4.1 **Variables explicatives catégorielles** : Dans cet exercice, nous reprenons l'exemple d'intention d'achat vu en classe et nous allons effectuer une analyse de régression pour évaluer l'effet de la variable `revenu` sur la variable `intention`.

(a) Ajustez le modèle en créant vous-même les variables indicatrices binaires que vous allez inclure dans le modèle (utilisez la catégorie 3 comme catégorie de référence). Écrivez formellement le modèle ajusté et interprétez les coefficients du modèle.

(b) À l'aide du modèle ajusté en (a), prédisez l'intention d'achat pour un individu dont le revenu est supérieur à 60 000$.

(c) Ajustez le modèle en spécifiant que la variable revenu est catégorielle (commande `class` en SAS , ou `as.factor` en R). Écrivez l'équation de la régression et interprétez les coefficients.

(d) Réajustez le modèle de régression une dernière fois en traitant la variable `revenu` comme une variable continue. Comparez les résultats et commentez sur la différence conceptuelle de traiter `revenu` comme une variable continue versus catégorielle.

(e) Ajustez un modèle de régression linéaire pour l'intention d'achat avec toutes les variables et interprétez l'effet de cette dernière.

(f) Testez l'effet global conditionnel des variables `revenu` et `educ`, étant donné les autres variables explicatives dans le modèle.

4.2 Le jeu de données `automobile` contient des informations sur 392 voitures. On considère un modèle linéaire liant la consommation d'essence (en miles au gallon) des voitures en fonction de leur puissance (en watts).

(a) Tracez un nuage de point illustrant la relation entre la consommation d'essence (`consommation`) et la puissance (`puissance`) et commentez.

(b) Ajustez un modèle linéaire avec `puissance` comme variable explicative. Commentez sur l'adéquation en regardant le R^2 et les diagrammes des résidus.

(c) Ajustez le modèle quadratique

$$\texttt{consommation} = \beta_0 + \beta_1 \texttt{puissance} + \beta_2 \texttt{puissance}^2 + \varepsilon$$

et commentez la qualité de l'ajustement et la significativité des coefficients. En SAS , le code suivant permet d'ajuster le modèle quadratique :

```
proc glm data=infe.automobile;
model consommation=puissance puissance*puissance/ss3 solution;
run;
```

et en R via

```
lm(consommation~puissance+I(puissance^2), data = automobile)
```

(d) Ajustez maintenant un modèle cubique et comparez au modèle précédent.

(e) Concluez quand au modèle le plus approprié pour les données sur la base de la significativité des paramètres. Faites également une analyse des résidus pour le modèle d'ordre un, le modèle quadratique et le modèle cubique et comparez-les.

4.3 **Interactions entre variables catégorielle et continue** Nous avons vu en classe comment modéliser et interpréter l'interaction entre une variable binaire et une variable continue. Cet exercice a pour but de vous expliquer comment ajuster et interpréter un modèle incluant un terme d'interaction entre une variable catégorielle et une variable continue. Pour cet exercice, nous allons travailler avec l'`intention` d'achat, mais uniquement avec deux variables explicatives, `educ` et `fixation`. La variable `educ` possède trois catégories, et donc cette variable va être modélisée à l'aide de deux variable indicatrices binaires `educ1` (respectivement `educ2`) vaut un si `educ=1` (`educ=2`)et zéro sinon.

(a) Ajustez un modèle de régression incluant les variables `educ` et `fixation` pour modéliser `intention`, sans interaction. Utilisez la catégorie trois de la variable `educ` comme catégorie de référence.

