Simulation Examen de Statistique Approfondie II **Corrigé **

Ces quatre exercices sont issus du livre d'exercices de François Husson et de Jérôme Pagès intitulé Statistiques générales pour utilisateurs*, éditions PUR.*

Exercice 1. Prix d'un appartement en fonction de sa superficie

1. Le coefficient de corrélation entre le prix et la superficie d'un appartement T3 correspond à la racine carré du coefficient de détermination $r = \sqrt{0,495} = 0,704$. Ce coefficient est relativement grand. Ces deux variables sont donc liées comme le montre le dessin.
2. Le modèle s'écrit :
$$\forall i \; Y_i = \alpha + \beta x_i + \epsilon_i$$
avec Y_i le prix de l'appartement i en euros, x_i la superficie en m^2 de l'appartement i et ϵ_i le résidu ; on fait désormais les hypothèses habituelles pour les résidus puisque l'on n'est pas à même de pouvoir réaliser les tests adéquats :
$$\forall i \; \mathcal{L}(\epsilon_i) = \mathcal{N}(0, \sigma) \quad \text{et} \quad Cov(\epsilon_i, \epsilon_j) = 0 \; \forall j \neq i$$
Une approche par un modèle de régression linéaire complète celle faite par un coefficient de corrélation : elle permet de prédire une valeur de Y à partir d'une valeur de x.
3. Pour tester si la superficie joue un rôle sur le prix des appatements de type 3, on teste l'hypothèse $H_0 \; : \; \beta = 0$ contre $H_1 \; : \; \beta \neq 0$. La statistique de ce test est $T = \frac{\hat{\beta}}{\hat{\sigma}_\beta}$ et, sous l'hypothèse H_0 cette statistique de test suit une loi de Student à $n-2$ degrés de liberté. La probabilité critique étant inférieure à 5 %, on rejette l'hypothèse nulle et on considère donc que la superficie d'un appartement de type T3 influe sur son prix. Ce test est équivalent à celui qui questionne la nullité du coefficient de corrélation. La superficie a donc une influence mais celle-ci est-elle importante ? Le coefficient de détermination r^2 s'interprète comme le pourcentage de vrariation observée qui est expliquée par le modèle. Ici il est de $49,5$ % ; on en conclue que l'influence est importante mais ne suffit pas à elle seule pour expliquer la variabilté que l'on observe. Il serait donc bon de considérer d'autres facteurs (emplacement, étage, balcon, ascenseur, vue,. . .).

4. L'estimation de la pente de la droite des moindres carrés est $\hat{\beta} = 6,657$ euros par m^2. On interprète ce coefficient de la manière suivante : un appartement coutera en moyenne $6,657$ euros supplémentaires pour une augmentation de $1\ \mathrm{m}^2$.
5. Le prix moyen d'un mètre carré se calcule comme le rapportde 591,95 et 68,74 soit 8,61 euros le mètre carré. Ce prix est différent de l'estimation de β car le prix des appartements n'est pas directement proportionnel à leur surface. Comme $\hat{\beta}$ est inférieur au prix moyen d'un mètre carré, proportionnellement à la surface, les petits appartements sont plus chers que les grands. Le modèle de régression stipule qu'il faut d'abord une mise de fond α pour louer un T3, et qu'ensuite le prix d'1 m^2 coûte β. Remarquons que ce coefficient α est significatif. Il n'est donc pas souhaitable de la retirer du modèle.
6. Pour déterminer les appartements "bon marché", on peut se fonder sur l'estimation des résidus du modèle : plus le résidu est faible (négatif et avec une forte valeur absolue), plus l'appartement a un prix faible par rapport à celui attendu pour sa superficie.

Exercice 2. Etude de l'appréciation de cocktails de jus de fruits en fonction de leurs saveurs fondamentales et de leur caractère pulpeux

1. Le tableau d'analyse de la variance complété est :

```
Analyse de variance

Source           DL        SC         CM        F        P
Régression        4   11,6075     2,9019     6,06    0,008
Erreur résid     11    5,2694     0,4790
Total            15   16,8768
```

   Pour tester l'hypothèse de nullité simultanée des 5 coefficients au seuil $\alpha = 5$ %, on calcule la valeur de la statistique de test $f_{obs} = \dfrac{CM_{\text{modèle}}}{CM_{\text{résiduelle}}} = 6,06$. Cette valeur étant supérieure au quantile $0,95$ de la loi de Fisher à 4 et 11 degrés de liberté (3,357), on rejette l'hypothèse H_0 et on considère qu'au moins un coefficient est non nul.

