3.1 Le jeu de donnée `croissance` contient la masse (en livres) et la taille (en pouces) de 19 adolescents. Notre objectif est de prédire la masse d'adolescents sur la base de leur taille à l'aide d'un modèle linéaire simple.

(a) Calculer les statistiques descriptives de l'échantillon (moyenne, écart-type, quartiles, etc.) des deux variables et inspecter visuellement l'effet de la taille sur la masse à l'aide d'un nuage de points.

Solution
Il y a apparence d'une relation linéaire forte et positive entre les variables masse et taille (à mesure que l'adolescent grandit, sa masse augmente).

(b) Calculer le coefficient de corrélation entre la taille et la masse et interprétez ce dernier.

Solution
Le coefficient de corrélation linéaire estimé est $r = 0.88$, ce qui confirme la présence d'une relation linéaire forte entre les variables.

(c) Ajustez un modèle de régression linéaire pour prédire la masse en fonction de la taille. Écrivez l'équation du modèle de régression ainsi que celle du modèle ajusté.

Solution
Soit Y la masse (en livres) et X la taille (en pouces) de l'ado. L'équation du modèle de régression est

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad \varepsilon_i \sim \text{IID}(0, \sigma^2).$$

Le modèle ajusté est $\widehat{y} = -143,0 + 3,9x$.

(d) Interprétez les paramètres du modèle, β_0 et β_1.

Solution
L'ordonnée à l'origine β_0 n'a pas d'interprétation, puisque qu'un humain ne peut pas avoir une taille nulle (encore moins une masse négative). La pente représente l'effet d'une augmentation de la `taille`, soit $\beta_1 = \mathsf{E}(Y \mid X = x+1) - \mathsf{E}(Y \mid X = x)$; si un ado croît d'un pouce, sa masse augmente en moyenne de 3.9 livres.

(e) Quelles sont les hypothèses nulles et alternatives des tests-t de la sortie? Écrivez les conclusions de ces tests.

Solution
L'hypothèse nulle pour le paramètre β_j ($j = 0,1$) est $\mathscr{H}_0 : \beta_j = 0$ contre l'alternative $\mathscr{H}_1 : \beta_j \neq 0$. On rejette l'hypothèse nulle que les deux paramètres sont nuls à un niveau de 5%, ce qui veut dire que la pente et l'ordonnée à l'origine sont chacune différente de zéro sachant l'autre. La variable `taille` a un effet significatif sur la `masse` de l'adolescent(e).

(f) Donne un intervalle à 95% pour la pente et interprétez ce dernier.

Solution
Un intervalle de confiance basé sur la loi t asymptotique de la statistique de Wald est $(2,81; 4,99)$. Puisque zéro n'est pas inclus dans l'intervalle, on conclut que l'effet linéaire de la taille sur la masse est significativement différent de zéro. 95% des intervalles ainsi construits devraient contenir le vrai effet moyen de la taille sur la masse.

(g) Prédisez la masse de deux adolescents hypothétiques de 70,5 et 71,5 pouces. Que représente la différence entre ces valeurs?

Solution
La différence entre les deux prédictions est $135,754 - 131,855 = 3,899$ livres, soit la valeur de $\widehat{\beta}_1$. C'est l'"effet de la taille sur la masse": en moyenne, la masse augmente de 3,899 livres pour chaque pouce supplémentaire.

Compiled 29/01/2020 at 22:54

(h) Fournissez des intervalles de confiance pour la masse moyenne et la masse prédite pour deux individus de taille 70,5 et 71,5 pouces. Commentez sur les différences entre ces deux intervalles.

Solution
L'intervalle de confiance à 95% pour la masse moyenne d'un individu de 70,5 pouces est $[121,437; 142,273]$ et de $[124,393; 147,115]$ for 71,5 pouces. L'intervalle de prédiction à 95% sont respectivement $[105,979; 157,730]$ et $[109,485; 162,023]$ — ils sont plus larges parce que la variabilité de la prédiction inclut celle de l'erreur ε; il y a de ce fait moins de variaiblité quand on estime la moyenne qu'une observation inconnue.

(i) Supposez qu'on ajuste un modèle de régression pour la masse (en kg) et la taille (en cm). Est-ce que le modèle linéaire serait toujours adéquat? Expliquez en quoi le changement d'unité affecterait les estimés de vos paramètres.

