.

(b) Interprétez les graphiques. Vous pourrez vous aider de l'extrait donné Table 1.

Dans le premier graphique, on constate que quelques points ont des résidus élevés. Il s'agit en effet de peintres bénéficiant d'une grande couverture dans le livre, mais de peu d'expositions au MoMA : ils sont donc "outliers" par rapport au modèle. Ils ne semblent cependant pas avoir une distance de Cook très élevée, et n'ont donc pas d'impact majeur sur l'estimation des coefficients. On constate de plus que la courbe de lissage par polynomes locaux n'est pas droite, ce qui peut indiquer que certains effets importants n'ont pas été pris en compte dans le modèle.
Dans le second graphique, trois peintres semblent avoir une distance de Cook élevée, tous espagnols. Picasso, avec près de trois pages dans le manuel et 30 exposition, est particulièrement représenté. Il est possible qu'il tire vers le haut l'estimation des coefficients. Il

serait interessant de regarder une régression de laquelle ce peintre 'hors norme" est retirée.