2. Le fait de centrer et de réduire les données ne modifie en rien la pertinence statistique du modèle : toutes les probabilités critiques sont inchangées, exceptée, bien sûr, la probabilité critique relative à la constante du modèle puisque cette dernière est nullepar construction (ces probabilités critiques seront commentées à la question 5.). L'intérêt du centrage-réduction est de rendre les coefficients $\hat{\beta}_j$ comparables entre eux. Ainsi est-il possible de voir quelles sont les variables les plus influentes (celles qui ont un coefficient le plus élevé). On considèrera donc que c'est la variable *acide* qui influe le plus sur l'appréciation.

3. Pour les variables *amer* et *pulpeux*, le signe de $\hat{beta_j}$ est différent du signe du coefficient de corrélation entre cette variable et la variable appréciation. Ces situations sont inattendues car à partir du coefficient de corrélation, on dira, par exemple, que plus le cocktail est amer moins il est apprécié ($r < 0$), tandis qu'avec le modèle on dira que si l'amertume augmente, alors l'appréciation augmente ($\hat{\beta}_j > 0$). Cette contradiction apparente provient du fait que, lorsque l'on interprète un coefficient du modèle, on considère implicitement toutes les autres variables égales par ailleurs. Or lorsqu'un cocktail est amer, il est également plutôt acide (corrélation de 0,858). Et un cocktail acide est moins apprécié, ce qui est déjà pris en compte dans le modèle (par la variable *acide*). Autrement dit, le modèle indique qu'àcidité, sucrosité, caractère pulpeux constants, les cocktails ont tendance à être plus appréciés lorsqu'ils sont plus amers.

   Ceci illustre une des caractéristiques de la régression multiple : lorsque les prédicteurs sont corrélés, leurs coefficients dans le modèle ne sont pas aisément interprétables.
4. Le test d'un paramètre se construit de la même façon que le test de β en régression simple. Les hypothèses testées sont $H_0 \; : \; \beta_j$ contre $H_1 \; : \; \beta_j \neq 0$. La statistique du test est : $\frac{\hat{\beta}_j}{\sigma_{\beta_j}}$. Sous l'hypothèse nulle, cette statistique suit une loi de Student à $n-p-1$ degrés de liberté (i.e. les degrés de liberté de la résiduelle), soit 11 dans cet exercice. Rèdle de décision : si la valeur absolue de la statistique de test est supérieure au quantile 0,975 de la loi de Student à 11 degrés de liberté, alors on décide l'hypothèse alternative, sinon on accepte l'hypothèse nulle. En pratique, il est plus facile de comparer la $p-$valeur au seuil α choisi.
5. Pris séparément, les tests de nullité d'un coefficient montrent qu'aucun coefficient n'est significativement différent de 0 au seuil 5 %. Ceci est en apparente contradiction avec le test de nullité simultanée de tous les coefficients (test F) puisqu'on avait considéré qu'au moins un coefficient était non nul. Mais dans le test d'un coefficient β_j, on teste si ce coefficient est nul, toutes les autres variables étant présentes dans le modèle. L'hypothèse testée peut se lire "la variable β_j n'apporte pas d'information supplémentaire pour expliquer Y, i.e. une information non déjà expliquée par les autres variables du modèle". Ainsi, dans l'exemple, chaque prédicteur n'a pas d'apport significatif compte tenu de la présence de tous les autres. On peut alors se demander s'il ne serait pas judicieux de se limiter à un sous-ensemble de prédicteurs. Pour cela, il existe des algorithmes (et donc des programmes) dit de "sélection de prédicteurs".

Exercice 3. Prédiction du maximum d'ozone à Rennes

1. Le modèle s'écrit :
$$\forall i \; Y_i = \beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2} + \ldots + \beta_1 2 x_{i_1 2} + \epsilon_i$$
avec les hypothèses usuelles pour les résidus :
$$\forall i \; \mathcal{L}(\epsilon_i) = \mathcal{N}(0, \sigma) \quad \text{et} \quad Cov(\epsilon_i, \epsilon_j) = 0 \; \forall j \neq i$$

