Modélisation statistique

#2.i Diagnostics graphiques des résidus

Dr. Léo Belzile
HEC Montréal

Postulats de validité du modèle

On postule que les erreurs $\varepsilon_i \sim \mathbf{No}(0, \sigma^2)$ sont indépendantes et identiquement distribuées.

Implications

- indépendance
- linéarité
- homoscédasticité (égalité des variances)
- normalité

Postulats revisités

1. **Indépendance**: les erreurs $\varepsilon_1, \ldots, \varepsilon_n$ sont indépendantes, idem pour les observations)
2. **Linéarité**: l'espérance des erreurs est $\mathsf{E}(\varepsilon_i) = 0$ pour tout $i = 1, \ldots, n$.
   - ✚ cela implique que le modèle pour la moyenne est correctement spécifié, d'où $\mathsf{E}(Y \mid \mathbf{X}) = \beta_0 + \beta_1 \mathbf{X}_1 + \cdots + \beta_p \mathbf{X}_p$
   - ✚ toutes les variables explicatives importantes sont incluses dans le modèle.
   - ✚ leur effet (présumé linéaire) est adéquatement représenté par le modèle.
3. **Homoscédasticité**: la variance des erreurs est **constante** $\mathsf{Va}(\varepsilon_i) = \sigma^2$ pour $i = 1, \ldots, n$.
   - ✚ corollaire: la variance Y_i est constante et ne dépend pas de $\mathbf{X}$.
4. **Normalité**: les termes d'erreurs $\boldsymbol{\varepsilon}$ suivent une loi normale.

Diagnostics graphiques (par défaut)

- Dans la procédure `glm`, l'option `plots=diagnostics residuals(smooth)` permet d'obtenir une panoplie de diagrammes pour analyser les résidus (et faires des graphiques des résidus ordinaires contre les variables explicatives).

Dans **SAS**, on peut sauvegarder la sortie de `glm` à l'aide de la commande`output`.

- Dans l'extrait de code qui suit, on sauvegarde
  - les valeurs ajustées `ajustees`
  - les résidus ordinaires `reso`
  - les résidus studentisés externes `rsc` dans une base de données temporaire `residus`.

- Code SAS ✚ Sortie SAS (1) ✚ Sortie SAS (2)

```
ods graphics on;
proc glm data=modstat.intention
 plots=diagnostics residuals;
class sexe educ revenu;
model intention=fixation emotion
 sexe age revenu educ statut / ss3 solution;
output out=residus predicted=ajustees r=reso rstudent=rsc;
run;
```

Graphiques

Dans le sens des aiguilles d'une montre, en partant du coin supérieur gauche:

- résidus ordinaires contre valeurs ajustées (linéarité)
- résidus studentisés contre valeurs ajustées (hétéroscédasticité)
- effet de levier (influence des observations sur l'estimateur)
- diagramme quantile-quantile des résidus (normalité)
- nuage de point de Y_i versus $\hat{Y}_i$ (linéarité)
- distance de Cook (détection des valeurs aberrantes)
- densité et histogramme des résidus ordinaires (normalité)

Conclusion

Dans l'exemple, tous les indicateurs sont verts et nous n'avons aucune raison de douter des postulats de notre modèle.

Cela conforte l'idée que les résultats de notre analyse (conclusions des tests d'hypothèse et intervalles de confiances) sont valides.

Indépendance

- Contexte ✚ Corrélogramme ✚ Code SAS

✚ Les données `trafficaerien` contient des données mensuelles du traffic aérien mondial dans les années 50.

✚ On ajuste un modèle log-linéaire avec mois (catégorielle) et année pour expliquer le nombre de passagers.

✚ La fonction d'autocorrélation (ACF) indique une dépendance résiduelle mensuelle et annuelle.

Postulat de linéarité

Plusieurs graphiques potentiels avec les résidus ordinaires...

