Statistiques - Modèles linéaires

JOFFRION Martin & KNOBLOCH Adrien

2023 - 2024

Question 1

- Hypothèse 1 : loi gaussienne des erreurs :

La forme slalomée de la répartition des résidus sur le qqplot indique un potentiel défault de Kurtosis, confirmé sur l'histogramme de répartition. Notre jeu de donnée souffre ainsi d'un excès de Kurtosis, acceptable au vu de la robustesse du modèle gaussien.

- Hypothèse 2 : indépendance des erreurs :

Notre modèle est déjà filtré sur une espèce, et l'ANOVA intègre les variables relevé et altitude, la seule variable encore susceptible de générer une dépendance des résidus est le statut, à savoir l'état de l'arbre au moment du prélèvement.

Le boxplot démontre l'absence de biais sur les résidus, et donc l'indépendance des individus par rapport à la variable statut.

- Hypothèse 3 : homoscédasticité des erreurs :

Sur le SL-plot, la moyenne lissée (courbe rouge pleine) des résidus standardisés est presque confondue avec la moyenne théorique attendue (courbe bleue), preuve de l'homoscédasticité des erreurs.

Le modèle *modAnovRegQuad2_Querc_BC3 avec effet de relevé et effet quadratique de l'altitude valide donc toutes les hypothèses du modèle linéaire.

Question 2

- Hypothèse 1 : loi gaussienne des erreurs :

A nouveau, la forme slalomée de la répartition des résidus sur le qqplot indique un potentiel défault de Kurtosis, aditionné à un légère asymétrie, confirmés sur l'histogramme de répartition. Notre jeu de donnée souffre d'un excès de Kurtosis et d'une légère asymétrie à gauche.

- Hypothèse 2 : indépendance des erreurs :

Notre modèle n'est plus filtré sur une espèce, mais l'ANOVA intègre toutes les variables potentiellement sources de dépendance des erreurs. Ainsi, les résidus ont un biais nul quant à la variable espèce, comme prouvé sur le boxplot.

- Hypothèse 3 : homoscédasticité des erreurs :

Sur le SL-plot, la moyenne lissée (courbe rouge pleine) des résidus standardisés s'éloigne fortement de la moyenne théorique attendue (courbe bleue). Les résidus ont une trop forte variance pour les relevés à faible diamètre et à fort diamètre moyen.

Inclure la variable espèce dans l'ANOVA n'est donc pas une solution pour combler l'influence des espèces. Le modèle ne valide pas complètement les hypothèses de départ. Au vu de la robustesse du test ANOVA, nous continuons l'analyse, et comme attendu, les résultats sont très médiocres.

Sur le boxplot final, on observe ainsi des moyennes empiriques de sous échantillons (points rouges) plus "éloignées" des sous échantillons que la moyenne empirique globlabe (point bleu), symptome de l'échec du modèle.

Question 3

- Hypothèse 1 : loi gaussienne des erreurs :

A nouveau, la forme slalomée de la répartition des résidus sur le qqplot indique un potentiel défault de Kurtosis important, aditionné à un asymétrie marquée, confirmés sur l'histogramme de répartition. Notre jeu de donnée souffre d'un excès de Kurtosis et d'une forte asymétrie à gauche, invalidant l'hypothèse de normalité.

- Hypothèse 2 : indépendance des erreurs :

Notre modèle filtré intègre toutes les variables potentiellement sources de dépendances des erreurs sauf celle du relevé. Nous allons donc tester la dépendance des résidus concernant cette variable.

Nous observons sur le boxplot obtenu de légers biais, notamment pour les relevés 12, 13 et 17 (triangle 1). L'indépendance n'est donc pas pleinement satisfaisante.

- Hypothèse 3 : homoscédasticité des erreurs :

Sur le SL-plot, nous constatons une variance trop faible pour les relevés à fort diamètre moyen. La moyenne lissée reste toutefois proche de la moyenne théorique.

Les résultats du test emboîté pour ce niveau d'agrégation ne sont pas concluants. Comparé au modèle complexe de relevé, le facteur F est très important (239.87) et la p-value très faible (<2.2e-16). La probabilité que le modèle de relevé augmente la valeur du coefficient de corrélation par rapport à notre modèle à triangle est très élevée, de l'odre de 99%.

La fusion des sous populations n'est donc pas efficace. Le modèle initial intégrant la variable relevé sans regroupement reste le plus cohérent.

Question 4

Comparé aux résultats obtenus avec la variable latitude, la variable *lasLog* (dernière coupe massive) explique bien mieux l'effet relevé. C'est notamment visible dans le Scale-Location plot, où la moyenne glissée suit presque parfaitement le moyenne théorique, à l'exception des individus aux forts diamètres.

Toutefois, l'effet des relevés persiste, comme illustré dans le dernier graphique.

La variable dernière coupe massive explique donc partiellement l'effet relevé.

Question 5

En sortie du graphique qqplot, nous obtenons une p-value de 0.9887. L'hypothèse nulle est donc acceptée : la distribution des résidus est uniforme.

En sortie du graphique testCategorical, nous obtenons une p-value de 0.9376. L'hypothèse nulle est donc acceptée : la distribution des résidus par relevé est uniforme.

En sortie du graphique testQuantiles, nous obtenons une p-value de 0.9946. L'hypothèse nulle est donc acceptée : la distribution des résidus est homogène

Les hypothèses du nouveau modèle sont ainsi vérifiées.

Analyse des résultats

Comme le cas étudié dans le cours, l'altitude a un intervalle de confiance "centré" autour de 0, avec une p-value supérieure au seuil de confiance de 95% fixé. L'hypothèse nulle est donc vraissemblablement vérifiée : l'altitude n'a pas d'impact sur la formation des cavités.

Dans le cas du DBH, les résultats sont similaires : l'intervalle de confiance est proche et "centré" autour de 0, avec une p-value de 0.7 largement supérieure à 0.05. A nouveau, l'hypothèse nulle est donc vraissemblablement vérifiée : le DBH n'a pas d'impact sur la formation des cavités.

Pour affirmer ces résultats, nous avons calculé le R^2 de McFadden, afin d'évaluer le progrès apporté en terme de vraisemblance par notre modèle avec les covariables DBH et altitude.

La valeur de R^2 obtenue est de 0.19, soit relativement faible. L'ajout des covariables DBH et altitude n'a donc pas significativement amélioré la qualité du modèle, confirmant ainsi les résultats obtenus précédemment. Dans le cours, le R^2 de McFadden obtenu avec un modèle avec altitude seulement était de 0.188, soit très légèrement inférieur à celui obtenu avec l'ajout du DBH, confirmant encore une fois le très faible impact du DBH sur le modèle.