Devoir 3

Aloui Mohamed, Jafuno Douba

14 décembre 2018

Question 1: Tailles des effectifs

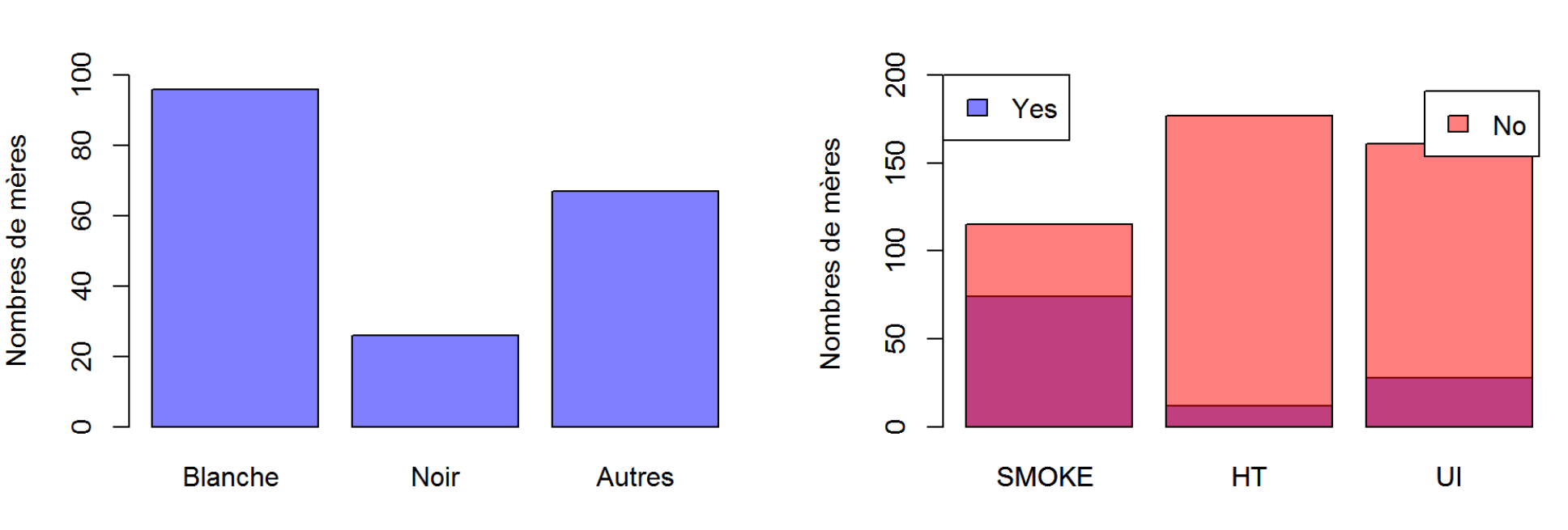


Tableau des effectifs de RACE

RACE	Effectif
Blanche	96
Noir	26
Autres	67

Tableau des effectifs de SMOKE HT et UI

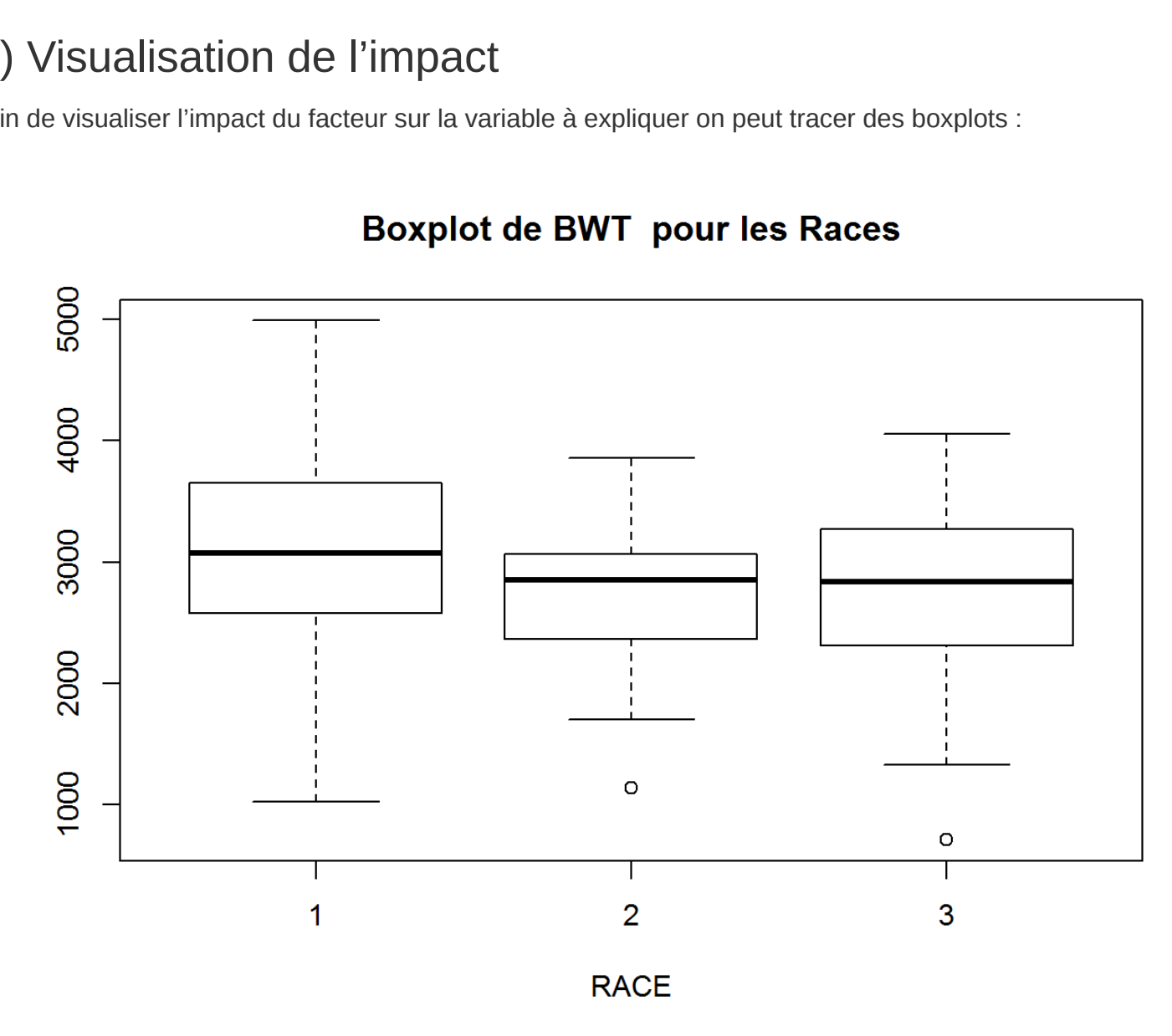
.	Smoke	HT	UI
Yes	74	12	28
No	115	177	161