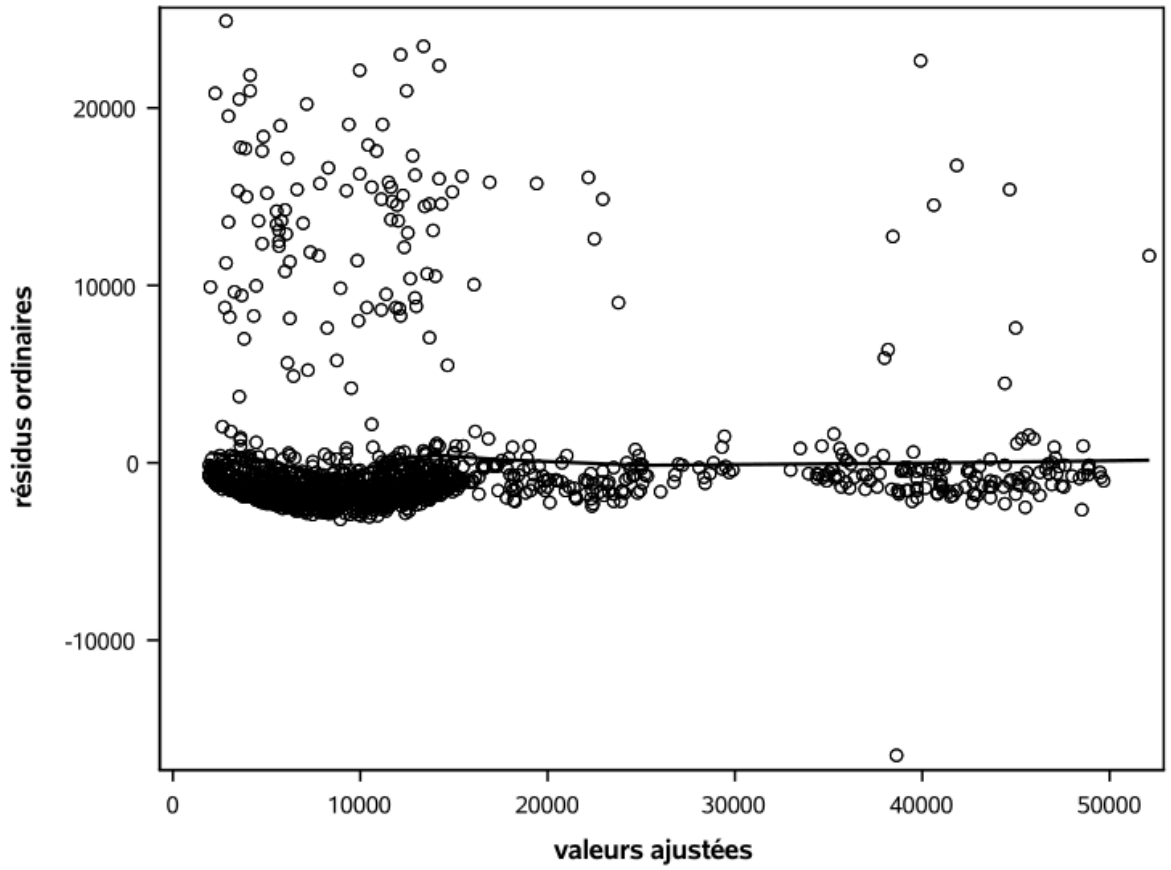
- contre les valeurs ajustées
- contre les variables explicatives
- contre des variables omises (pas incluses dans le modèle pour la moyenne)
- diagrammes de régression partielle

Données assurance

On considère un modèle linéaire avec `age`, `sexe`, `region` et une interaction entre `fumeur/obese` et `imc`.

