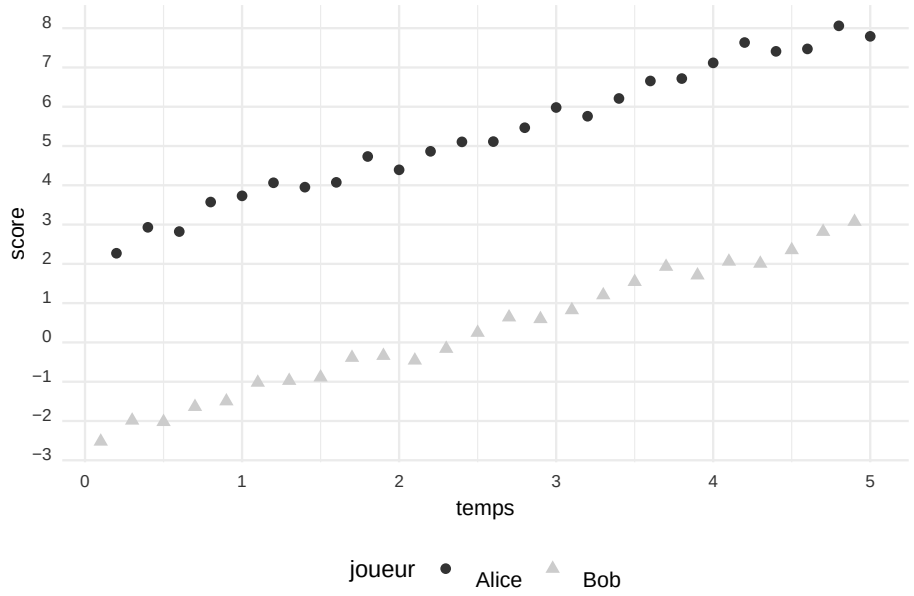

Figure 1: Scores d'Alice et de Bob en fonction du nombre d'heures de jeu.

2.1 Bob et Alice remarquent que leurs scores à un jeu de société peuvent être modélisés en fonction du nombre d'heures de jeu à l'aide du modèle linéaire

$$\texttt{score}_i = \beta_0 + \beta_1 \texttt{temps}_i + \beta_2 \texttt{joueur}_i + \beta_3 \texttt{temps}_i \texttt{joueur}_i + \varepsilon_i,$$

où ε_i est un terme d'erreur de moyenne nulle et $\texttt{joueur}_i$ est une variable binaire égale à un si le ie score est celui d'Alice et zéro pour celui de Bob.
En vous basant sur la Figure 1, que peut-on dire quant aux signes des coefficients $\widehat{\beta}_1, \ldots, \widehat{\beta}_3$?

2.2 On considère un modèle de régression pour expliquer l'impact de l'éducation et du nombre d'enfants sur le salaire des femmes, à savoir:

$$\texttt{logsalaire}_i = \beta_0 + \beta_1 \mathrm{X}_{i1} + \beta_2 \mathrm{X}_{i2} + \varepsilon_i,$$

où

$$\mathrm{X}_1 = \begin{cases} 0, & \text{si la femme n'a pas de diplôme secondaire,} \\ 1, & \text{si la femme a un diplôme secondaire mais pas de diplôme collégial,} \\ -1, & \text{si la femme a un diplôme collégial.} \end{cases}$$

$$\mathrm{X}_2 = \begin{cases} 0, & \text{si la femme n'a pas d'enfant,} \\ 1, & \text{si la femme a 1 ou 2 enfants,} \\ -1, & \text{si la femme a 3 enfants ou plus.} \end{cases}$$

Selon ce modèle, quelle serait la **différence** moyenne en log-salaire entre (i) une femme qui possède un diplôme collégial et qui a trois enfants et (ii) la moyenne de toutes les femmes dans l'échantillon, en supposant que la taille de chacun des neuf groupes est la même (plan équilibré)?