Compiled 09/02/2020 at 23:06

scénario	`fixation`	`educ`	**intention d'achat moyenne selon le modèle**
1	x	1	
2	x	2	
3	x	3	
4	$x+1$	1	
5	$x+1$	2	
6	$x+1$	3	
7	0	1	
8	0	2	
9	0	3	

TABLE 1 – Valeurs ajustées pour l'intention d'achat pour les neuf scénarios

i. Écrivez l'équation du modèle de régression estimé.
ii. Selon l'équation du modèle, calculez les trois équations des droites estimant la relation entre `intention` et `fixation` lorsque `educ=1`, `educ=2` et `educ=3`, respectivement.
iii. La sortie SAS inclut un graphique montrant l'effet de fixation sur intention selon les trois groupes d'éducation (coloré selon le groupe). Que pensez-vous de la qualité de la modélisation? Selon vous, quelle(s) caractéristique(s) devrait avoir le modèle "idéal" pour ces données, que le présent modèle n'a pas?

(b) Ajoutez au précédent modèle une interaction entre `educ` et `fixation` (si vous utilisez SAS , avec la commande `class`). Le modèle postulé est

$$\texttt{intention} = \beta_0 + \beta_1\texttt{educ1} + \beta_2\texttt{educ2} + \beta_3\texttt{fixation} + \beta_4\texttt{educ1}\times\texttt{fixation} + \beta_5\texttt{educ2}\times\texttt{fixation} + \varepsilon \tag{E1}$$

i. Écrivez l'équation du modèle de régression estimé et commentez sur la significativité des termes d'interaction.
ii. Calculez les trois équations des droites estimant la relation entre `intention` et `fixation` lorsque `educ=1`, `educ=2` et `educ=3`, respectivement.
iii. En examinant le graphique de la sortie SAS montrant l'effet de fixation sur intention selon les trois groupes d'éducation (code de couleur), comparez l'ajustement de ce modèle avec le modèle sans interaction et commentez.

(c) La partie suivante traite de l'interprétation des coefficients du modèle en présence d'une interaction.
i. Remplissez le tableau des valeurs de l'intention d'achat dans les neuf scénarios suivants.
ii. À l'aide des scénarios 3 et 6 du tableau, interprétez le coefficient β_3 du modèle de régression (pente de la variable `fixation`)
iii. À l'aide des scénarios 7 et 9 du tableau, interprétez le coefficient β_1 du modèle de régression (pente de la variable `educ1`)
iv. À l'aide des scénarios 8 et 9 du tableau, interprétez le coefficient β_2 du modèle de régression (pente de la variable `educ2`)

4.4 La série chronologique `trafficaerien` donne le nombre total mensuel de passagers internationaux (en milliers) pour la période 1949 à 1960.

(a) Ajustez un modèle linéaire avec l'année comme variable explicative. Quelle est l'interprétation de l'ordonnée à l'origine et de la pente? Considérez un modèle équivalent dans lequel la variable explicative année est décalée par 1949, soit $t-1949$. Comment est-ce que cette transformation affecte l'interprétation des coefficients?

(b) Considérez l'ajout d'un effet mensuel en traitant cette variable comme une variable catégorielle (prenez janvier comme référence). Écrivez l'équation du modèle théorique et ajustez ce dernier. Est-ce que vous notez

Compiled 09/02/2020 at 23:06

une amélioration de l'ajustement?

(c) Utilisez le modèle avec la variable catégorielle mensuelle et l'année pour prédire le nombre de passagers mensuels en décembre 1962.

(d) Présentez des diagnostics graphiques pour valider les hypothèse du modèle linéaire. Que remarquez-vous?