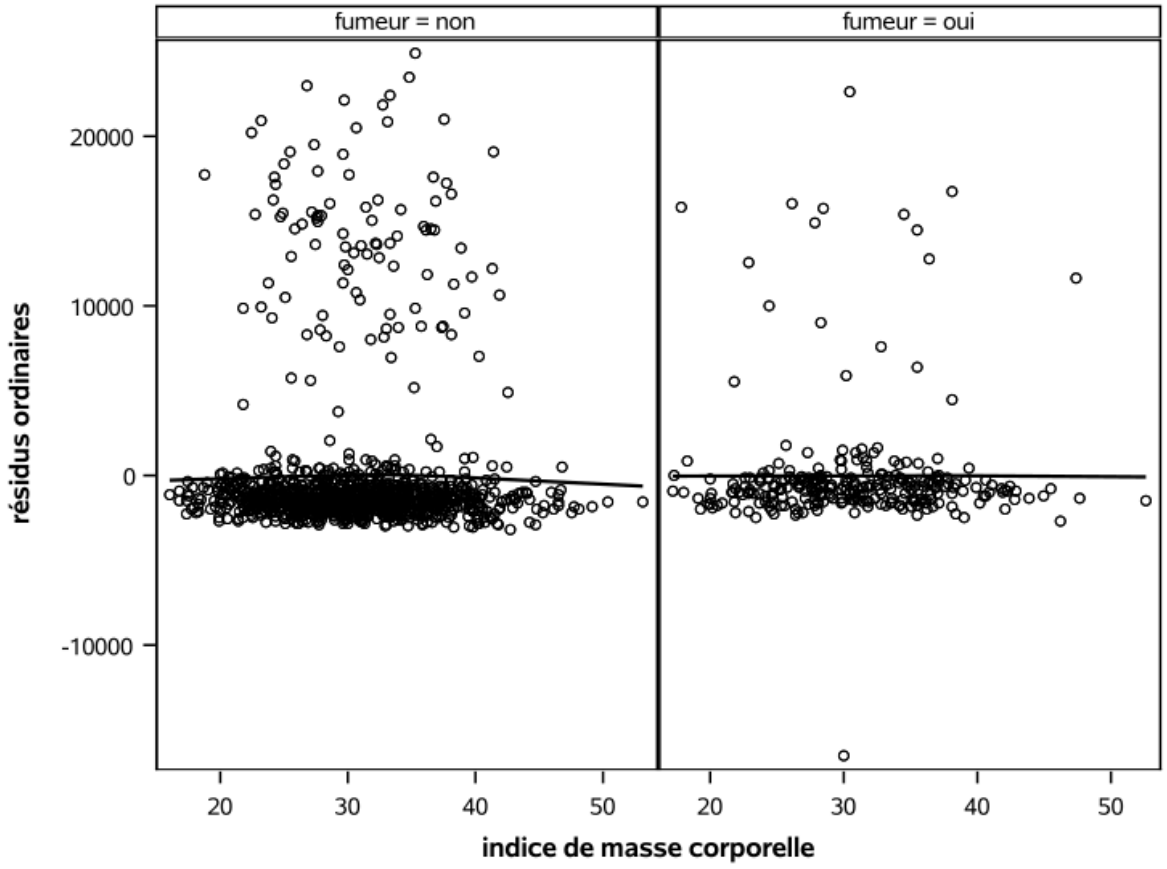
- Les diagrammes nous indiquent que le modèle est inadéquat, mais les apparences sont parfois trompeuses:
  - à cause de valeurs abnormalement élevées (trop grand frais), la moyenne estimée des non-fumeurs est plus grande que la majorité des observations
  - l'ajustement est bon hormis pour ces valeurs inattendues: cela a des répercussions, notamment pour le diagramme quantile-quantile.
  - une transformation logarithmique pourrait potentiellement réduire l'impact de ces valeurs, ou alors une régression robuste qui pondère à la baisse les points trop différents.