2.3 On considère le log du prix de vente de maisons en fonction de leur localisation (urbain ou rural), de la surface du

garage (en pieds carrés), et d'un indicateur dénotant la présence ou l'absence de garage. Le modèle linéaire postulé est le suivant

$$\texttt{logprix} = \beta_0 + \beta_1 \texttt{garage} + \beta_2 \texttt{surface} + \beta_3 \mathbf{1}_{\texttt{loc=urbain}} + \varepsilon,$$

où ε est un terme d'erreur de moyenne nulle et `garage` une variable indicatrice

$$\texttt{garage} = \begin{cases} 0, & \text{si la maison a un garage } (\texttt{surface} > 0); \\ 1, & \text{si la maison n'a pas de garage } (\texttt{surface} = 0). \end{cases}$$

On suppose que le modèle a été ajusté par la méthode des moindres carrés et qu'on obtient $\widehat{\beta}_1 > 0$ et $\widehat{\beta}_2 > 0$. Laquelle des affirmations suivantes est **toujours** correcte?

(a) Toutes choses étant égales par ailleurs, les maisons avec garage sont en moyenne plus chères que celles sans garage.
(b) Toutes choses étant égales par ailleurs, les maisons sans garage sont toujours moins chères que celles avec garage.
(c) Toutes choses étant égales par ailleurs, les maisons avec garage sont en moyenne moins chères que celles sans garage.
(d) La localisation (urbain versus rural) est négativement corrélée avec la surface du garage.
(e) Aucune de ces réponses.

2.4 On considère un modèle de régression simple pour le prix d'une voiture électrique en fonction de son autonomie (distance); le modèle est

$$\texttt{prix}^{\texttt{USD}} = \beta_0^i + \beta_1^i \texttt{distance}^{\texttt{mi.}} + \varepsilon^i,$$

où ε est un terme d'erreur de moyenne nulle. Vos amis ont collecté des données où la variable prix est mesurée en dollars américains (`USD`) et la variable distance est mesurée en miles (`mi.`), et ont ajusté le modèle de régression afin d'obtenir les estimés $(\widehat{\beta_0^i}, \widehat{\beta_1^i})$.
Vous aimeriez connaître les estimés du modèle où la variable prix est exprimée en dollars canadiens (`CAD`) et la variable distance est exprimée en kilomètres (`km`), c'est-à-dire

$$\texttt{prix}^{\texttt{CAD}} = \beta_0^m + \beta_1^m \texttt{distance}^{\texttt{km}} + \varepsilon^m.$$

Sachant que 1 USD vaut 1.39 CAD et que 1 mile égale 1.61 km, quelle est la valeur de C dans l'équation $\widehat{\beta_1^i} = C\widehat{\beta_1^m}$?

2.5 Les données `eolienne` contiennent des mesures de la production électrique d'éoliennes sur 25 périodes non-consécutives de 15 minutes. Nous sommes intéressés à modéliser la relation entre la production électrique et la vitesse du vent moyenne (mesurée en miles à l'heure) pendant la période de mesure.

(a) Ajustez un modèle linéaire avec la vitesse du vent comme covariable et produisez un graphique des résidus contre les valeurs ajustées. Est-ce que vous remarquez une structure résiduelle qui n'est pas prise en compte dans votre modèle? Essayez aussi un modèle avec la réciproque de la vitesse du vent comme variable explicative. Commentez sur l'adéquation des deux modèles.
(b) Prédisez, en utilisant les deux modèles à tour de rôle, la production électrique sachant que la vitesse du vent moyenne dans une période donnée est de 5 miles à l'heure. Fournissez également des intervalles de prédiction pour vos estimés.
(c) [⋆] La production électrique de l'éolienne devrait être inexistente en l'absence de vent, mais cette réalité n'est

Compilé le 06/08/2020 à 12:14

pas capturée par le premier modèle liant la production électrique à la vitesse du vent. Mettez votre modèle à jour en retirant l'ordonnée à l'origine (avec `~-1` dans R ou l'option `noint` dans SAS avec `prog glm`. Qu'arrive-t-ils i vous retirez l'ordonnée à l'origine?

(d) Produisez un diagramme quantile-quantile des résidus studentisés externes et commentez sur l'hypothèse de normalité.