Solution
Un modèle linéaire serait toujours approprié, puisque le changement est une transformation linéaire, qui ne fait qu'affecter les estimés des paramètres – graphiquement, cela revient à étirer l'axe des abcisses et celui des ordonnées. Le changement d'unité de la taille (X) ne change que β_1, tandis que le changement en Y impacte à la fois β_0, β_1 et σ^2: puisque l'on multiplie X par 2,54 pour obtenir la taille en cm, $\widehat{\beta}_1$ décroît par ce même facteur. Une livre vaut 0,4536 kg, on obtient donc $Y_i = 0,4536 \times (\beta_0 + X_i/2,54 + \varepsilon_i)$; les estimés des coefficients dans le nouveau système de coordonnées sont $\widehat{\beta}_0 = -64,877$ kg et $\widehat{\beta}_1 = 0.696$ kg/cm et l'écart-type des résidus passe à $s = 5.092$ au lieu de 11.23.

3.1 Les données `eolienne` contiennte 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 1 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.

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-ils i vous retirez l'ordonnée à l'origine?

Compiled 29/01/2020 at 22:54

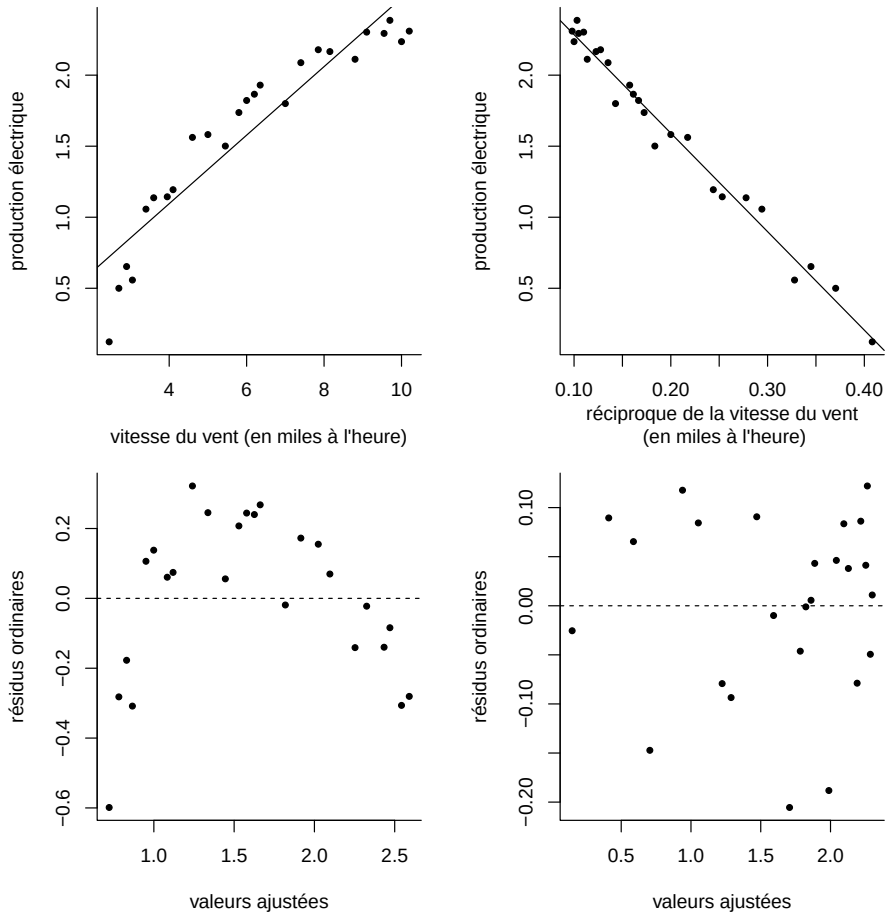

Figure 1: Panneau supérieur: droite de régression ajustée pour la production électrique en fonction de la vitesse du vent (gauche) des vents et de la réciproque (droite). Panneau inférieur: graphique des résidus et des valeurs ajustées.

Solution
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.

Solution
Il n'y a aucun indice dans la Figure 2 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.

