Fiche bilan SAE

Nom de la SAE	**SAE Modèle Linéaire**		semestre / Période	S4
volume horaire consacré par l'étudiant	avec enseignant	3h	en autonomie	2h
coéquipiers :	Ayoub Errahmani		Franck Tankapanya	
	Sami Said			

Sujet spécifique	**Expliquer ou prédire une variable quantitative à l'aide de plusieurs facteurs**
Objectifs	Importer et Comprendre les Données Examiner les relations entre les variables Modéliser les relations à travers la régression Linéaire Multiple Optimiser et évaluer le modèle
Livrables	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
## -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
## ClubArsenal                2.4978     0.2256  11.074  < 2e-16 ***
## ClubAston Villa            1.4286     0.2461   5.805 6.53e-09 ***
## ClubAtalanta               1.3125     0.2261   5.806 6.49e-09 ***
## ClubAthletic Bilbao        2.0245     0.2369   8.547  < 2e-16 ***
## ClubAtletico Madrid        2.4608     0.2369  10.388  < 2e-16 ***
## ClubAugsburg               1.4061     0.2634   5.338 9.48e-08 ***
## ClubAuxerre                1.6795     0.3831   4.384 1.17e-05 ***
```

Fiche bilan SAE

Les moyennes par groupe de nos variables ne sont pas nulles 2.2e-16 avec test de Fisher. La contrainte est b=0 intercept n'est plus la.

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 
##            ClubBlackpool                ClubBochum 
```

Tester au risque 5

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

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

2 rows

Au moins une équipe a marqué plus que 0 but ; dans notre cadre cette question n'était pas intéressente car on l'observe lors des estimations question b) b=0 on teste si les moyennes de chaque groupe est nulle ## MEME MODELE MAIS SOUS CONTRAINTE

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

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

```
##
## Call:
## lm(formula = Goals ~ C(Club, base = 2), 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|)
## (Intercept)                          1.891566   0.214409   8.822  < 2e-16
## C(Club, base = 2)AC Ajaccio         -0.627198   0.421359  -1.489 0.136630
## C(Club, base = 2)Alaves             -0.251566   0.709819  -0.354 0.723036
## C(Club, base = 2)Almeria            -0.434423   0.393685  -1.103 0.269832
## C(Club, base = 2)Angers             -0.529497   0.493284  -1.073 0.283099
## C(Club, base = 2)Arles-Avignon      -1.235316   0.635361  -1.944 0.051875
## C(Club, base = 2)Arsenal             0.606212   0.311200   1.948 0.051431
## C(Club, base = 2)Aston Villa        -0.462995   0.326399  -1.418 0.156061
## C(Club, base = 2)Atalanta           -0.579066   0.311565  -1.859 0.063102
## C(Club, base = 2)Athletic Bilbao     0.132944   0.319504   0.416 0.677346
## C(Club, base = 2)Atletico Madrid     0.569218   0.319504   1.782 0.074834
## C(Club, base = 2)Augsburg           -0.485506   0.339626  -1.430 0.152866
## C(Club, base = 2)Auxerre            -0.212079   0.439004  -0.483 0.629036
## C(Club, base = 2)Barcelona           2.201117   0.319075   6.898 5.41e-12
## C(Club, base = 2)Bari               -0.737720   0.471249  -1.565 0.117491
## C(Club, base = 2)Bariti             -1.891566   3.390098  -0.558 0.576872
```

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

Estimation des espérances de Y

```
summary(mod_c2)$coefficients[,1]
```

```
##                          (Intercept)
##                          1.891566265
##            C(Club, base = 2)AC Ajaccio
##                         -0.627198449
##                C(Club, base = 2)Alaves
##                         -0.251566265
##               C(Club, base = 2)Almeria
##                         -0.434423408
##                C(Club, base = 2)Angers
##                         -0.529497300
##          C(Club, base = 2)Arles-Avignon
```

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

	En s'adaptant aux spécificités (données, enjeux, méthodes) d'un domaine d'application particulier (santé, marketing, assurance, qualité, socio-démographie...)

Ma démarche

Savoirs / connaissances	Savoir-faire	Savoir-être
Statistiques descriptives, corrélation, régression linéaire, sélection de variables.	Observation Analyse Interprétation	Travail d'équipe Communication

Evaluation du résultat

- Ce que je trouve bien réalisé, pourquoi ?
  Tout à été bien réalisé car nous sommes parvenus au résultat final souhaitée

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

Lors de cette SAE, l'objectif était de découvrir et d'exploiter un nouveau modèle d'analyse de plusieurs variables nommé ANOVA. Ayant choisi un jeu de données sur le football, notre contrainte ici était d'adapter notre modèle en fonction de notre jeu de données. Nous avons alors créé un modèle en fonction des clubs et de leur nombre de buts marqués. Notre modèle nous a alors permis d'effectuer une comparaison entre un club précis et tous les autres en fonction de leur nombre de buts marqués et d'en déduire des conclusions. Le club ciblé ici était l'AC Milan ; grâce à cela, nous avons pu voir quels sont les clubs qui marquent plus ou moins de buts que l'AC Milan.

Eléments de preuve, ce que je peux montrer
(Choisir des éléments précis à mettre annexe)
1)

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

Estimation des espérances de Y

```
summary(mod_c2)$coefficients[,1]
```

```
##                            (Intercept)
##                              1.891566265
##              C(Club, base = 2)AC Ajaccio
##                             -0.627198449
##                 C(Club, base = 2)Alaves
##                             -0.251566265
##                C(Club, base = 2)Almeria
##                             -0.434423408
##                 C(Club, base = 2)Angers
##                             -0.529497300
##           C(Club, base = 2)Arles-Avignon
```

2)

Tester au risque 5

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

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<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%.