- Sortie SAS + Code SAS



Linéarité

- Sortie SAS (1) ✚ Sortie SAS (2) ✚ Code SAS



Homoscédasticité

- Sortie SAS (1) ✚ Sortie SAS (2) ✚ Code SAS

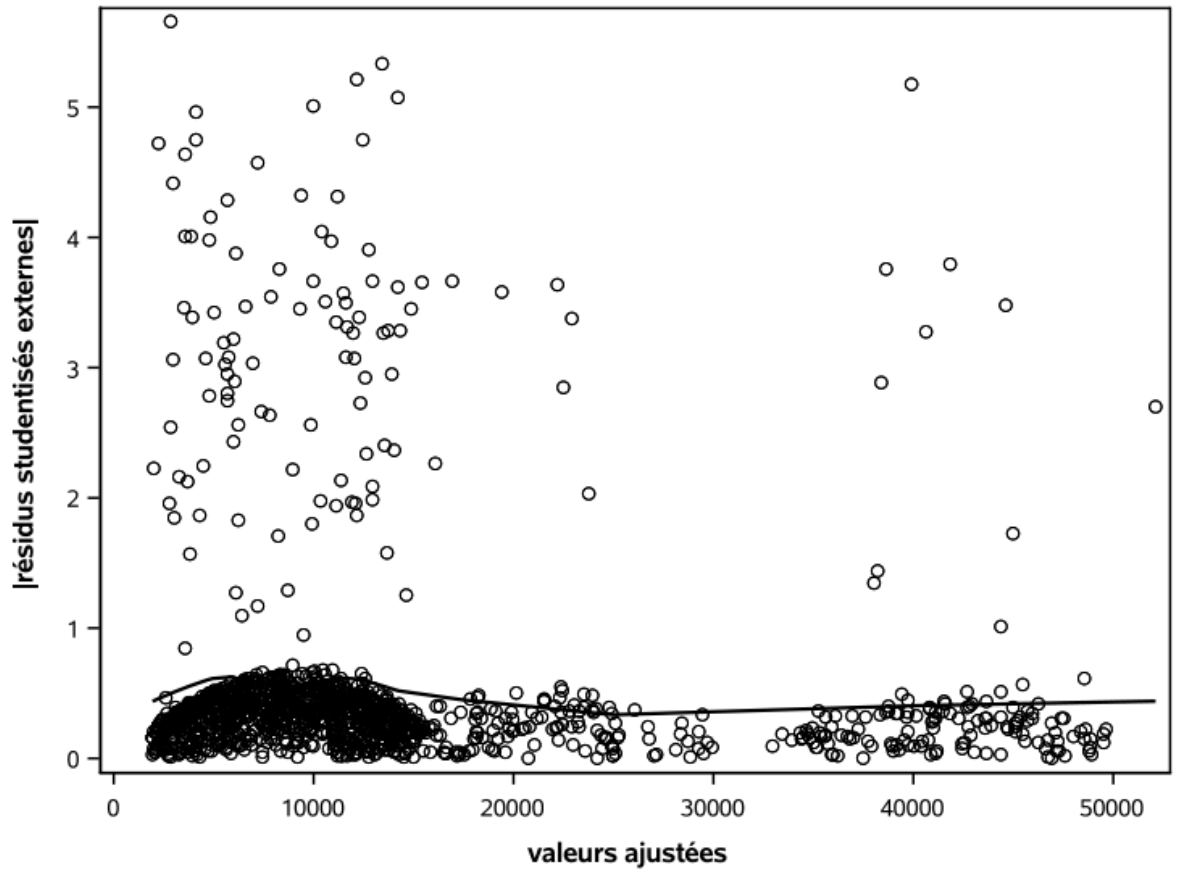


Diagramme quantile-quantile

Pour créer un diagramme quantile-quantile à la mitaine

- trier les données (résidus studentisés externes)
- calculer les positions théoriques $i/(n+1)$, $i = 1, \ldots, n$
- calculer la transformation inverse $F^{-1}\{i/(n+1)\}$, où F^{-1} est la fonction quantile de la loi postulée.
- ajouter des intervalles de confiance ponctuels approximatives (avec des statistiques d'ordre)
  - $U_{(j)} \sim \mathsf{Be}(j, n+1-j)$
  - on calcule les quantiles 0.025 et 0.975 de la loi $\mathsf{Be}(j, n+1-j)$
  - transformer ces variables à l'échelle Student
  - éliminer la traîne

Normalité

- Sortie SAS ✚ Code SAS (1) ✚ Code SAS (2)