Grace aux graphiques et aux tableaux ci dessus on voit bien que pour chaque variable la taille des groupes est différentes, autrement dit le plan d'expérience n'est pas équilibré en chaque variable.

Question 2: Modèle à un facteur

1) Visualisation de l'impact

Afin de visualiser l'impact du facteur sur la variable à expliquer on peut tracer des boxplots :



On observe que les trois boxplots sont légèrement different avec une echelle du poids à la naissance plus grande quand il s'agit d'une femme de la race 1 que pour les deux autres races.

2) RACE variable qualitative

On considère le modèle: $\mathrm{BWT}_i = \beta_0 + \beta_1 \mathrm{RACE} + \varepsilon$ ici la variable RACE est interpréter comme une variable quantitative. , Mais en réalité la variable RACE est de type qualitative, ce modèle n'a donc aucun sens d'un point de vue d'interprétation.

Observons alors ce qui se passe en la définissant comme une variable qualitative, c'est à dire en la déclarant comme une variable de type factor.

3) Tableau des coefficients de la Fonction lm

.	Estimate	Std. Error	t value	Pr(>abs(t)
(Intercept)	3103.74	72.88	42.586	< 2e-16
RACE2	-384.05	157.87	-2.433	0.01594
RACE3	-299.72	113.68	-2.637	0.00908

On observe qu'il s'agit d'un modèle de type analyse de la variance à 1 facteur en effet le tableau des **Coefficients** est composé de trois lignes (**Intercept**, **RACE2** et **RACE3**).

Avec les observations $\mathrm{BWT}_{i,j}$ on considère donc le modèle suivant:

$$\mathrm{BWT}_{ij} = \mu_1 + \alpha_i + \varepsilon_{ij}, \qquad i = 1, \ldots, I \quad j = 1, \ldots, n_i.$$

où l'indice i indique la race (1, 2 ou 3), μ_i les moyennes de ses groupes avec la contrainte $\alpha_1 = 0$, on a ε_{ij}~ $N(0, \sigma^2)$. Autrement dit, les α_i pour $i = 2, \ldots, I$ vérifient

$$\alpha_i = \mu_i - \mu_1,$$

on voit alors que tout les autres paramètres (pour $i > 1$) dépendent du niveau $i = 1$

La colonne **Estimate** contient les estimateurs de μ_1 (**Intercept**: 3103.74), de α_2 (**RACE2**: -384.05) et de α_3 (**RACE3**:-299.72).

La colonnes **t value** représente la valeur observée de la statistique du test d'hypothèse H_0:μ_1=α_2=α_3= 0 contre $H_1 : \mu_1 \neq 0, \alpha_2 \neq 0$ et $\alpha_3 \neq 0$

Les tests de significativité des coefficients (**Pr(>|t|)**) donnent ici des p-valeurs inférieures à 0.05: 0.01594 et 0.00908

La très petite p-valeur (2e-16) pour la constante indique que la constante (l'intercept) doit apparaître dans le modèle.

Tout cela montre bien que les coefficients **RACE 2** et **RACE 3** sont significatives sur le poids de naissance du bébé (**BWT**) ainsi l'hypothèse nulle H_0 de chacun des tests est rejetée au profit de l'hypothèse alternative H_1

4) Tableau de l'analyse de la variance

.	Df	Sum Sq	Mean Sq	F value	Pr(>F)
RACE	2	5070608	2535304	4.9719	0.007879
Residuals	186	94846445	509927		

Df signifie degree of freedom. **Sum Sq** et **Mean Sq** renseignent les différentes SCE et CME.

La colonne **F value** donne la valeur de la statistique de test F en sur ce jeu de données.

La dernière colonne **Pr(>F)** est la plus importante car elle permet de conclure pour notre test.

En fait, **Pr(>F)** est la p-value du test de l'absence d'effet dû au facteur (RACE) dans le cas du poids de naissance du bébé (BWT).

Rappelons qu'en général, si la p-value est petite (< 0.01), on rejette H_0 et on décide H_1. **Ici, on voit que la p-value du test de l'absence d'effet dû à la RACE est de 0.007879, ce qui est clairement en faveur de l'hypothèse H_1.**

Il n'y a pas de doute que la RACE de la mère influence significativement sur le poids de naissance du bébé.

Question 3: Lm avec les variables SMOKE, UI et HT

Coefficients pour Smoke Modèle 1:

.	Estimate	Std. Error	t value	Pr(>abs(t)
(Intercept)	3054.96	66.93	45.642	< 2e-16
SMOKEY	-281.71	106.97	-2.634	0.00916

Coefficients pour HT Modèle 2:

.	Estimate	Std. Error	t value	Pr(>abs(t)
(Intercept)	2972.31	54.35	54.685	<2e-16
HTY	-435.56	215.71	-2.019	0.0449

Coefficients pour UI Modèle 3:

.	Estimate	Std. Error	t value	Pr(>abs(t)
(Intercept)	3030.61	55.25	54.857	< 2e-16
UIY	-580.18	143.53	-4.042	7.73e-05

Pour le modèle 1 (resp 2, resp 3) On voit qu'il s'agit d'un modèle de type analyse de la variance à 1 facteur en effet le tableau des **Coefficients** est composé de deux lignes (**Intercept**, **SMOKEY**(resp **HTY**, resp **UIY**).