3.1 Le jeu de données `auto` contient les caractéristiques de 392 voitures. Nous sommes intéressés à modéliser la `consommation` d'essence (en miles par gallon) de voitures en fonction de leur `puissance` (en watts).

(a) Ajustez un modèle linéaire pour les données, regardez le diagramme quantile-quantile des résidus studentisés externes et commentez sur l'adéquation du postulat de normalité.

Compiled 29/01/2020 at 22:54

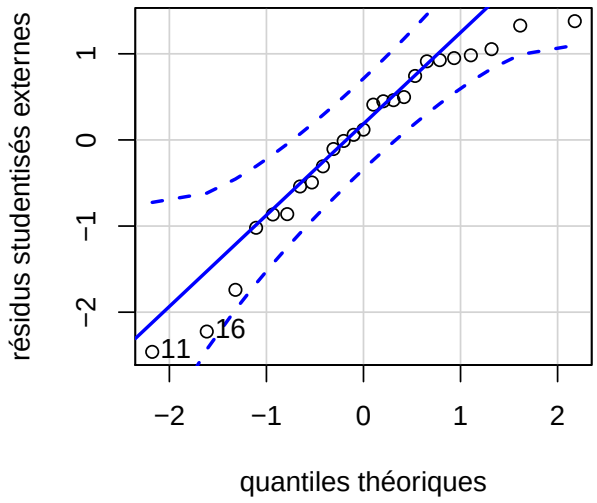

Figure 2: Diagramme quantile-quantile Student des résidus studentisés externes, avec intervalles de confiance ponctuels à 95% simulés.

Solution

Le postulat de normalité ne tient pas, avec des résidus systématiquement plus grands qu'attendus sous une loi Student (voir Figure 3). La queue lourde des résidus est essentiellement due à la surface de réponse non-linéaire qui n'est pas bien prise en compte par le modèle de régression.

(b) Produisez un nuage de points des résidus contre la variable puissance et commentez sur l'adéquation de l'hypothèse de linéarité.

Solution

Il y a une traîne résiduelle d'apparence quadratique entre la consommation d'essence et la puissance qui peut être visualisée dans la Figure 3. Cette dernière indique que le modèle est inadéquat.

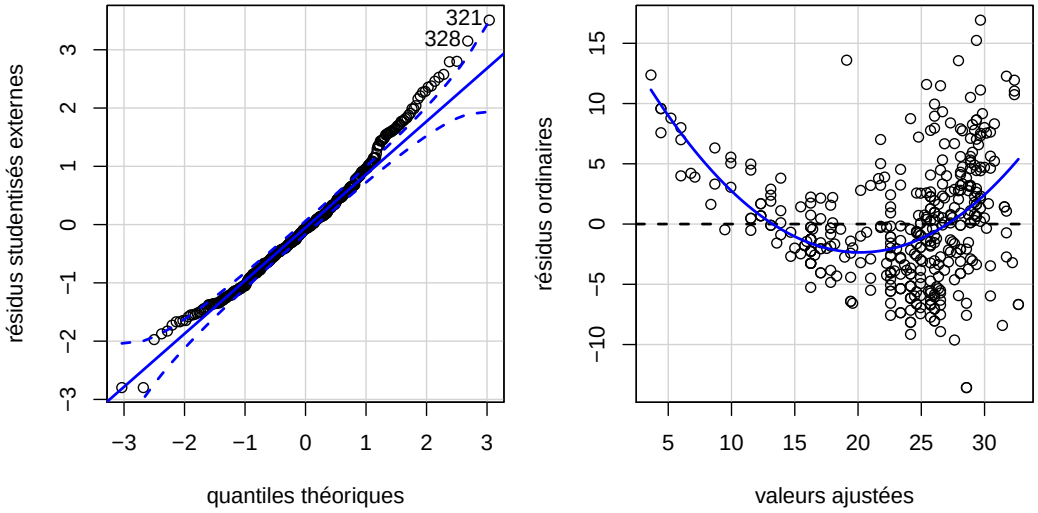

Figure 3: Gauche: diagramme quantile-quantile des résidus studentisés externes pour les données `auto`. Droite: graphique des résidus ordinaires contre les valeurs ajustées, avec courbe de lissage locale LOESS.

Compiled 29/01/2020 at 22:54