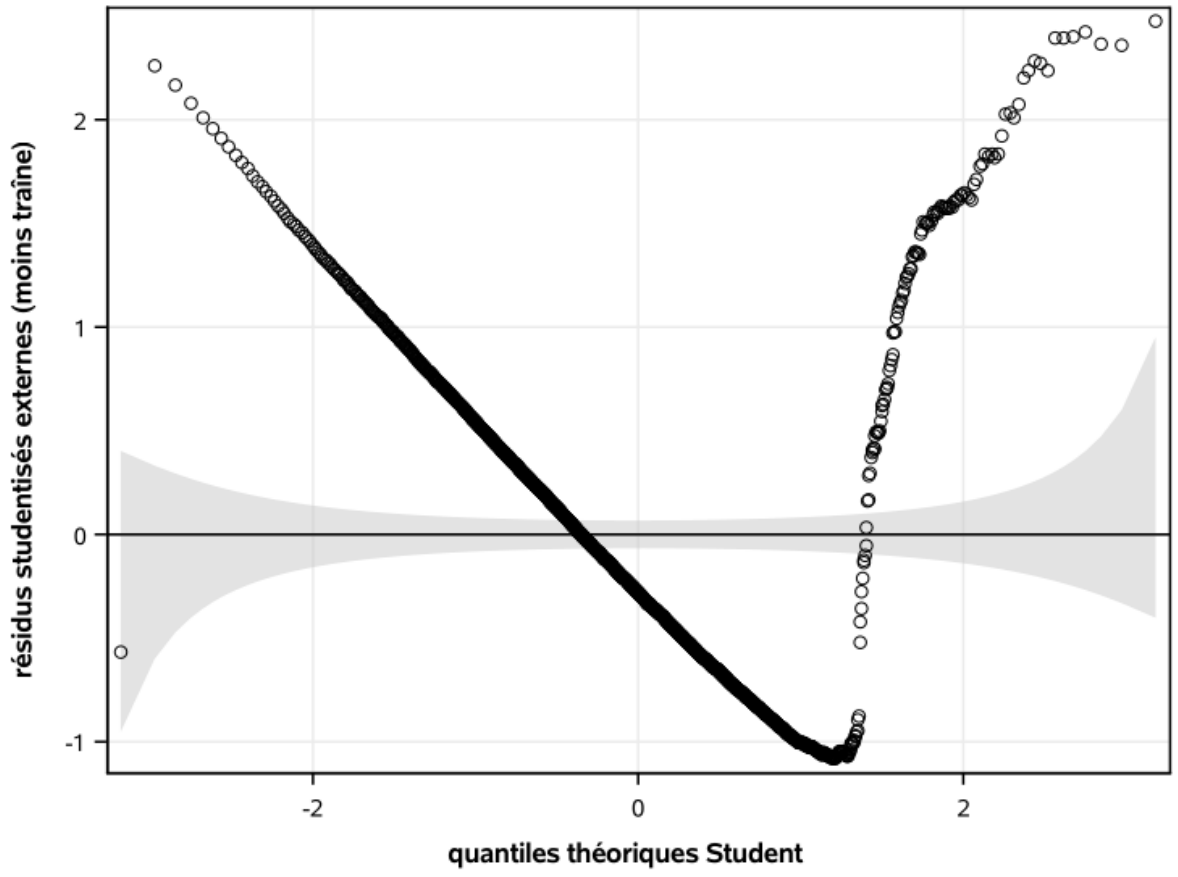


Diagramme quantile-quantile

- La procédure `univariate` permet de faire un diagramme QQ pour quelques lois, incluant la loi normale.
- On pourrait utiliser cette dernière en misant sur l'approximation de la loi Student par une loi normale si les degrés de liberté $n - p - 2$ est grand.

```
/* Histogramme des résidus students et densité */
proc sgplot data=residus;
histogram rsc;
density rsc / type=kernel;
keylegend / position=bottom;
run;

proc univariate data=residus noprint;
qqplot rsc / normal(mu=est sigma=est l=2)
square;
run;
```