Fiche bilan SAE

Nom de la SAE	**SAE Modèle linaire**		Semestre / Période	Semestre 4
Volume horaire consacré par l'étudiant	Avec enseignant	3h	En autonomie	3h
Coéquipiers :	Sami Said		Mohamed Belarbi	
	Franck Tankapanya			

Sujet spécifique	**Expliquer ou prédire une variable quantitative à partir de plusieurs facteurs**
Objectifs	• Importer et Comprendre les Données • Examiner les Relations entre Variables • Modéliser les Relations à travers la Régression Linéaire Multiple • Optimiser et Évaluer le Modèle

Fiche bilan SAE

Livrables	

```
options(repos = c(CRAN = "https://cran.rstudio.com"))
install.packages("readxl")
```

```
## Installation du package dans 'C:/Users/Feky/AppData/Local/R/win-library/4.3'
## (car 'lib' n'est pas spécifié)
```

```
## le package 'readxl' a été décompressé et les sommes MD5 ont été vérifiées avec succès
##
## Les packages binaires téléchargés sont dans
##  C:\Users\Feky\AppData\Local\Temp\RtmpSsNG0x\downloaded_packages
```

```
library(readxl)
```

```
## Warning: le package 'readxl' a été compilé avec la version R 4.3.3
```

```
foot<-read_excel("C:/FRANCK/taff/football_2009_2016.xlsx" )
data=subset(foot, select=c(Goals,Club))
```

Proposer un modele goals en fonction du club

```
mod_foot <- lm(Goals~Club -1, data = data)
summary(mod_foot)
```

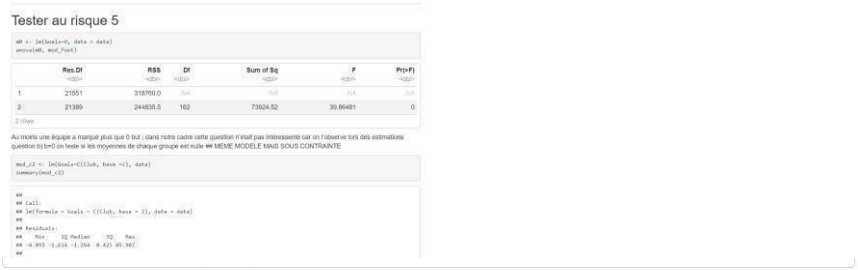
```
##
## Call:
## lm(formula = Goals ~ Club - 1, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.893 -1.614 -1.264  0.425 45.907
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## ClubAC Ajaccio             1.2644     0.3627   3.486 0.000492 ***
## ClubAC Milan               1.8916     0.2144   8.822  < 2e-16 ***
## ClubAlaves                 1.6400     0.6767   2.424 0.015373 *
## ClubAlmeria                1.4571     0.1582   4.613 1.01e-05 ***
## ClubAngers                 1.3621     0.4683   3.066 0.002172 **
## ClubArles-Avignon          0.6563     0.5981   1.097 0.272568
## ClubArsenal                2.4978     0.2256  11.076  < 2e-16 ***
## ClubAston Villa            1.4288     0.2461   5.805 6.53e-09 ***
## ClubAtalanta               1.3125     0.2263   5.800 6.49e-09 ***
## ClubAthletic Bilbao        2.0245     0.2369   8.547  < 2e-16 ***
## ClubAtletico Madrid        2.4688     0.2369  10.388  < 2e-16 ***
## ClubAugsburg               1.4063     0.2634   5.338 9.48e-08 ***
## ClubAuxerre                1.6795     0.3831   4.384 1.17e-05 ***
## ClubBarcelona              4.0927     0.2263  17.120  < 2e-16 ***
## ClubBari                   1.1538     0.4376   2.750 0.005971 **
## ClubBariti                 0.0000     1.3833   0.000 1.000000
## ClubBayer Leverkusen       2.4375     0.2442   9.983  < 2e-16 ***
```