Pour Smoke resp(HT),resp(UI) on considère le modèle 1 (resp 2, resp 3) suivant:

$$\mathrm{BWT}_{ij} = \mu_1 + \alpha_i + \varepsilon_{ij}, \qquad i = 1, \ldots, I \quad j = 1, \ldots, n_i.$$

où l'indice i indique (1 pour No et 2 pour Yes), μ_i les moyennes de ses groupes avec la contrainte $\alpha_1 = 0$, on a ε_{ij}~ $N(0, \sigma^2)$. Autrement dit, les α_i pour $i = 2, \ldots, I$ vérifient

$$\alpha_i = \mu_i - \mu_1,$$

on voit alors que le paramètre pour $i = 2(Y)$ dépend du niveau $i = 1(N)$

La colonne **Estimate** contient les estimateurs de μ_1 **Intercept**: 3054.96 (resp 2972.31, resp 3030.61), et α_2 **SMOKEY**: -281.71 (resp **HTY**: -435.56, resp **UIY**: -580.18)

La colonnes **t value** représente la valeur observée de la statistique du test d'hypothèse H_0:μ_1=α_2= 0 contre $H_1 : \mu_1 \neq 0$ et $\alpha_2 \neq 0$

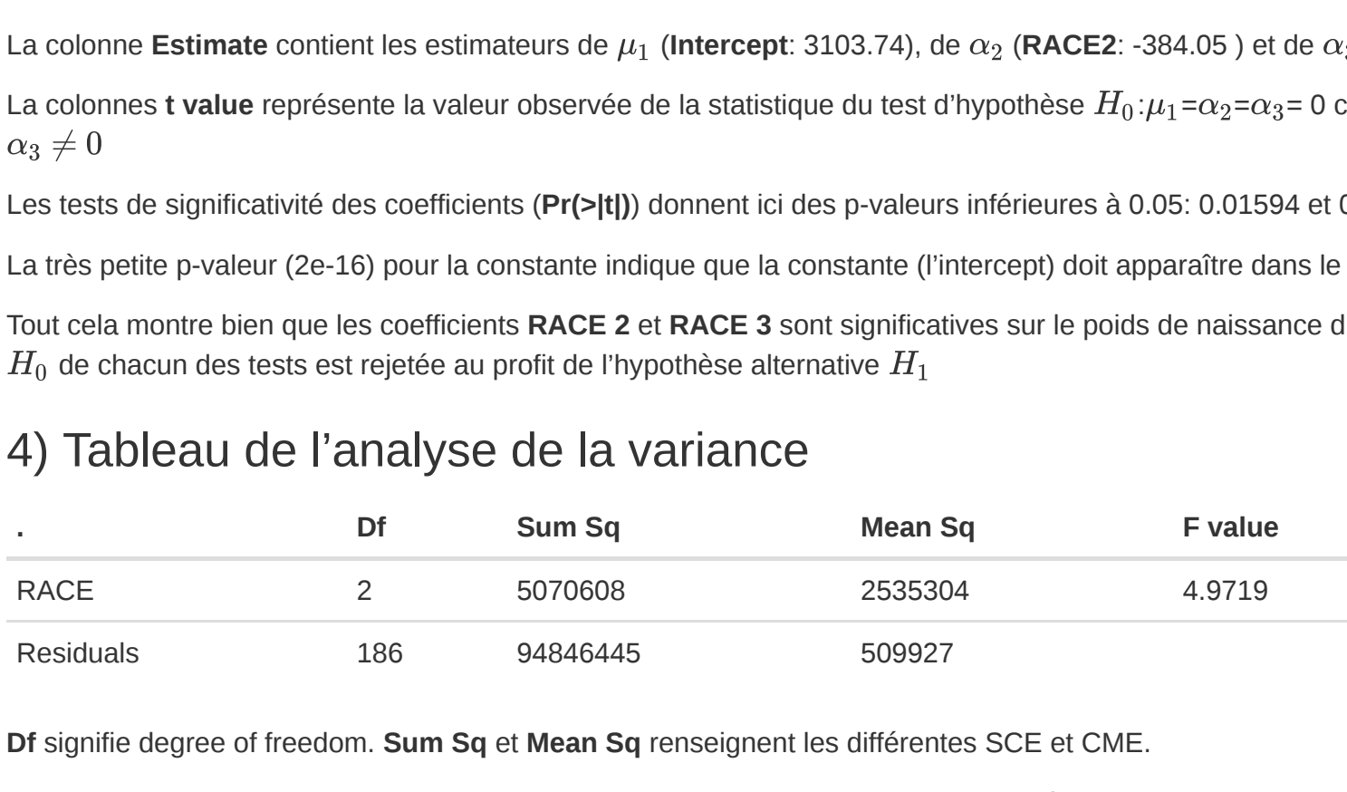
Les tests de significativité des coefficients (**Pr(>|t|)**) donnent ici des p-valeurs inférieures à 0.05: 0.00916 (resp 0.0449, resp 7.73e-05)

La très petite p-valeur (2e-16) pour la constante indique que la constante (l'intercept) doit apparaître dans le modèle.

Tout cela montre bien que **le test du coefficient SMOKEY (resp HTY , resp UIY) est significative sur le poids de naissance du bébé (BWT)** ainsi l'hypothèse nulle H_0 de chacun des tests est rejetée au profit de l'hypothèse alternative

Question 4: Modèle à deux facteurs

1) Analyse de l'impact des facteurs



Nous pouvons voir que la courbe des moyennes de poids de naissance associées aux non-fumeurs est plus élevée que celle des fumeurs quelque soit la race de la mère. **Il semble donc avoir un impact du tabagisme (SMOKE) sur le poids de naissance**.