(e) Il est plausible que la croissance du traffic soit exponentielle durant la période à l'étude. Essayez d'ajuster un modèle linéaire avec le log du nombre de passagers comme variable réponse. Produisez et rapportez les diagnostics graphiques suivants : (1) un nuage de points des valeurs ajustées et des résidus ordinaires (2) un nuage de point des résidus studentisés externes en fonction du temps (3) un diagramme quantile-quantile des résidus studentisés externes et (4) un nuage de points des résidus décalés, soit un graphique de e_i en fonction de e_{i+1} pour $i = 1, \ldots, n-1$. Est-ce que les postulats du modèle linéaire semblent valides? Commentez

4.5 Le jeu de données `Ratemyprofessor` fourni des notes sur 366 enseignants (159 femmes et 207 hommes) dans une université du *Midwest* américain. Chaque enseignant inclut dans la base de donnée avait reçu un minimum de 10 évaluations (potentiellement sur une période s'étalant sur plusieurs années). Les étudiant(e)s fournissaient des notes sur une échelle de 5 : les variables `serviabilite`, `clarte` et `facilite` sont des moyennes d'autres échelles de Likert sur $[1,5]$, des valeurs basses indiquant de mauvais scores. Les données contiennent ces notes moyennes et d'autres informations sur les enseignant(e)s. Le but de l'analyse est de prédire la qualité en fonction des autres variables. Le Table 2 contient les coefficients (avec erreurs-type), des mesures d'adéquation pour huit modèles différents.

(a) Rapportez le score de qualité moyen des enseignantes de l'échantillon.

(b) À l'aide du modèle 8, prédisez le score de qualité moyen pour un homme dont les scores de `serviabilite`, `clarte` et `facilite` sont tous égaux à 4.

(c) Quelles sont les hypothèses nulle et alternative associées à la statistique F globale qui vaut 62228.971 dans le modèle 4. Donnez la conclusion de ce test d'hypothèse. *Indice : le 95% quantile de la loi nulle est 3.021.*

(d) Donnez un intervalle de confiance à 95% approximatif pour le paramètre `clarte` dans le modèle 4, de la forme $\widehat{\beta}_j \pm 1.96\text{se}(\widehat{\beta}_j)$. Est-ce que le modèle 2 est une simplification adéquate du modèle 4?

(e) Contrastez les coefficients estimés pour les modèles 2 et 4. Est-ce que ces estimés sont cohérents avec les graphiques de la Figure 1?

(f) Expliquez pourquoi on ne devrait pas considérer le modèle 7, et ce peu importe si le coefficient associé à l'interaction interaction `homme:serviabilite` est significatif.

(g) Quelles sont les postulats du modèle linéaire? Commentez sur la validité sur la base des graphiques présentés dans les Figures 1 and 2.

Compiled 09/02/2020 at 23:06

	modèle 1	modèle 2	modèle 3	modèle 4
`constante`	3.532 (0.066)	0.033 (0.038)	0.221 (0.040)	−0.020 (0.011)
`homme (sexe)`	0.077 (0.088)			
`serviabilite`		0.975 (0.010)		0.538 (0.007)
`clarte`			0.952 (0.011)	0.466 (0.007)
R^2	0.002	0.962	0.952	0.997
degrés de liberté	364	364	364	363
statistique F (test global)	0.755	9322.673	7299.061	62228.971
somme du carré des résidus (RSS)	255.479	9.620	12.161	0.745
s^2	0.702	0.026	0.033	0.002
AIC	913.088	-287.129	-201.361	-1221.679

	modèle 5	modèle 6	modèle 7	modèle 8
`constante`	−0.029 (0.011)	−0.030 (0.012)	0.323 (0.057)	−0.054 (0.016)
`homme (sexe)`		0.002 (0.005)	−0.397 (0.076)	0.048 (0.021)
`serviabilite`	0.536 (0.007)	0.535 (0.007)		0.541 (0.008)
`clarte`	0.465 (0.007)	0.465 (0.007)	0.863 (0.016)	0.466 (0.007)
`facilite`	0.007 (0.004)	0.007 (0.004)	0.062 (0.014)	0.007 (0.004)
`homme:serviabilite`			0.116 (0.020)	−0.013 (0.006)
R^2	0.997	0.997	0.959	0.997
degrés de liberté	362	361	361	360
statistique F (test global)	41739.797	31236.209	2107.165	25272.111
somme du carré des résidus (RSS)	0.738	0.738	10.515	0.727
s^2	0.002	0.002	0.029	0.002
AIC	-1222.912	-1221.120	-248.592	-1224.244

TABLE 2 – Coefficients (erreurs-type) et mesures d'adéquation pour différents modèles ajustés aux données `Ratemyprofessor`.

