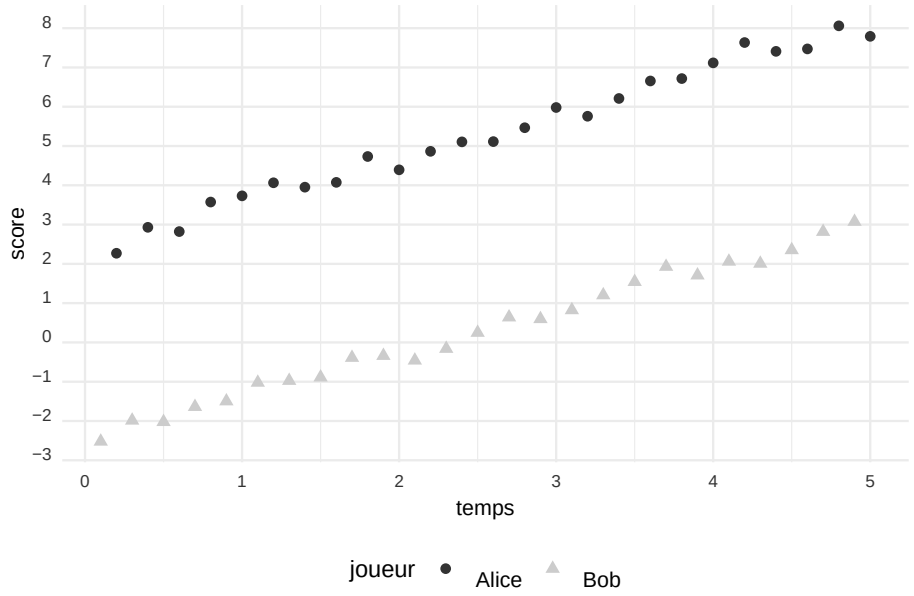

Figure 1: Scores d'Alice et de Bob en fonction du nombre d'heures de jeu.

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 $\widehat{\beta}_1, \ldots, \widehat{\beta}_3$?

Solution
On peut déduire l'ordonnée à l'origine et la pente à partir de la Figure 1 et reparamétriser le modèle. L'équation de la pente pour Alice est $2.5 + 1.1\texttt{temps}$ et celle de Bob est $-2.5 + 1.1\texttt{temps}$. Le paramètre $\widehat{\beta}_0$ correspond à l'ordonnée à l'origine de la catégorie de référence, -2.5, et la pente $\widehat{\beta}_1$ est celle de la personne de référence, 1.1. Les autres paramètres sont la différence moyenne entre l'ordonnée à l'origine/la pente du score d'Alice moins ceux de Bob, à savoir ($\widehat{\beta}_2 = 5, \widehat{\beta}_3 = 0$). Il suffit maintenant de considérer le signe des coefficients et on déduit que $\widehat{\beta}_0 < 0, \widehat{\beta}_1 > 0, \widehat{\beta}_2 > 0$ et $\widehat{\beta}_3 = 0$.

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:

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

Compilé le 24/09/2020 à 12:56

où

$$X_1 = \begin{cases} 0, & \text{si la femme n'a pas de diplôme secondaire,} \\ 1, & \text{si la femme a un diplôme secondaire mais pas de diplôme collégial,} \\ -1, & \text{si la femme a un diplôme collégial.} \end{cases}$$

$$X_2 = \begin{cases} 0, & \text{si la femme n'a pas d'enfant,} \\ 1, & \text{si la femme a 1 ou 2 enfants,} \\ -1, & \text{si la femme a 3 enfants ou plus.} \end{cases}$$

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)?

Solution
On modélise la moyenne de chaque groupe (analyse de variance à deux facteurs, sans interaction). Puisqu'on a le même nombre de femmes dans chaque catégorie, la moyenne globale est la somme de chacune des catégories, soit $\widehat{\beta}_0$. L'équation de la moyenne ajustée pour la catégorie de référence en (i) est $\widehat{\beta}_0 - \widehat{\beta}_1 - \widehat{\beta}_2$ et la différence est donc $-\widehat{\beta}_1 - \widehat{\beta}_2$.

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

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

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.

Solution
Aucune de ces réponses. Les estimateurs des moindres carrés minimisent la distance moyenne, donc celle-ci ne peut être systématique et l'énoncé «toujours moins chères» ne tient pas la route; idem pour la localisation (on ne sait rien sur la corrélation entre localisation et les autres variables explicatives. Le toutes choses étant égales par ailleurs ne veut rien dire parce qu'on ne peut fixer garage sans affecter superficie: si la variable `garage` = 1 en l'absence de garage, alors `superficie` = 0 pour cette même maison. La différence de prix entre deux maisons

d'une même localisation est $\beta_1 - \beta_2 \texttt{surface}$ et sans connaître la surface moyenne des maisons avec garage et la proportion de maisons, on ne peut rien conclure puisque $\beta_1 > 0$ et $\beta_2 > 0$.

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 $\widehat{\beta_1^i} = C\widehat{\beta_1^m}$?