Estimation des espérances de Y

```
summary(mod_foot)$coefficients[,1]
```

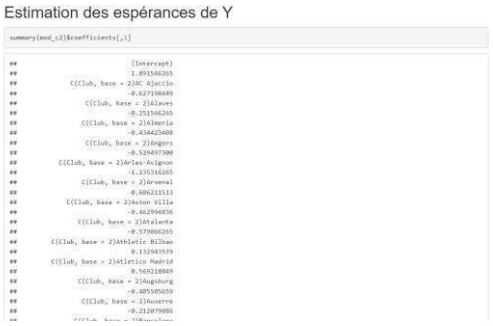
```
##             ClubAC Ajaccio               ClubAC Milan
##                  1.2643678                  1.8915663
##                 ClubAlaves                ClubAlmeria
##                  1.6400000                  1.4571429
##                 ClubAngers          ClubArles-Avignon
##                  1.3620690                  0.6562500
##               ClubArsenal             ClubAston Villa
##                  2.4977778                  1.4285714
##               ClubAtalanta         ClubAthletic Bilbao
##                  1.3125000                  2.0245098
##         ClubAtletico Madrid               ClubAugsburg
##                  2.4607843                  1.4060606
##               ClubAuxerre              ClubBarcelona
##                  1.6794872                  4.0926829
##                  ClubBari                 ClubBariti
##                  1.1538462                  0.0000000
##        ClubBayer Leverkusen          ClubBayern Munich
##                  2.4375000                  3.3730570
##             ClubBirmingham             ClubBlackburn
##                  1.4509804                  1.4204545
```

Fiche bilan SAE



On a mis alpha égale a 0, donc le Club de l'AC Milan est passé en Intercept. Ce qui a comparé les estimations des nombres de buts des différents clubs en la comparant a celle de l'AC Milan0. On observe donc que les clubs ayant une estimation positive ont marqué plus de buts en moyenne que L'AC Milan. Tandis que ceux qui ont une estimation négative ont marqué moins de buts en moyenne que l'Ac Milan. Mais si l'on veut déduire que la différence de buts entre l'Ac Milan et un des clubs est significative on doit avoir une p-valeur inférieure a 0,05. On voit donc que le Fc Barcelone a marqué en moyenne 2,2 buts par match de plus que l'AC Milan et la p-valeur est de 5.41e-12, donc on rejette H0. On peut donc affirmer que la différence de buts marqués en moyenne entre le Fc Barcelone et l'Ac Milan est significative. Dans le cas inverse, si on s'intéresse au club de Cordoba on peut voir que l'estimation de buts moyens marqués par Cordoba est de 0.77 but de moins que l'Ac Milan. Puis la p-valeur est de 0.03 donc on rejette H0, on peut onc affirmer que la différence de buts entre cordoba et l'Ac Milan est significative.

la contrainte est alpha=0 intercept nest plus la



Tester au risque 5

```
m1 <- lm(Goals~1, data)
anova(m1, mod_c2)
```

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	21550	251544.4	*NA*	*NA*	*NA*	*NA*
2	21389	244835.5	161	6708.914	3.640344	9.897197e-49

2 rows

Au moins un club a mis plus ou moins de buts par rapport au Milan AC avec une p valeur de 2.2e-16 on rejette H0. La pvaleur extrêmement faible de 2.2e-16 obtenue lors de la comparaison entre le modèle avec seulement l'intercept et le modèle incluant la variable 'Club' en tant que facteur suggère que l'ajout du facteur 'Club' apporte une contribution significative à l'explication de la variance dans les objectifs marqués. Ainsi, nous rejetons l'hypothèse nulle selon laquelle tous les clubs marquent plus ou moins le même nombre de buts par rapport au Milan AC, avec un risque d'erreur de 5%.

Bilan de la SAE
(reproduire le tableau autant de fois que de compétences mobilisées dans la SAÉ)

Compétence	**Modéliser les données dans un cadre statistique**
Apprentissages critiques sollicités	Comprendre l'impact du type de données sur le choix de la modélisation à mettre en œuvre
	Apprécier les limites de validité et les conditions d'application d'un modèle
	Réaliser l'importance de la mise en œuvre d'une procédure de test statistique pour valider ou non une hypothèse
Composantes essentielles à respecter	En utilisant le modèle de données adapté aux besoins

	En maîtrisant la qualité du modèle
	En choisissant le modèle adapté à la situation

Ma démarche

Savoirs / connaissances	Savoir-faire	Savoir-être
Statistiques descriptives, corrélation, régression linéaire, sélection de variables.	• Manipuler • Programmer • Interpréter	• Travail d'équipe, • Persévérance dans L'analyse de données • Communication claire des résultats.