2.6 Dans le cadre d'une étude réalisée au Tech3Lab, des cobayes devaient naviguer sur un site internet qui contenait, entre autres choses, une publicité pour des bonbons. Pendant la navigation, un oculomètre mesurait l'endroit où se posait le regard du sujet. On a ainsi pu mesurer si le sujet a regardé la publicité et combien de temps il l'a regardé. De plus, un logiciel d'analyse des expressions faciales (FaceReader) a également été utilisé pour mesurer l'émotion du sujet pendant qu'il regardait la publicité. À la fin de l'expérience, un questionnaire mesurait l'intention d'achat du sujet pour ces bonbons, ainsi que des variables socio-démographiques. Seuls les 120 sujets qui ont regardé la publicité sont inclus dans les données `intention`, qui contient les variables suivantes:

- `intention`: variable discrète entre 2 et 14; plus elle est élevée, plus le sujet exprime l'intention d'acheter ce produit. Le score a été construit en additionnant les réponses de deux questions sur une échelle de Likert allant de fortement en désaccord (1) à fortement en accord (7).
- `fixation`: durée totale de fixation de la publicité (en secondes).
- `emotion`: une mesure de la valence durant la fixation, soit le ratio de la probabilité d'une émotion positive sur la probabilité d'une émotion négative
- `sexe`: sexe du sujet, soit homme (0) ou femme (1).
- `age`: âge (en années).
- `revenu`: variable catégorique indiquant le revenu annuel du sujet; un parmi (1) $[0, 20\,000]$; (2) $[20\,000, 60\,000]$ ou (3) 60 000 et plus.
- `educ`: variable catégorique indiquant le niveau d'éducation le plus élevé obtenu, soit (1) secondaire ou moindre; (2) collégial, ou (3) universitaire.
- `statut`: statut matrimonial, soit célibaire (0) ou en couple (1).

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 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.

2.7 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.

2.8 **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.

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`)

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

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`)

2.9 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 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

2.10 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 2?

(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 2 and 3.

	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`.

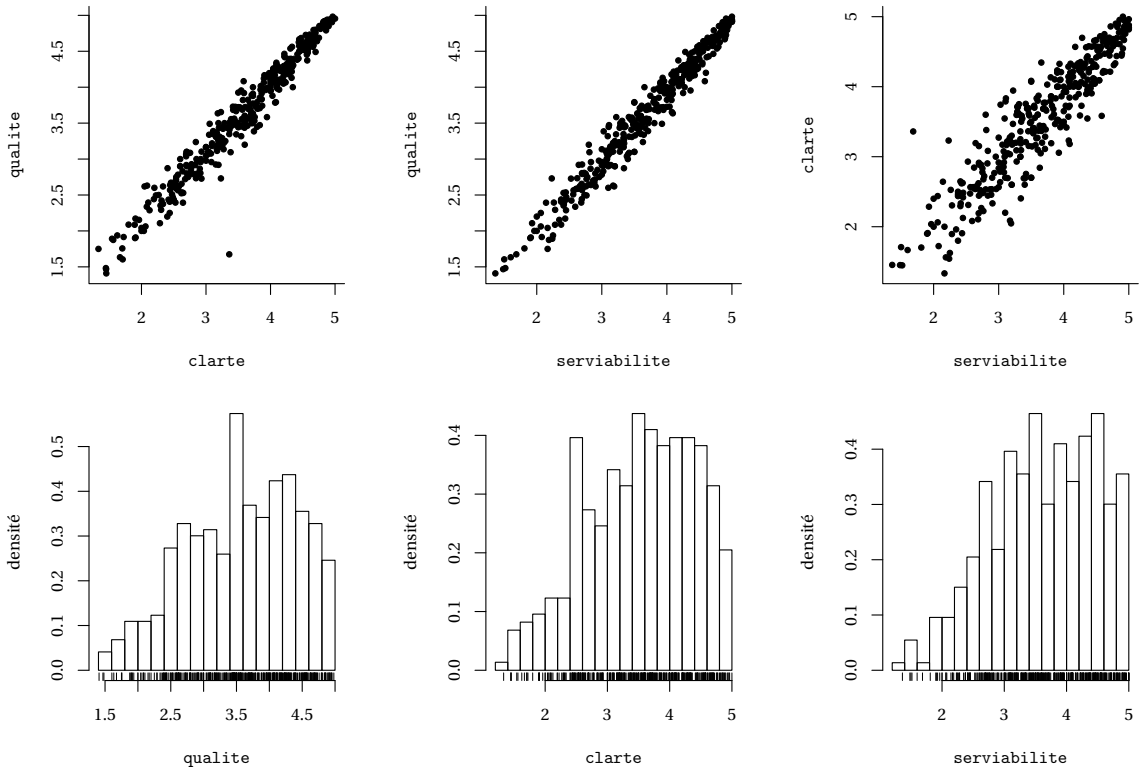

Figure 2: Panneau supérieur: nuage de point des paires (les corrélation linéaires de gauche à droite sont égales à 0.98,0.98 et 0.92). Panneau inférieur: histogramme des scores moyennes des indicateurs `qualite`, `serviabilite` et `clarte`.