Solution
Attention: la conversion est contre-intuitive 10 USD = 13.9 CAD, donc il faut multiplier le montant USD par 1.39 pour obtenir l'équivalent en dollars canadiens. En substituant les nouvelles variables dans l'équation, on obtient

$$\texttt{prix}^{\texttt{CAD}} = 1.39\texttt{prix}^{\texttt{USD}} = 1.39\beta_0^i + \frac{1.39}{1.61}\beta_1^i \texttt{distance}^{\texttt{km}} + 1.39\varepsilon^i.$$

pour l'équation en fonction de $\widehat{\beta_1^m} = 1.39/1.61\beta_1^i$; on déduit $C = 1.158 = 1.61/1.39$. Pour ce genre d'exercices, on peut facilement simuler des données et faire les conversions d'unités et valider le résultat.

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.

Solution
Les graphiques dans la Figure 2 montrent la droite des valeurs ajustés pour les deux modèles de régressions et les diagrammes des résidus. Il y a une structure résiduelle dans le modèle `production ~ velocite`, puisque les plus petites valeurs des résidus apparaissent pour les plus petites et plus grandes valeurs de vitesse du vent, suggérant que l'effet est nonlinéaire. Il y a par contraste moins de structure dans le modèle avec la réciproque, qui capture davantage de la variabilité puisque son coefficient de détermination R^2 est de 0,98, comparativement à 0,87 pour le premier modèle. Notez que, dans le second modèle, l'ordonnée à l'origine correspond à des vents de force infinie.

(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.

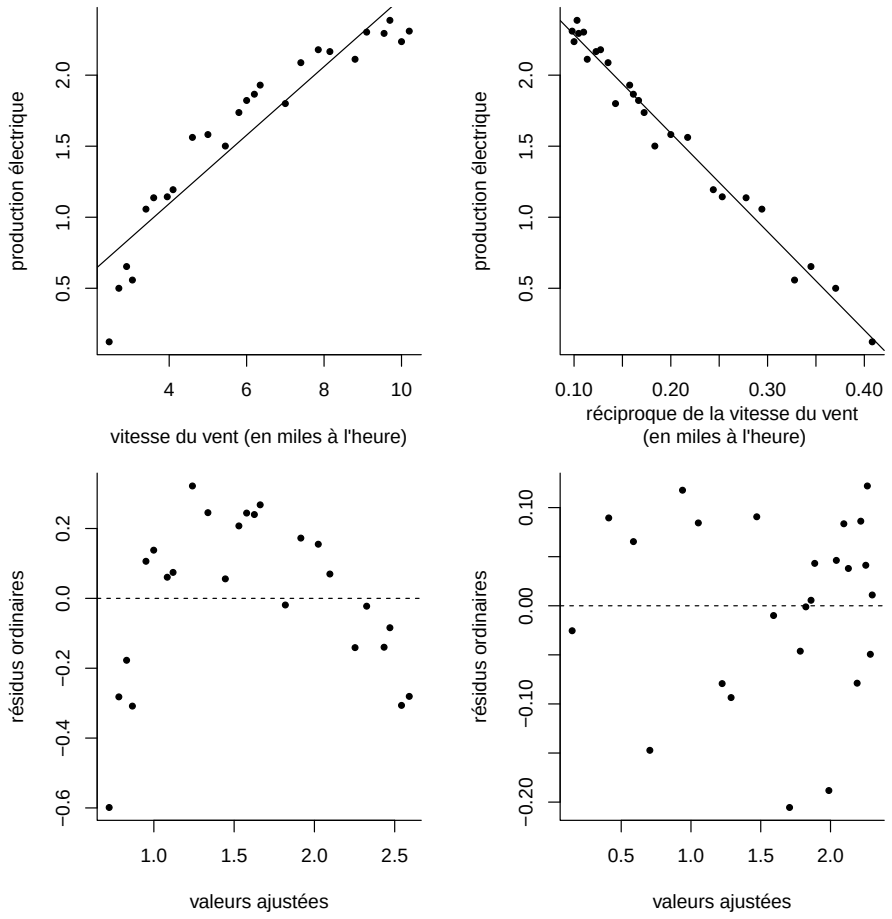

Figure 2: Panneau supérieur: droite de régression ajustée pour la production électrique en fonction de la vitesse du vent (gauche) des vents et de la réciproque (droite). Panneau inférieur: graphique des résidus et des valeurs ajustées.

Solution
La production prédite est de 1.34 unité pour le premier modèle, contre 1.59 pour le deuxième qui utilise la vitesse du vent inverse. Les deux intervalles se chevauchent, mais le second [1.39, 1.79] est considérablement plus étroit que le premier de [0.84, 1.84].

(c) [⋆] La production électrique de l'éolienne devrait être inexistente en l'absence de vent, mais cette réalité n'est 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-il si vous retirez l'ordonnée à l'origine?