2) Tableau des coefficients de lm

On considère comme dans le cours le modèle à deux facteurs avec interaction donné par

$$\mathrm{BWT}_{ijk} = \mu_{11} + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij},$$

avec les contraintes $\alpha_1 = 0$, $\beta_1 = 0$, $\gamma_{1j} = 0, j = 1, \ldots, J$ et $\gamma_{i1} = 0, i = 1, \ldots, I$. On le tableau des coefficients suivant :

.	Estimate	Std. Error	t value	Pr(>abs(t)
(Intercept)	3428.8	103.0	33.278	< 2e-16
SMOKEY	-600.0	140.0	-4.286	2.94e-05
RACE2	-574.3	199.5	-2.878	0.00448
RACE3	-614.5	138.2	-4.445	1.52e-05
SMOKEY:RACE2	249.5	309.0	0.807	0.42047
SMOKEY:RACE3	542.9	258.9	2.097	0.03734

La colonne Estimate contient les estimateurs de μ_{11} (Intercept) ,α_2 (SMOKEY) (Yes du facteur SMOKE) , β_2, β_3 (RACE2, RACE3) et γ_{22}, γ_{23} (SMOKEY:RACE2,SMOKEY:RACE3)

On observe que pour chaque test de signification des coefficients nous avons une p-value inférieur à .05 sauf pour le coefficient **SMOKEY:RACE2** (0.42047) dont le test n'a pas vraiment d'impact significatif sur le poids à la naissance et **on ne peut pas rejeter l'hypothèse nulle**.

3) Analyse de la variance

D'après le cours le choix des contraintes sur les paramètres n'a pas d'importance, car les sous-espaces vectoriels intervenant dans l'analyse de la variance sont les mêmes quelques soient les contraintes d'identifiabilité. Ainsi, on obtient:

.	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SMOKE	1	3573406	3573406	7.6503	0.0062584
RACE	2	8768299	4384149	9.3861	0.0001316683
SMOKE:RACE	2	2097537	1048769	2.2453	0.1088037
Residuals	183.0	85477810	467092		

Comme dans le cas à un facteur la dernière colonne donne les p-valeurs des différents tests, on voit que les tests sur les facteurs **SMOKE** et **RACE** sont significatifs (car 0.0062584 et 0.0001316683 sont inférieur à 0.05).

Mais la p-valeur du test d'interaction **SMOKE:RACE** est de 0.1088037 ce qui est trop élevé .

Donc, dans ce cas on ne peut pas rejeter l'hypothèse nulle au niveau habituel de 0.05 et nous considérons donc un modèle sans interaction.

4) Analyse de la variance sans interaction

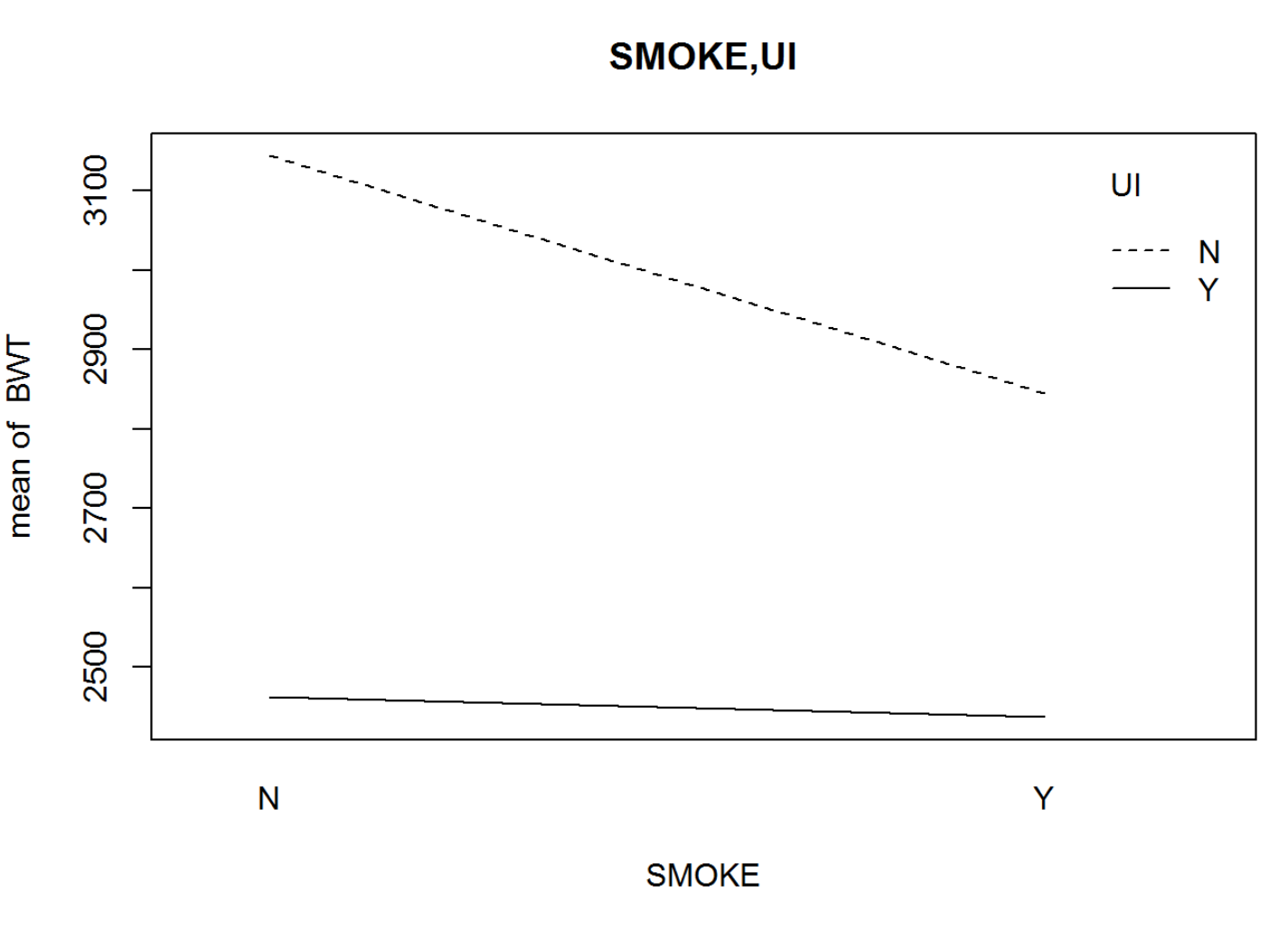
On définit donc le modèle sans interaction nous obtenons le tableau suivant:

.	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SMOKE	1	3573406	3573406	7.5487	0.0065995
RACE	2	8768299	4384149	9.2614	0.0001468
Residuals	185	87575348	473380		

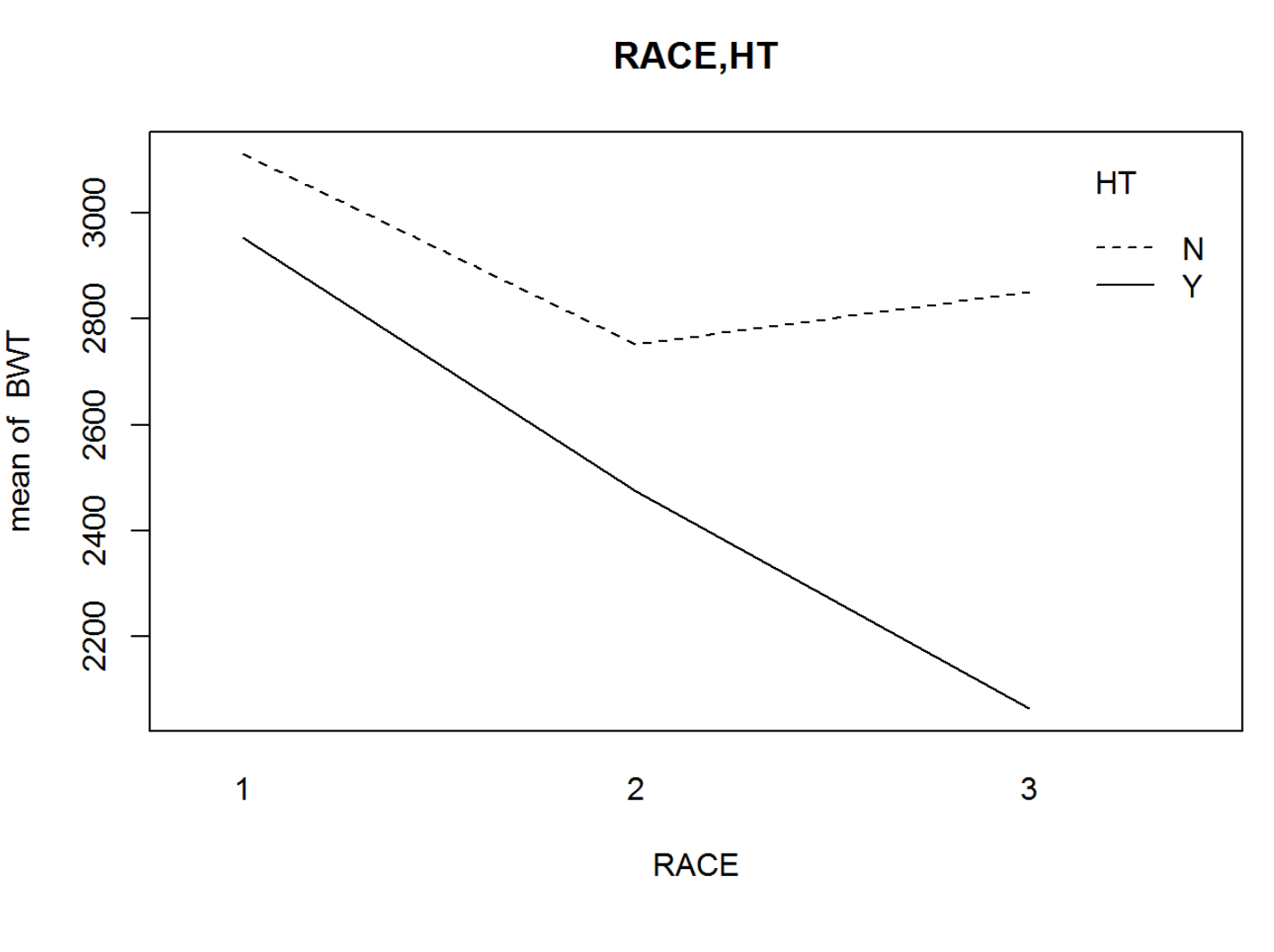
Question 5

1) Interpretation graphique

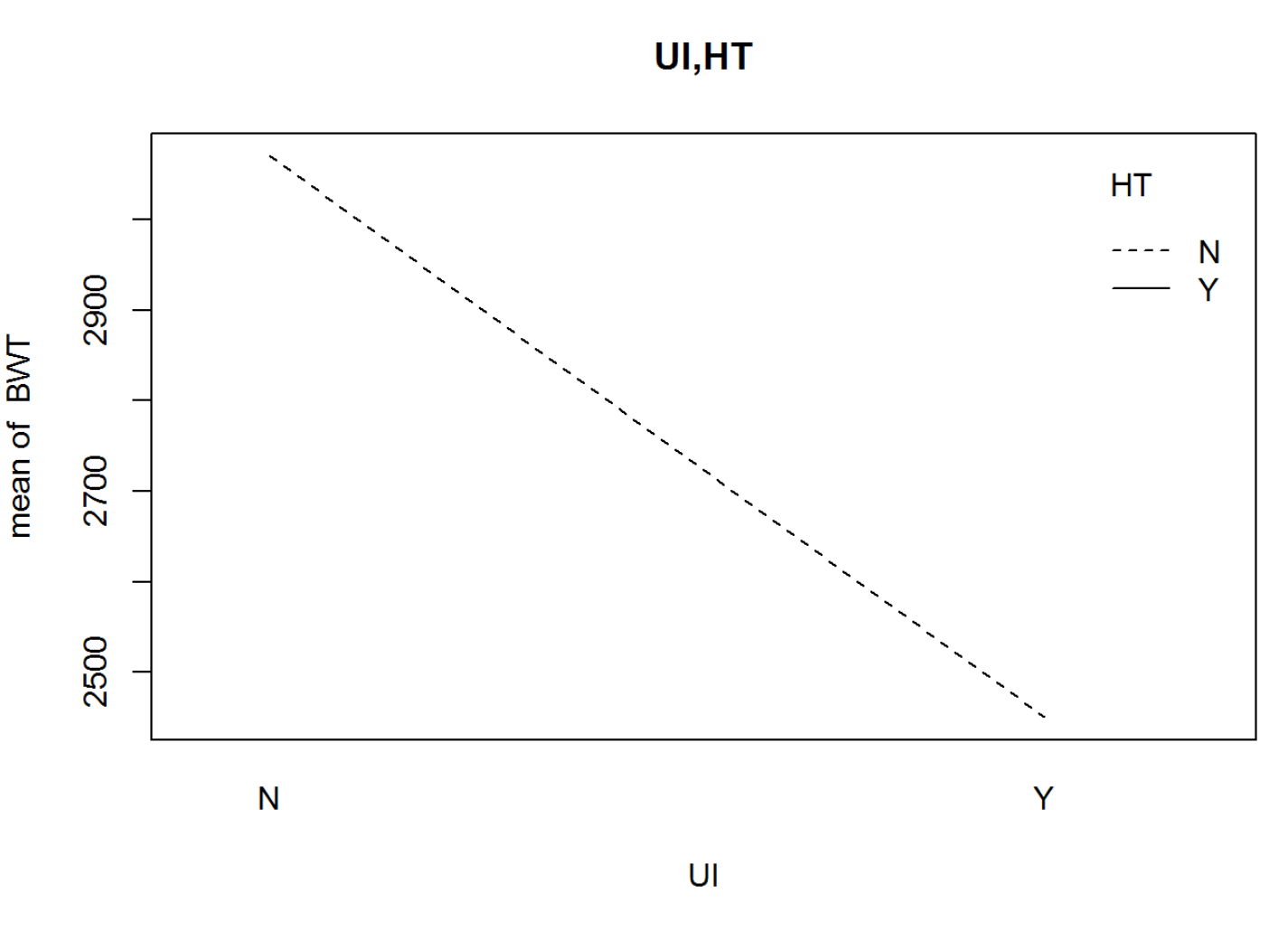
Pour analyser l'impact des facteurs considérons le meme type de graphique que celui de la question 4 nous obtenons les graphiques suivant:



Nous pouvons voir que la courbe des moyennes de poids de naissance associées aux mères qui n'ont pas d'irritabilité utérine est plus élevée que celle de celles qui en ont, quelque soit le niveau de tabagisme de la mère. **Il semble donc avoir un véritable impact de l'uterus irritable sur le poids de naissance**.



Nous pouvons voir que la courbe des moyennes de poids de naissance associées aux mères qui n'ont pas eu d'antécédents d'hypertensions est plus élevée que celle de celles qui en ont eu, quelque soit la race de la mère. **Il semble donc avoir un impact d'antécédents d'hypertensions sur le poids de naissance**.



Nous pouvons seulement voir la courbe des moyennes de poids de naissance associées aux mères qui n'ont pas eu d'antécédents d'hypertensions que la mère à un uterus irretable ou non. **Il semble tout de meme donc avoir un impact d'antécédents d'hypertensions sur le poids de naissance**