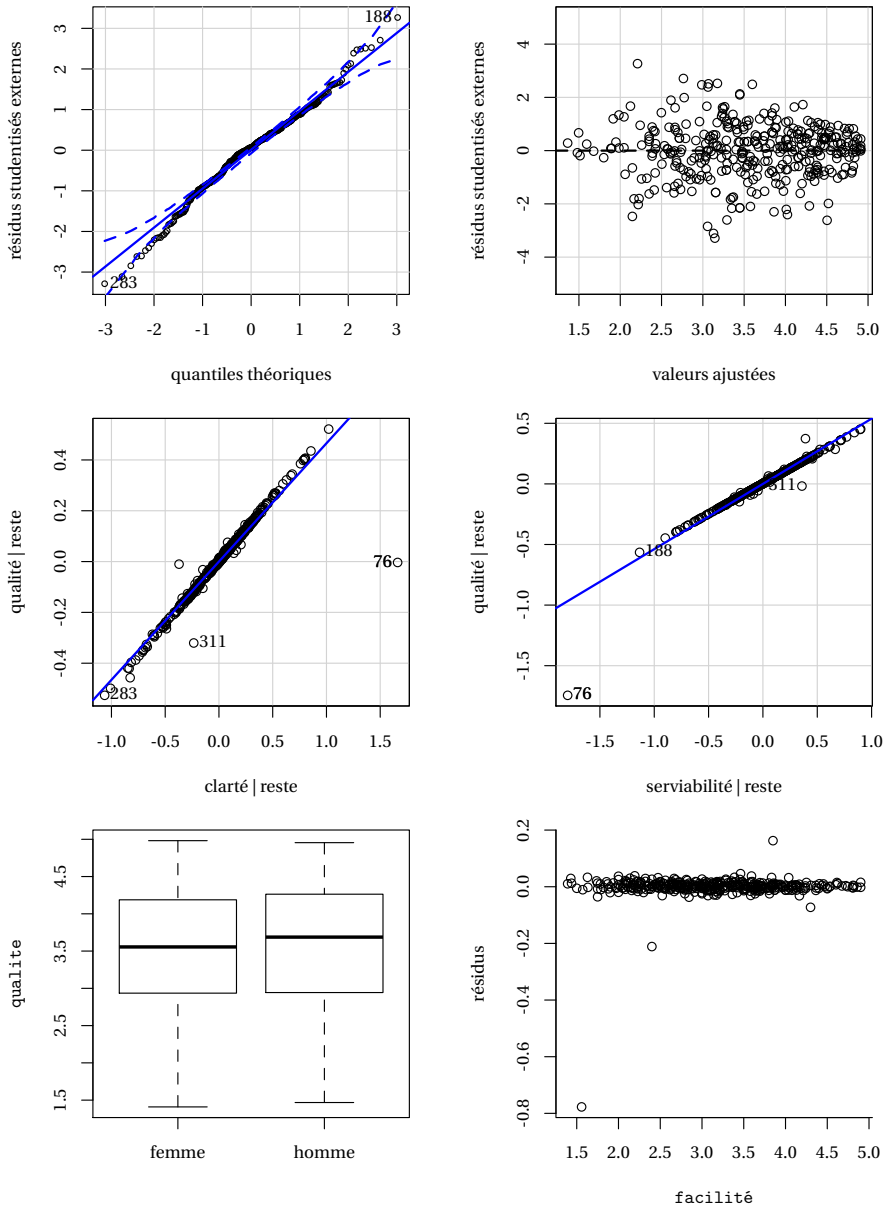

Figure 3: Diagnostics graphiques pour le modèle 4 ajusté aux données `Ratemyprofessor`. Panneau supérieur gauche: diagramme quantile-quantile des résidus studentisés externes, avec intervalles de confiance ponctuels à 95% (traitillés), en excluant l'observation 76. Panneau supérieur droit: diagramme 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`.