Solution
Si vous retirez l'ordonnée à l'origine du modèle, la moyenne des résidus ordinaire n'est plus zéro et R utilise un autre critère que le R^2 — le résultat du test-F pour la significativité globale du modèle n'a plus aucun sens. Même si le coefficient de l'ordonnée à l'origine, β_0, n'est pas significativement différent de zéro, on pourrait justifier sa présence par les erreurs de mesures de Y — le modèle n'est pas conçu pour être extrapolé au-delà de l'étendue de la variable explicative.

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

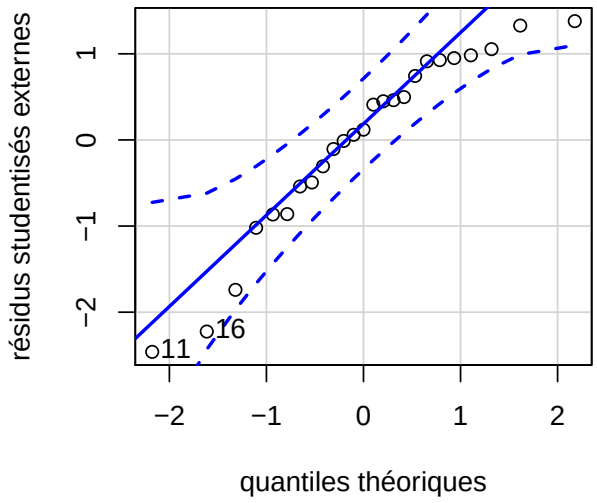

Figure 3: Diagramme quantile-quantile Student des résidus studentisés externes, avec intervalles de confiance ponctuels à 95% simulés.

Solution

Il n'y a aucun indice dans la Figure 3 qui laisse à penser que l'hypothèse de normalité n'est pas respecté, même si les plus grandes et plus petites valeurs sont plus petites que prévues (caractéristiques d'une loi asymmétrique). Tous les points sont à l'intérieur des intervalles de confiance ponctuels.

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.

Solution
Soit $\texttt{revenu}_i$ ($i = 1,2$) des variables binaires égales à 1 si `revenu` $= i$ et zéro autrement. Le modèle linéaire est

$$\texttt{intention} = \beta_0 + \beta_1 \texttt{revenu}_1 + \beta_2 \texttt{revenu}_2 + \varepsilon,$$

où β_0 est la moyenne du groupe de référence (`revenu` = 3) et β_1, β_2 sont des effets différentiels pour les groupes 1 et 2, c'est-à-dire la différence moyenne entre les individus de la classe du revenu i ($i = 1,2$) par rapport à celle du groupe 3.

(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$.

Solution
L'intention d'achat prédite est 7.116, soit la moyenne de revenu du groupe 3. Cette valeur est l'ordonnée à l'origine.

(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.

Solution
Le modèle est identique à celui ajusté en (a) si la catégorie de référence est la même (`revenu=3`).

(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.

Solution
La variable `revenu` est continue et l'ordonnée à l'origine n'a pas d'interprétation. Puisque les groupes à faible revenu ont une intention d'achat plus élevée, la pente β_1 représente la différence entre les groupes 1–2 et 2–3. Cette différence, -1.24, est constante selon le modèle. Plutôt que $k-1$ paramètres supplémentaires pour une variable catégorielle à k niveaux, il y a un seul paramètres représentant la différence entre groupe. Si on change les étiquettes numériques, le modèle ajusté changerait possiblement.

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

Solution
Les coefficients β des variables catégorielles représentent la différence moyenne entre catégories selon la classe de revenu (relativement à la catégorie de base (groupe 3), *ceteris paribus*. Les coefficients estimés sont 1.7 et 0.24 pour les différences avec les groupes 1 et 2, respectivement. Dans le modèle qui n'incluait que revenu, les estimés étaient plus grands, soit 2.48 et 1.19.

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

Solution
Le test-F pour `revenu` vaut 4.86 et la valeur-p associée est 0.0095. On rejette l'hypothèse nulle que le revenu n'est pas un prédicteur utile à niveau 5%. Au contraire, `educ` n'est pas statistiquement significatif (statistique de test-F égale à 1.45, valeur-p de 0.24).

2.7 Le jeu de données `automobile` contient des informations sur 392 voitures. On considère un modèle linéaire liant l'autonomie (en miles au gallon) des voitures en fonction de leur puissance (en watts).

(a) Tracez un nuage de point illustrant la relation entre l'autonomie d'essence (`autonomie`) 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

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

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=modstat.automobile;
model autonomie=puissance puissance*puissance/ss3 solution;
run;
```

et en R via

```
data(automobile, package = "hecmodstat")
lm(autonomie~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.