2) Analyse de la variance

A) Modèle avec l'intéraction

SMOKE et UI

.	Df	Sum Sq	Mean Sq	F value	Pr(>F)
UI	1	8028747	8028747	16.7840	6.259e-05
SMOKE	1	2949940	2949940	6.1668	0.01391
UI:SMOKE	1	442511	442511	0.9251	0.33740
Residuals	185	88495854	478356		

La dernière colonne donne les p-valeurs des différents tests on voit que les tests sur les facteurs **UI** et **SMOKE** sont significatifs (car 6.259e-05 et 0.01391 sont inférieur à 0.05).

Mais la p-valeur du test d'interaction **UI:SMOKE** est de 0.33740 ce qui est trop élevé .

Donc, dans ce cas on ne peut pas rejeter l'hypothèse nulle au niveau habituel de 0.05 et nous considérons donc un modèle sans interaction.

RACE et HT

.	Df	Sum Sq	Mean Sq	F value	Pr(>F)
HT	1	2132014	2132014	4.2323	0.04108
RACE	2	4716128	2358064	4.6811	0.01041
HT:RACE	2	884006	442003	0.8774	0.41759
Residuals	183	92184904	503743		

La dernière colonne donne les p-valeurs des différents tests on voit que les tests sur les facteurs **HT** et **RACE** sont significatifs(car 0.04108 et 0.01041 sont inférieur à 0.05).

Mais la p-valeur du test d'interaction **HT:RACE** est de 0.41759 ce qui est trop élevé .

Donc, dans ce cas on ne peut pas rejeter l'hypothèse nulle au niveau habituel de 0.05 et nous considérons donc un modèle sans interaction.