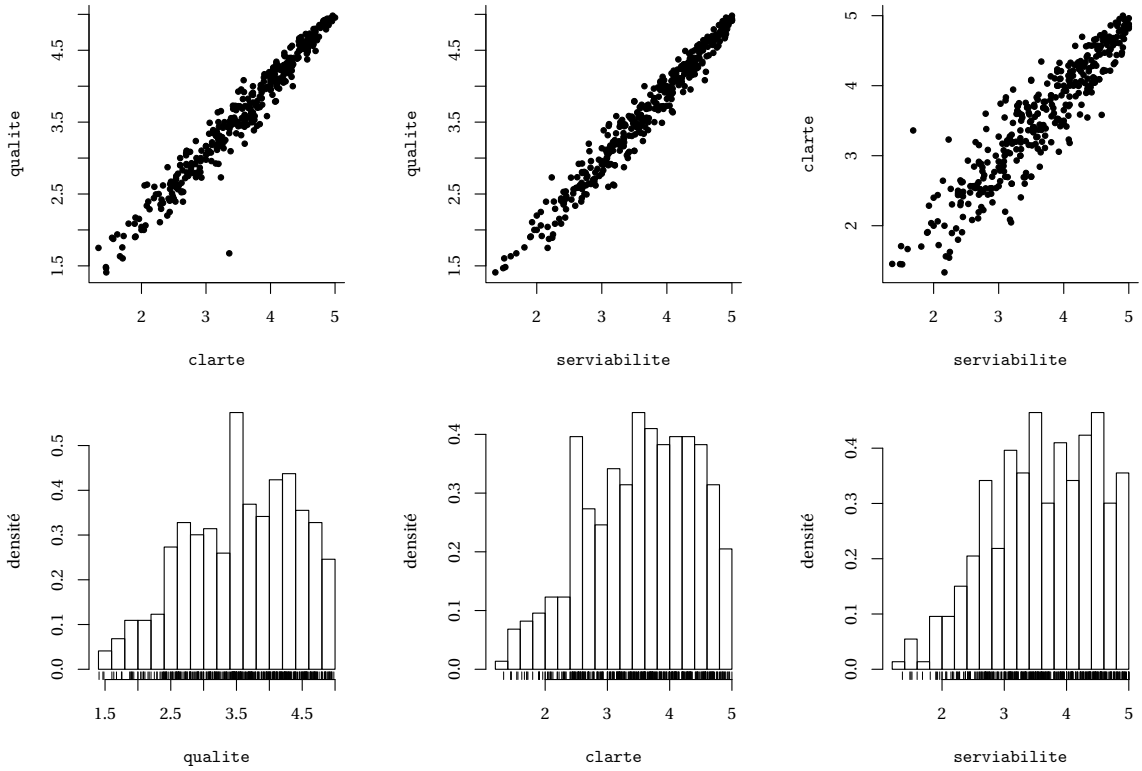

FIGURE 1 – Panneau supérieur : nuage de point des paires (les corrélation linéaires de gauche à droite sont égales à 0.98,0.98 et 0.92). Panneau inférieur : histogramme des scores moyennes des indicateurs `qualite`, `serviabilite` et `clarte`.