Solution

Le nuage de point montre que la relation entre puissance et distance parcourue est négative et que les voitures plus puissantes consomment davantage d'essence pour une même distance parcourue. La relation est nonlinéaire.
En ajustant le modèle linéaire simple avec `puissance`, on obtient un coefficient linéaire significatif et un R^2 de 0.6. L'analyse du graphique des résidus ordinaires en fonction des valeurs ajustées montre que le modèle n'arrive pas à capter la nonlinéarité, contrairement aux modèles d'ordre supérieur. Le terme quadratique du modèle d'ordre 2 est statistiquement significatif et une analyse des résidus montre qu'il n'y a pas de structure résiduelle dans la moyenne (malgré une apparence d'hétéroscédasticité). La différence d'ajustement entre le modèle cubique et le modèle quadratique est à peine visible aux extrémités de `puissance`; le coefficient n'est pas significatif à niveau 5% (valeur-p de 0.36) et on conclut que le modèle quadratique est une simplification adéquate du modèle cubique. Les queue des résidus sont légèrement plus lourdes qu'attendues pour un échantillon normal.

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.

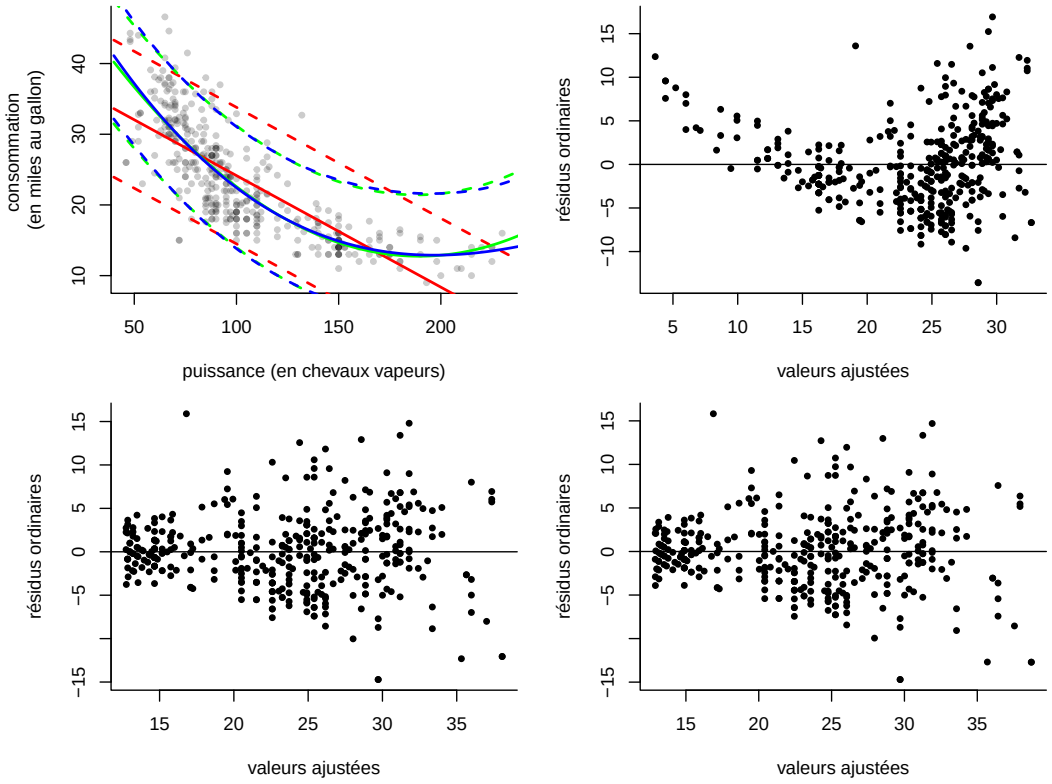

Figure 4: Nuage de point de la distance parcourue par volume d'essence en fonction de la puissance du véhicule pour les données `auto`; les droites superposées montrent l'ajustement du modèle de degré 1 (rouge), du degré 2 (vert) et du degré 3 (bleu), avec des intervalles de prédictions ponctuels à 95%. Les graphiques des résidus en fonction des valeurs ajustées sont pour le modèle linéaire (en haut à droite), le modèle quadratique (en bas à gauche) et le modèle cubique (en bas à droite).

Solution

$$\widehat{\texttt{intention}} = 5.53 + 1.41\texttt{educ1} - 1.4\texttt{educ2} + 1.097\texttt{fixation}.$$

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.

Solution

Lorsque `educ=1` la variable `educ1` vaut 1 et `educ2` vaut zéro et réciproquement quand `educ=2`. Les trois équations sont

$$\widehat{\texttt{intention}} = \begin{cases} 6.92 + 1.09\texttt{fixation}, & \text{si } \texttt{educ=1} \\ 6.93 + 1.09\texttt{fixation}, & \text{si } \texttt{educ=2} \\ 5.53 + 1.09\texttt{fixation}, & \text{si } \texttt{educ=3}. \end{cases}$$

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 ?

Solution
On peut voir que l'ajustement des trois droites sur les trois types de points (pour chaque catégorie de `educ`) n'est pas bon. La traîne du groupe 2 est trop forte, celle du groupe 3 trop faible. Un meilleur modèle inclurait possiblement inclure trois droites avec trois pentes différentes.

(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.