UI et HT

.	Df	Sum Sq	Mean Sq	F value	Pr(>F)
HT	1	2132014	2132014	4.4694	0.03584
UI	1	9059202	9059202	18.9912	2.169e-05
Residuals	186	88725836	477021		

La dernière colonne donne les p-valeurs des différents tests on voit que **les tests sur les facteurs HT et UI sont significatifs** (car 0.03584 et 2.169e-05 sont inférieur à 0.05).

Le coefficient **UI:HT** n'apparait pas dans le tableau en effet **notre jeux de donné ne contient pas de femme ayant à la fois un uterus irritable et des antécédents d'hypertations, il aurait fallu en avoir quelque-une c'est une condition nécessaire sur les données pour l'analyse de la variance on devra alors considérer uniquement le modèle sans interractions.**.

B) Modele sans interraction

SMOKE et UI

.	Df	Sum Sq	Mean Sq	F value	Pr(>F)
UI	1	8028747	8028747	16.7908	6.225e-05
SMOKE	1	2949940	2949940	6.1693	0.01388
Residuals	186	88938365	478163		

RACE et HT

.	Df	Sum Sq	Mean Sq	F value	Pr(>F)
HT	1	2132014	2132014	4.2380	0.04093
RACE	2	4716128	2358064	4.6873	0.01033
Residuals	185	93068910	503075		

UI et HT

.	Df	Sum Sq	Mean Sq	F value	Pr(>F)
HT	1	2132014	2132014	4.4694	0.03584
UI	1	9059202	9059202	18.9912	2.169e-05
Residuals	186	88725836	477021		

Dans les 3 tableaux la dernière colonne donne les p-valeurs des différents tests on voit que les tests sur les coefficients **SMOKE**,**UI** resp (**HT** et **RACE**) resp (**HT** et **UI**) sont significatives car leurs p-values est inférieures à 0.05.

3) Coefficient R^2 et R_a^2 associés à chaque modèle

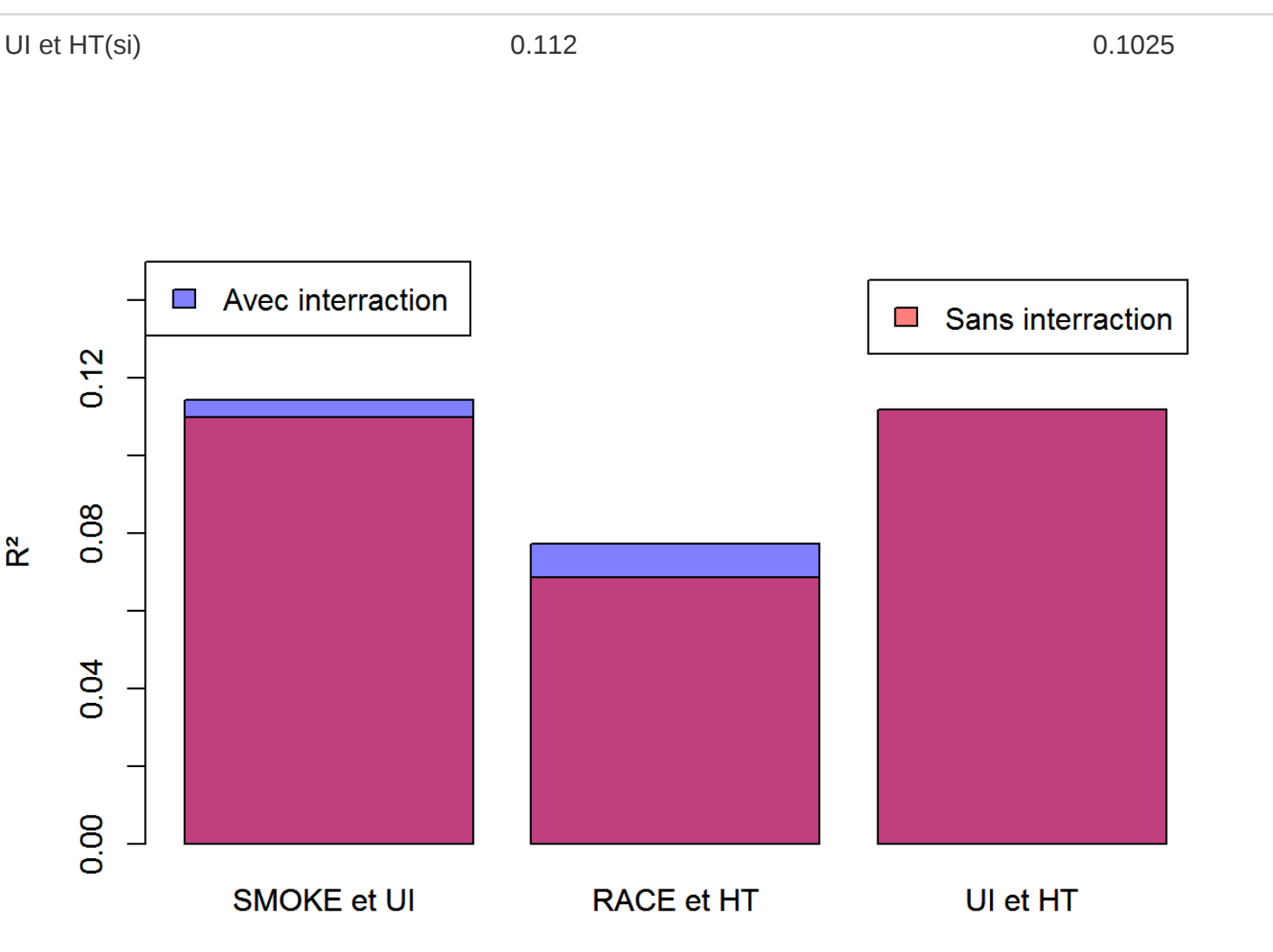
Tout d'abord $R^2 = 1 - \frac{SCR}{SCT}$ et $R^2_{adjusté} = 1 - \frac{SCR/n-k}{SCT/n-1}$ = $1 - \frac{CMR}{CMT}$ avec SCR la somme des carrés des résidus. SCT est la somme des carrés totaux. $CMR = \frac{SCRC}{n-|k|}$ et $CMT = \frac{SCT}{n-1}$ les carrés moyens avec k le nombre de facteur de notre modele et n-1 et n-|k| les degrés de liberté de SCR et SCT

