# 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 $\sigma^2$.
   (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 $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ ?
   (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 $\boldsymbol{\beta}$.
   (c) En déduire l'expression de $\hat{\sigma}^2_{ml}$ l'estimateur du maximum de vraisemblance de la variance $\sigma^2$.
   (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* $\mathbb{E}[\|\mathbf{Z}\|^2] = tr(\mathbb{V}ar[\mathbf{Z}] + \mathbb{E}[\mathbf{Z}]\mathbb{E}[\mathbf{Z}]^T)$.

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 $\mathbf{Y}'$ ?
   (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* $\mathbf{X}^T$*.*
   (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 $\sigma^2$.
   (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 $\mathbf{Q}$ ?
   (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 $\mathbf{Q} = \mathbf{P}^\perp$.
   (f) En déduire que : $\log f_{\mathbf{Z}}(\mathbf{Z}; \sigma^2) = -\frac{n-p}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|\mathbf{P}^\perp\mathbf{Y}\|^2$.
   (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 $\mathbf{K}$ ?
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 $L\left(\mathbf{Y};\sigma^2\right)$ dépend-elle de $\boldsymbol{\beta}$ ?

(b) Montrez que l'on peut écrire :

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

(c) En déduire :

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

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

## 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 ?

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

(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.

(d) À 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
```

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 ?

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

(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.

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.

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

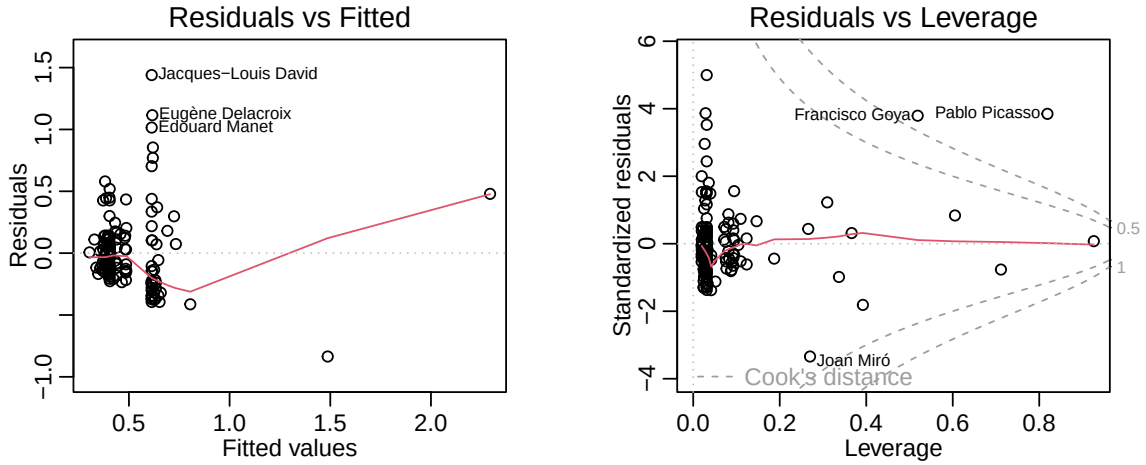
```
## Analysis of Variance Table
##
## Response: space_ratio
##                          Df  Sum Sq Mean Sq F value    Pr(>F)
## moma_count                1  0.5336 0.53361  6.2276  0.013639 *
## nationality               5  3.9404 0.78808  9.1973 1.134e-07 ***
## moma_count:nationality    5  1.6945 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 ?

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

(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.

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 ?

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