Solution

$$\widehat{\texttt{intention}} = 4.58 + 1.45\texttt{educ1} + 3.32\texttt{educ2} + 1.75\texttt{fixation} - 0.11\texttt{educ1} \cdot \texttt{fixation} - 1.25\texttt{educ2} \cdot \texttt{fixation}.$$

L'interaction est significative à niveau 5% (valeur-p de 0.02); le modèle avec interaction est plus adéquat que le modèle sans interaction.

ii. Calculez les trois équations des droites estimant la relation entre `intention` et `fixation` lorsque `educ=1`, `educ=2` et `educ=3`, respectivement.

Solution

$$\widehat{\texttt{intention}} = \begin{cases} 6.03 + 1.64\texttt{fixation}, & \text{si } \texttt{educ=1} \\ 7.91 + 0.5\texttt{fixation}, & \text{si } \texttt{educ=2} \\ 4.59 + 1.75\texttt{fixation}, & \text{si } \texttt{educ=3}. \end{cases}$$

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.

Solution
On peut voir que l'ajustement des trois types de points sur les trois droites est meilleur; dans le premier graphiques, les droites ajustées pour les groupes 1 et 2 se chevauchaient, mais la pente du groupe 2 est maintenant plus faible.

(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.

Solution

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`)

Solution
Si on soustrait les lignes (6) et (3) on obtient le coefficient β_3, qui représente donc l'augmentation moyenne de l'intention d'achat lorsque fixation augmente de 1 seconde et que la catégorie d'éducation est égale à 3 (universitaire).

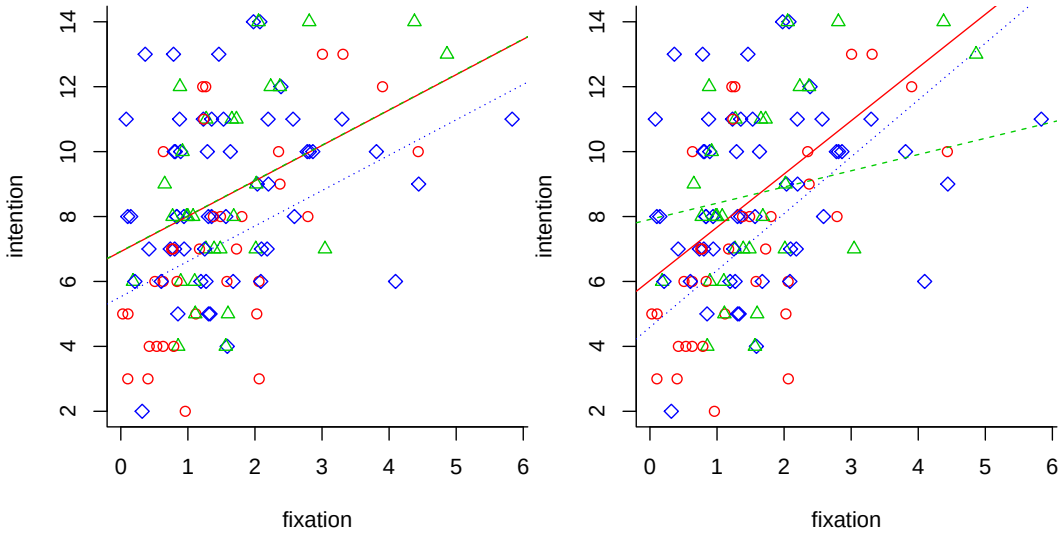

Figure 5: Droites ajustées pour le modèle sans interaction (gauche) et avec l'interaction (droite). Les observations et les droites sont colorées par groupe, (cercles et ligne rouges pour `educ=1`, triangles et traitillés verts pour `educ=2`, losanges et pointillés bleus pour `educ3`).

scénario	`fixation`	`educ`	**intention d'achat moyenne selon le modèle**
1	x	1	$(\beta_0+\beta_1)+(\beta_3+\beta_4)x$
2	x	2	$(\beta_0+\beta_2)+(\beta_3+\beta_5)x$
3	x	3	$\beta_0+\beta_3 x$
4	$x+1$	1	$(\beta_0+\beta_1)+(\beta_3+\beta_4)(x+1)$
5	$x+1$	2	$(\beta_0+\beta_2)+(\beta_3+\beta_5)(x+1)$
6	$x+1$	3	$\beta_0+\beta_3(x+1)$
7	0	1	$\beta_0+\beta_1$
8	0	2	$\beta_0+\beta_2$
9	0	3	β_0

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`)

Solution
Si on soustrait les lignes (9) et (7) on obtient le coefficient β_1, qui représente donc la différence moyenne d'intention entre les catégories d'éducation 1 et 3, lorsque la personne ne fixe pas le produit (`fixation` est égale à 0).

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`)

Solution
Si on soustrait les lignes (9) et (8) on obtient le coefficient β_2, qui représente donc la différence moyenne d'intention entre les catégories d'éducation 2 et 3, lorsque la personne ne fixe pas le produit (`fixation` est égale à 0).

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?

Solution
Avec les données originales, l'ordonnée à l'origine β_0 représente le traffic mensuel moyen en 1AD. Si on décale cette variable de 1949, l'ordonnée à l'origine devient le traffic en 1949. Dans les deux cas, la pente β_1 représente l'augmentation annuelle du traffic aérien mensuel.