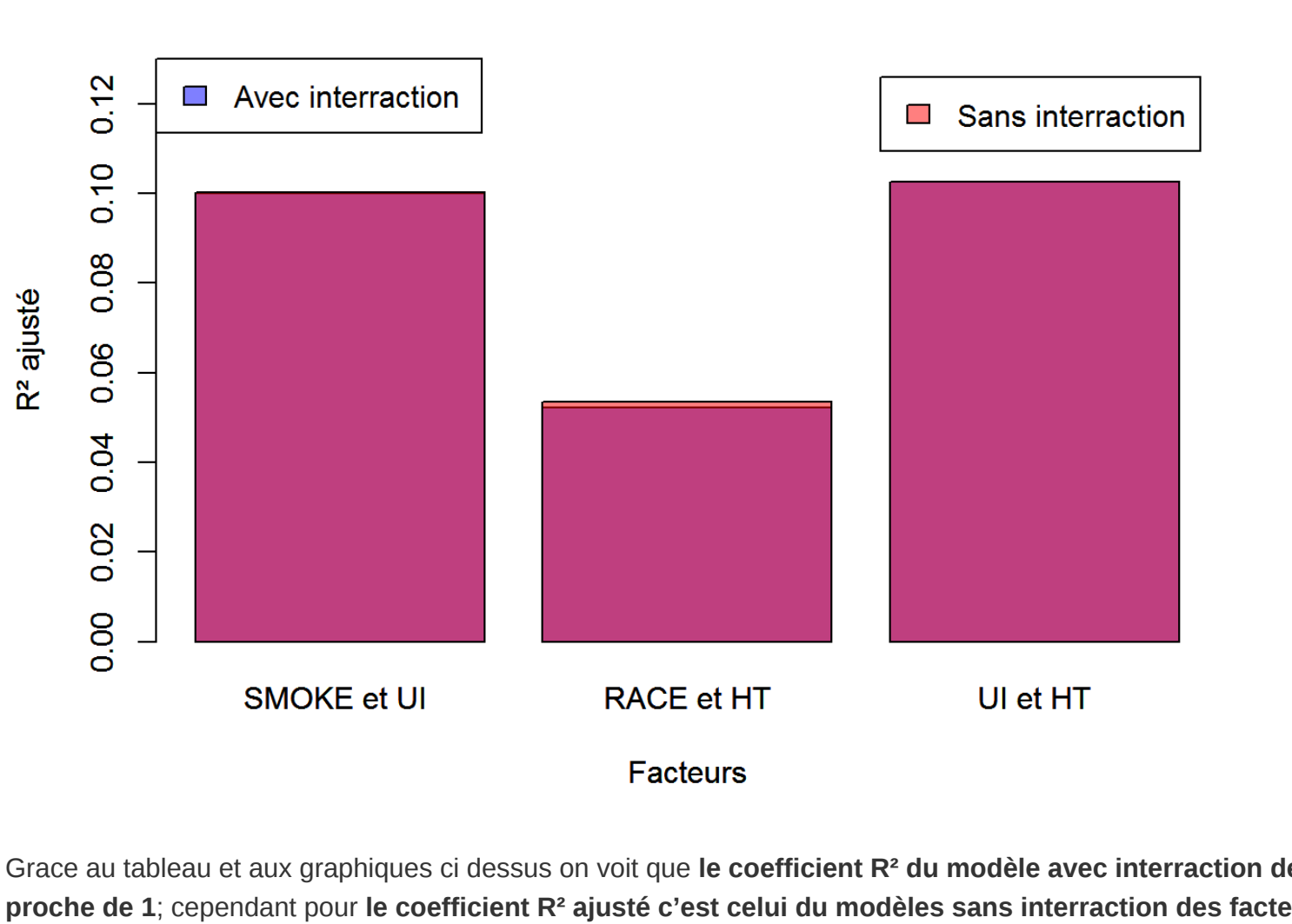
La SCR mesure l'ajustement du modèle aux données. Elle est autant plus petite que l'ajustement est bon. Naturellement, on a intérêt de trouver un modèle dont la SCR est faible. Or, minimiser la SCR revient à maximiser le coefficient de détermination R^2. On cherche donc le modèle qui a le coefficient R^2 le plus proche de 1. Or, ce critère augmente toujours avec le nombre de variables pour une suite de modèles emboîtés, Le R^2 ajusté (Adjusted R-Squared) va alors tenir compte de ce nombre et sera donc plus correct. Au final on cherchera donc le modèle qui a **le coefficient R^2 ajusté le plus proche de 1**.

Observons le tableau et les graphiques suivants tout en ne considérant pas le modèle avec intéraction de UI et HT :

ai: Avec interraction si: Sans interraction

.	Multiple R-squared	Adjusted R-squared
SMOKE et UI (ai)	0.1143	0.09994
RACE et HT(ai)	0.07739	0.05218
SMOKE et UI (si)	0.1099	0.1003
RACE et HT(si)	0.06854	0.05343
UI et HT(si)	0.112	0.1025





Grace au tableau et aux graphiques ci dessus on voit que **le coefficient R² du modèle avec interraction des facteurs SMOKE et UI est le plus proche de 1**; cependant pour **le coefficient R² ajusté c'est celui du modèles sans interraction des facteur UI er HT qui est le plus proche de 1** (0.1025)

On en deduit que **le modèles sans interraction des facteurs UI et HT est le meilleur modèle** puis vient le modèle sans interraction celui avec les facteurs **SMOKE et UI**.