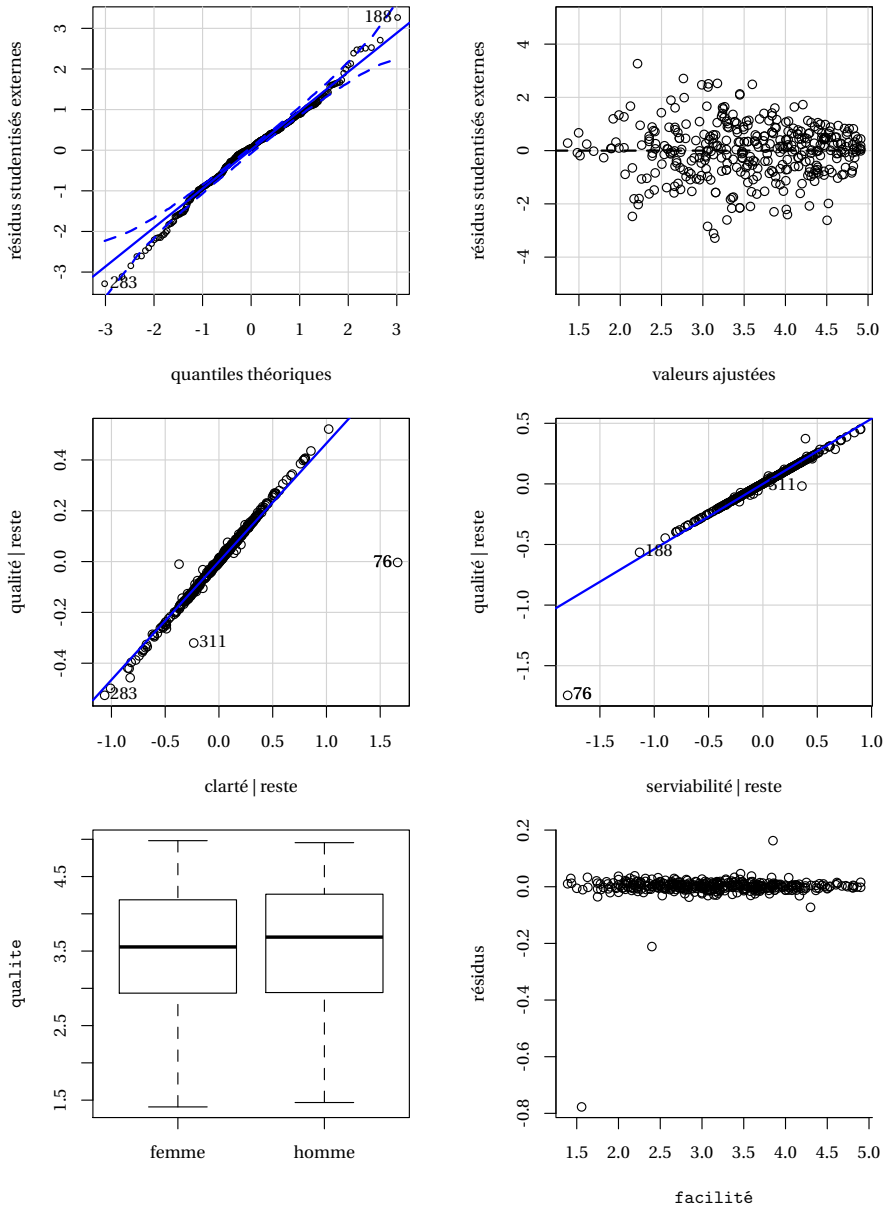

FIGURE 2 – Diagnostics graphiques pour le modèle 4 ajusté aux données `Ratemyprofessor`. Panneau supérieur gauche : diagramme quantile-quantile des résidus studentisés externes, avec intervalles de confiance ponctuels à 95% (traitillés), en excluant l'observation 76. Panneau supérieur droit : diagramme des résidus ordinaires contre les valeurs ajustées. Milieu : diagrammes de régression partielle pour `clarte` et `serviabilite`. Panneau inférieur gauche : boîte à moustache de l'indice `qualite` en fonction du sexe. Panneau inférieur droit : résidus ordinaires $\boldsymbol{e}$ versus la variable omise `facilite`.

Compiled 09/02/2020 at 23:06