(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?

Solution
Le modèle est

$$Y_i = \beta_0 + \beta_1 \texttt{time}_i + \sum_{j=2}^{12} \mathrm{S}_{ij}\beta_j + \varepsilon_i, \qquad i = 1,\ldots,n$$

où $\mathrm{S}_1,\ldots,\mathrm{S}_{12}$ sont des indicateurs binaires pour les douze mois de l'année; par exemple, S_2 est égale à 1 seulement pour le mois de février et zéro sinon.
Pour vérifier si l'amélioration de l'ajustement est significative, on fait un test F comparant le modèle sans et avec les 11 variables indicatrices. La statistique F vaut 27.589 et sa loi nulle est $\mathcal{F}(11,131)$; la valeur-p correspondante est négligeable et donc on conclut que l'ajout d'un effet mensuel mène à une amélioration notable 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.

Solution
Le nombre mensuel moyen prédit pour décembre 1962 est 501 263 passagers aériens.

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

Solution

On remarque que le modèle réussit bien à capter les caractéristiques essentielles. Le graphique des résidus en fonction des valeurs ajustées montre une relation quadratique; en regardant la courbe ajustée, on peut remarquer que l'amplitude de l'effet saisonnier croît avec les années. Cela fait en sorte que les valeurs prédites sont trop grandes au début des années 1950 et trop petites vers 1960. Cette augmentation de la variance pourrait être compensée par une transformation logarithmique, qui stabiliserait la variance. Une autre alternative serait d'inclure une interaction entre l'année et le mois, mais cela implique l'ajout de 11 paramètres. L'effet levier des premiers et derniers points est important, ce qui nous pousse à retourner à la planche à dessin.
En regardant les résidus studentisés externes en fonction du temps (plutôt qu'en fonction des valeurs ajustées), on peut remarquer une hétéroscédasticité persistante. L'hypothèse d'indépendance est également peu plausible au vu de la nature des données (série chronologique).

(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

Solution

La transformation logarithmique résoud le problème de non-normalité et le modèle capture mieux l'effet nonlinéaire et l'hétéroscédasticité détectés dans le modèle additif. Un postulat qui n'est toujours pas valide ici est celui d'indépendance entre erreurs: les résidus sont positivement corrélés. Ignorer cette dépendence résulte en des erreurs-type pour les paramètres qui sont trop précises par rapport à la réalité (la quantité d'information effective à disposition est moindre que n).

2.10 Le jeu de données `Ratemyprofessor` contient les 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.

Solution

L'ordonnée à l'origine du modèle 1, 3.532 points, donne la moyenne du score de qualité pour les femmes.

(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.

Solution

Le modèle est paramétrisé avec des effets différentiels, et les femmes sont dans la catégorie de référence. L'ordonnée à l'origine pour les hommes est $(-0.054 + 0.048)$, et il faut ajouter à ce terme la contribution des scores et l'interaction `homme:serviabilite`. Cela donne un score moyen prédit de

$$(-0.054 + 0.048) + (0.541 + 0.466 + 0.0007 - 0.013) \times 4 = 3.9728.$$

(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.*

Solution

L'hypothèse nulle est que tous les paramètres $\boldsymbol{\beta}$ sont nuls à l'exception du paramètre de l'ordonnée à l'origine, soit $(\beta_{\texttt{serviabilite}}, \beta_{\texttt{clarte}}) = \mathbf{0}_2$ versus la contre-hypothèse $(\beta_{\texttt{serviabilite}}, \beta_{\texttt{clarte}}) \neq \mathbf{0}_2$. On rejette l'hypothèse nulle que le modèle avec une seule moyenne décrit adéquatement les données (le quantile 95% de la loi nulle Fisher $\mathcal{F}(2,363)$, 3.021, est plus petit que 62228.971, la valeur observée de la statistique F pour le test de significativité globale).

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

Solution

L'intervalle de confiance pour le paramètre β est basé sur le test-t et la loi asymptotique est Student; puisque $n-p$ est grand, on peut remplacer le 97.5% de la loi Student par celui de la loi normale, 1.96. On obtient l'intervalle de confiance approximatif

$$\widehat{\beta}_2 \pm \mathsf{se}(\widehat{\beta}_2) \times 1.96 = 0.466 \pm 0.007 \times 1.96 = [0.4523, 0.4797].$$

Puisque l'intervalle ne contient pas zéro, le coefficient est significatif et le modèle 2 n'est pas une simplification adéquate du modèle 4 — le test-t pour un coefficient et le test-F mènent à la même inférence.

(e) Contrastez les coefficients estimés pour les modèles 2 et 4. Est-ce que ces estimés sont cohérents avec les graphiques de la Figure 6?