Evaluation du résultat

- Ce que je trouve bien réalisé, pourquoi ?

Fiche bilan SAE

La méthodologie complète de modélisation, depuis l'analyse initiale jusqu'à la sélection de variables, permet une compréhension approfondie des dynamiques salariales. La collaboration en équipe a également enrichi l'approche analytique.

- Ce que je n'ai pas bien compris ; ce qui serait à améliorer pour une prochaine fois : pourquoi ? comment ?

L'un des aspects qui pourrait ne pas être immédiatement clair est la manière dont les hypothèses sous-jacentes à l'ANOVA ont été vérifiées. L'ANOVA repose sur certaines hypothèses importantes comme l'homogénéité des variances et la normalité des distributions des groupes. Pour une prochaine fois, il serait bénéfique d'inclure des tests spécifiques (par exemple, le test de Levene pour l'homogénéité des variances) ou des visualisations (QQ-plots pour la normalité) pour valider ces hypothèses. Ceci est crucial pour s'assurer que les conclusions tirées du modèle sont valides.

Eléments de preuve, ce que je peux montrer *(Choisir des éléments précis à mettre annexe)*

1) Extrait de code

Fiche bilan SAE

```
install.packages("readxl")
library(readxl)
foot<-read_excel("D:/BUT/2eme annee/modele linaire/football_2009_2016.xlsx" )
data=subset(foot, select=c(Goals,Club))

# proposer un modele goals en fonction du club les moyennes par groupe de nos variables ne sont pas nul
mod_foot <- lm(Goals~Club -1, data = data)
summary(mod_foot)

#1 la contrainte est b=0 intercept nest plus la


#estimation des esp´erance de Y

summary(mod_foot)$coefficients[,1]

# Tester au risque 5

m0 <- lm(Goals~0, data = data)
anova(m0, mod_foot)

# au moin une equipe a marque plus que 0 but ; dans notre cadre cette question était pas intéressente ca
#b=0 on teste si les moyenne de chaque groupe est nulle

#PARTIE 2

#MEME MODELE MAIS SOUS CONTRAINTE

mod_c2 <- lm(Goals~C(Club, base =2), data)
summary(mod_c2)

# on a mis alpha égale a 0, ducoup le Club de l'AC Milan est passer en Intercept.
#Ce qui a compare les estimations des nombres de buts
#des différents clubs en la comparant a celle de l'AC Milan.
#On observe donc que les club ayant une estimation positive on marqué plus de but en moyenne que L'AC Mil
#Tandis que ceux qui ont une estimation négative on marqué moins de but en moyenne que l'Ac Milan.
#Mais si l'on veut déduire que la différence de but entre l'Ac Milan et un des clubs est significative on
# On voit donc que le Fc Barcelone a marqué en moyenne 2,2 but par match de plus que l'AC Milan et la p-v
#On peut donc affirmer que la différence de but marqué en moyenne entre le Fc Barcelone et l'Ac Milan est
# Dans le cas inverse si on s'intéresse au club de Cordoba on peut voir que l'estimation de but moyen mar
#Puis la p-valeur est de 0.03 donc on rejette h0 est on peut affirmer que la différence de but entre cord




  #1 la contrainte est alpha=0 intercept nest plus la

#estimation des esp´erance de Y

summary(mod_c2)$coefficients[,1]

# Tester au risque 5
```

2) Extrait de sorti de code réalisé :

Fiche bilan SAE

```
##
## Call:
## lm(formula = Goals ~ Club - 1, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.093 -1.614 -1.264  0.425 45.907
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## ClubAC Ajaccio               1.2644     0.3627   3.486 0.000492 ***
## ClubAC Milan                 1.8916     0.2144   8.822  < 2e-16 ***
## ClubAlaves                   1.6400     0.6767   2.424 0.015373 *
## ClubAlmeria                  1.4571     0.3302   4.413 1.02e-05 ***
## ClubAngers                   1.3621     0.4443   3.066 0.002172 **
## ClubArles-Avignon            0.6563     0.5981   1.097 0.272548
```

```
m0 <- lm(Goals~0, data = data)
anova(m0, mod_foot)
```

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	21551	318760.0	*NA*	*NA*	*NA*	*NA*
2	21389	244835.5	162	73924.52	39.86481	0

2 rows