2. Les hypothèses testées sont :
$$H_0 \ : \ \forall j, \ \beta_j = 0 \qquad \text{contre} \qquad H_1 \ : \ \exists \ j_0 \mid \beta_{j_0} \neq 0$$
La statistique de test utilisée est :
$$F = \frac{CM_{\text{modèle}}}{CM_{\text{résiduelle}}} = \frac{SCE_{\text{modèle}}/p}{SCE_{\text{résiduelle}}/(n-p-1)}$$
Sous l'hypothèse nulle, F suit une loi de Fisher à p et $n-p-1$ degrés de liberté. Décision : soit on compare $f_{obs} = 30,84$ au quantile $1-\alpha$ de la loi de Fisher à 12 et 78 degrés de liberté, soit (ce qui revient au même mais est plus commode) on compare la probabilité critique à α. Ici la probabilité critique est inférieure à 0,0001 donc on rejette l'hypothèse H_0 au seuil de 5 %. Par conséquent, on considère qu'au moins une variable explicative influe sur la concentration d'ozone.
3. Nous allons détailler le test $H_0 \ : \ \beta_j$ contre $H_1 \ : \ \beta_j \neq 0$. Si les résidus suivent une loi normale, alors $\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \sigma_{\hat{\beta}_1})$. Puisque l'on ne connaît pas la vraie valeur de $\sigma_{\hat{\beta}_1}$, on l'estime et on a :
$$\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\beta_j}} \sim \text{Student(n-p-1)}$$
Sous l'hypothèse nulle, $T = \frac{\hat{\beta}_1}{\sigma_{\beta_j}}$ suit une loi de Student à $n-p-1$ degrés de liberté. Il suffit donc de calculer t_{obs} et de le comparer au quantile $1 - \dfrac{\alpha}{2}$ de la loi de Student à $n-p-1$ degrés de liberté. $t_{obs} = \frac{-1,872}{1,236} = -1,514$, donc $|t_{obs}| = 1,514 < 1,99 = t_{78}(0,975)$ (on peut aussi remarquer que la $p-$valeur associée à ce test est égale à 0,134 ; or $0,134 > 0,05$). Par conséquent le test n'est pas significatif au niveau 5 % et on accepte l'hypothèse nulle. On considère donc que la température à 6h00 n'apporte pas, de manière significative, d'information supplémentaire sur le maximum d'ozone si toutes les autres variables sont déjà dans le modèle. En revanche la température à 18h00 apporte une information complémentaire par rapport aux autres variables explicatives du modèle car la probabilité critique associée au test de nullité du coefficient de la température à 18h est de 0,02.
4. Voici le tableau qui manquait à l'énoncé :

```
La réponse est maxO3
```

Vars	R-carré	R-carré(ajust)	C-p	S	T6	T9	T12	T15	T18	Ne6	Ne9	Ne12	Ne15	Ne18	Vx	maxO3v
1	58,1	57,7	100,6	22,891												X
1	38,3	37,6	189,4	27,784						X						
2	71,4	70,7	43,3	19,036						X						X
2	69,4	68,7	52,0	19,672					X							X

3	78,7	78,0	12,4	16,512					X	X						X
3	77,7	76,9	16,8	16,891	X				X							X
4	80,7	79,8	5,4	15,800	X				X	X						X
4	80,2	79,3	7,5	15,997					X	X					X	X
5	81,7	80,6	2,9	15,480	X				X	X					X	X
5	81,4	80,3	4,5	15,632		X			X	X					X	X
6	82,0	80,7	3,8	15,466	X		X		X	X					X	X
6	81,9	80,6	3,9	15,476	X				X	X			X		X	X
7	82,3	80,8	4,4	15,423	X				X	X			X	X	X	X
7	82,2	80,7	4,9	15,470	X		X		X	X			X		X	X
8	82,5	80,8	5,5	15,423	X		X		X	X			X	X	X	X
8	82,4	80,7	5,7	15,450	X			X	X	X			X	X	X	X
9	82,5	80,6	7,3	15,497	X	X	X		X	X			X	X	X	X
9	82,5	80,6	7,3	15,500	X		X		X	X	X		X	X	X	X
10	82,6	80,4	9,0	15,569	X	X	X		X	X	X		X	X	X	X
10	82,6	80,4	9,2	15,584	X	X	X		X	X		X	X	X	X	X
11	82,6	80,2	11,0	15,667	X	X	X	X	X	X	X		X	X	X	X
11	82,6	80,2	11,0	15,667	X	X	X		X	X	X	X	X	X	X	X
12	82,6	79,9	13,0	15,767	X	X	X	X	X	X	X	X	X	X	X	X

Le modèle ci-dessous est celui qui propose la réduction de variables explicatives la plus appropriée puisqu'à la fois le R^2 ajusté et le C_p sont bons. En effet $R^2 = 80,3$ ce qui est très proche de la valeur maximale et $C_p = 4,5$ ce qui est plus proche de $p = 5$ que pour la plupart des autres modèles.

```
La réponse est maxO3
```

Vars	R-carré	R-carré(ajust)	C-p	S	T6	T9	T12	T15	T18	Ne6	Ne9	Ne12	Ne15	Ne18	Vx	maxO3v
5	81,4	80,3	4,5	15,632		X			X	X					X	X