Solution

On a rejetté l'hypothèse nulle que le modèle 2 est une simplification adéquate du modèle 4. Les deux variables explicatives, `serviabilite` et `clarte`, sont très fortement corrélées comme illustré dans le panneau supérieur droit de la Figure 6. Si on suppose que le modèle 4 est le "vrai" modèle, celui qui a servi à générer les données, il est clair que la variable codeserviabilite va capturer la majeure partie de l'effet de `clarte` parce qu'ils sont presque parfaitement colinéaires — c'est ce qui resort d'une comparaison des coefficients. Dans les deux cas, les coefficients de `serviabilite` ou de `clarte` sont néanmoins significativement différents de zéro, une fois qu'on a contrôlé pour les autres; cela est illustré dans les diagrammes de régression partielle (milieu de la Figure 7).

(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.

Solution

C'est un modèle dans lequel l'interaction `homme:serviabilite` est incluses, mais sans l'effet principal. Cela veut dire que la pente de `serviabilite` est nulle pour les femmes. Cette contrainte n'est pas justifiée et, si on change la catégorie de référence pour hommes, on obtiendrait un modèle différent et l'inférence changerait.

(g) Quelles sont les postulats du modèle linéaire? Commentez sur la validité sur la base des graphiques présentés dans les Figures 6 and 7.

Solution

Les postulats sont (1) l'indépendance des erreurs (2) la linéarité, (3) homoscédasticité et (4) la normalité des erreurs.

i. Linéarité, ou spécification du modèle de moyenne: l'ajustement est globalement excellent, la relation entre `qualite` et `clarte`, ou entre `qualite` et `serviabilite` est linéaire à en croire la Figure 6. Le panneau supérieur droit de la Figure 7 ne montre aucune structure moyenne résiduelle, tandis que le diagramme du panneau inférieur droit montre qu'il n'y a pas de relation linéaire avec `facilite` (si la droite de régression a une pente non-nulle, cela serait exclusivement dû aux valeurs aberrantes).

ii. Homoscédasticité: aucune différence de variance entre hommes et femmes (boîtes à moustches, panneau inférieur gauche de la Figure 7). Il y a un peu d'hétéroscédasticité; les étudiant(e)s sont plus unanimes pour les enseignant(e)s qui obtiennent de très bons scores que pour d'autres qui ont de moins bon scores (graphique des résidus studentisés externes, panneau supérieur droite de la Figure 7).

iii. Indépendance: plausible de par le plan d'experience pourvu que les enseignant(e)s exclus (qui sont dans de plus petites classes ou pour qui le taux de réponse est plus faible), ne soient pas systématiquement meilleurs ou moins bons.

iv. Normalité: le diagramme quantile-quantile dans le panneau supérieur gauche de la Figure 7 semble correct; la queue de gauche est plus lourde à cause de l'asymmétrie de la réponse et du fait que le support de la loi est bornée, mais l'impact de cette trouvaille est nul car la taille de l'échantillon est grande. On s'attend de toute façon que 1 point sur 20 sorte des intervalles de confiance ponctuels à 95%, mais comme les quantiles sont ordonnés, les points adjacents risquent également d'être hors de l'intervalle.

v. Valeur aberrante: la valeur 76 est clairement inhabituelle et biaise la droite de régression. Sans précision sur son effet de levier ou sa distance de Cook, impossible d'en dire plus.

	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`.

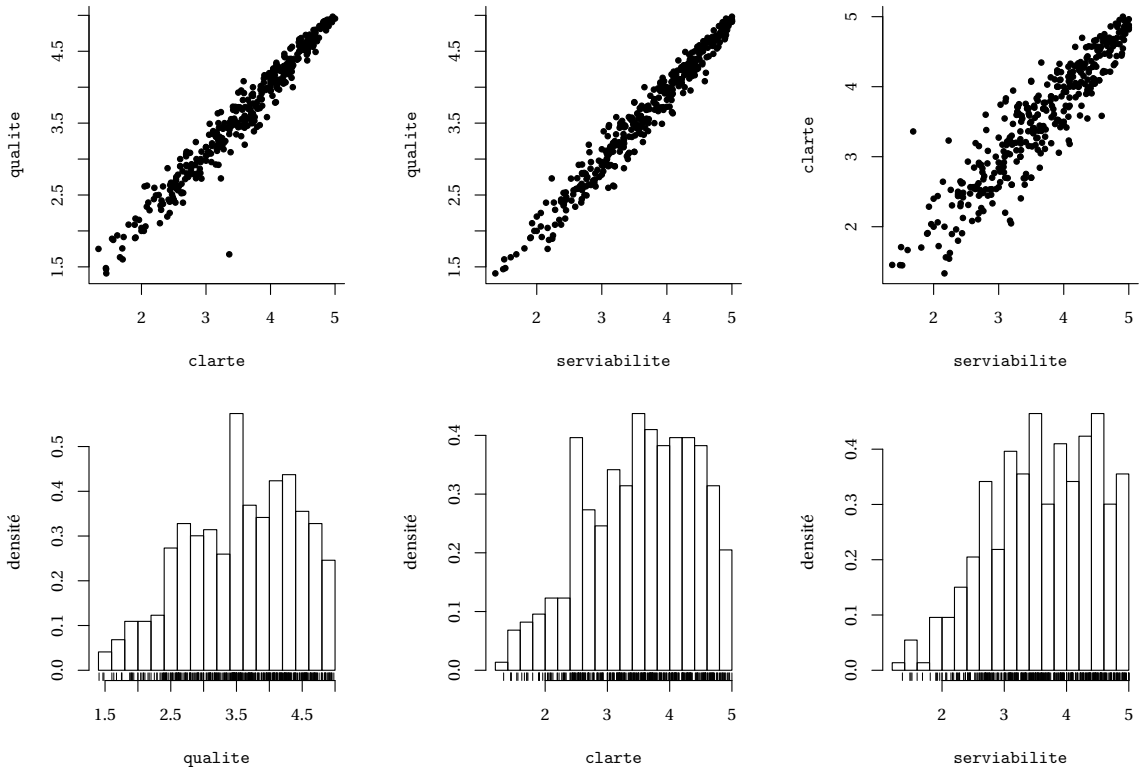

Figure 6: Panneau supérieur: nuage de point des paires (les corrélation linéaires de gauche à droite sont égales à 0.98,0.98 et 0.92). Panneau inférieur: histogramme des scores moyennes des indicateurs `qualite`, `serviabilite` et `clarte`.

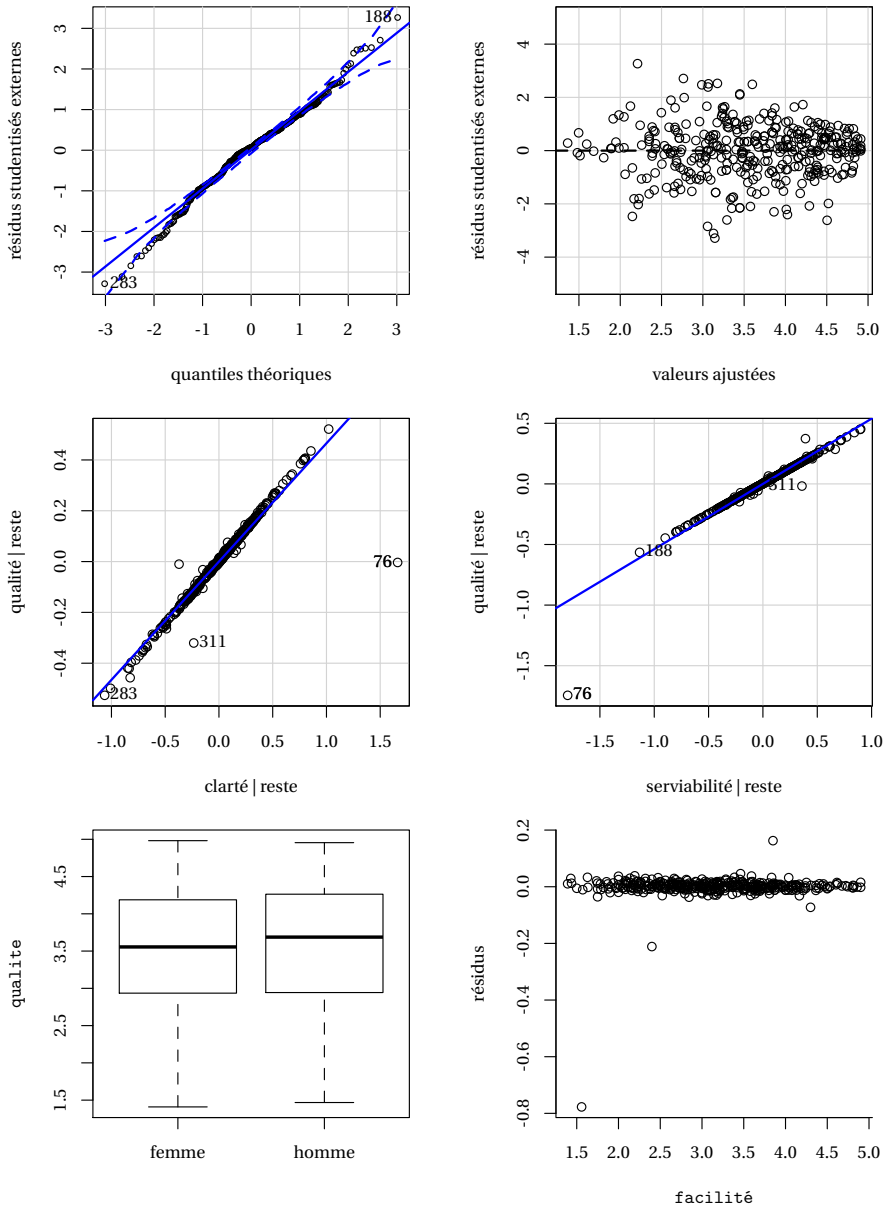

Figure 7: Diagnostics graphiques pour le modèle 4 ajusté aux données `Ratemyprofessor`. Panneau supérieur gauche: diagramme quantile-quantile des résidus studentisés externes, avec intervalles de confiance ponctuels à 95% (traitillés), en excluant l'observation 76. Panneau supérieur droit: diagramme 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`.