Université Laval
Faculté des Sciences et de Génie

Automne 2025
Mathématiques & Statistique

STT-7125 Théorie et applications des méthodes de régression
Devoir 1

Ce devoir compte pour 20% de la note finale.
Vous pouvez travailler individuellement ou en équipe de deux ou trois personnes.
À déposer dans la boîte de dépôt, au plus tard le dimanche 2 novembre 2025 à 23h59.

Ce devoir comporte 3 questions indépendantes.

Question 1 (40 points)

L'objectif de cet exercice est d'analyser les contenus dans le fichier salaires.txt qui se trouve sur le portail du cours.

Ce jeu de données concerne 65 employés de différentes PME dans le domaine des technologies de l'information. La variable réponse est le `salaire` de l'employé. Les variables explicatives sont:

- `diplome` : dernier diplôme obtenu (1: baccalauréat, 2: maîtrise, 3: doctorat),
- `experience` : nombre d'années d'expérience depuis le dernier diplôme obtenu,
- `nombre` : nombre de personnes superviées par l'employé en question.

Analyser ces données en suivant les étapes suivantes:

(a) Présenter des statistiques descriptives avec des graphiques et des conclusions.

(b) Est-ce que ces données présentent une forme de multicollinéarité ?

(c) Sélectionner le modèle le plus adéquat pour ces données selon les procédures *forward*, *backward* et *stepwise*. Le modèle le plus complet considéré sera celui qui comprte toutes les variables explicatives, ainsi que leurs interactions. Est-ce qu'on obtient le même modèle par les 3 méthodes? Si on obtient des modèles différents, on rentiendra celui obtenu par la méthode *stepwise*. Interpréter les coefficients de régression du modèle retenu.

(d) Créer une nouvelle variable `supervision` qui vaut 1 si l'employé supervise des personnes et 0 sinon. Refaire **(c)** avec `supervision` à la place de `nombre`. Est-ce qu'il vaut mieux considérer la variable `nombre` ou la variable `supervision` ?

(e) Présenter graphiquement les résidus et les leviers du modéle final retenu.

(f) Valider les hypothèses du modéle retenu. Est-ce qu'il est nécessaire de transformer la variable réponse ?

(g) Est-ce qu'il y a des observations influentes ou aberrantes ?

Question 2 (30 points)

Considérons le jeu de données downs de la librarie `GLMsData` du logiciel R. Ce jeu de données contient 30 observations de 3 variables :

- Age : âge de la mère.
- Births : nombre de naissances.
- DS : nombre de naissances avec le syndrome de Down (Trisomie 21).

Ainsi, la première ligne de ce jeu de données se lit : il y a eu 13555 naissances dans le groupe de mères âgées, en moyenne, de 17.5 ans et parmi ces naissances, 16 avaient le syndrome de Down. On veut savoir si l'âge de la mère est associé à la naissance avec un le syndrome de Down et si oui, comment?

(a) Présenter des statistiques descriptives du jeu de données avec des graphiques.

(b) Définir clairement la variable réponse. Quelle est la distribution exacte de cette variable et quelle est la distribution la mieux adaptée à ce problème ?

(c) On veut considérer un modèle avec une composante systématique η_i qui s'exprime en fonction de l'âge moyen age_i à l'aide d'un polynôme. Utiliser successivement des statistiques de test Wald et des critères AIC pour déterminer le degré du polynôme le plus adapté à ces données.

(d) Interpréter le modèle obtenu retenu au **(c)**. Représenter graphiquement les valeurs prédites par le modèle en fonction de l'âge. Interpréter la figure obtenue et la comparer aux graphiques produits au **(a)**. Est-ce qu'il y a un âge où le risque d'avoir une naissance avec le syndrome de Down est minimal ? maximal ?

(e) Procéder à la validation du modèle retenu au **(c)**: est-ce qu'il y a une extra-variabilité? est-ce qu'il y a des observations influentes ou aberrantes ?

Question 3 (30 points)

Considérons le jeu de données earinf de la library `GLMsData` du logiciel R. Ce jeu de données contient 287 observations d'infections d'oreilles chez les nageurs. On s'intŕesse à la variable réponse Infec. C'est une variable indicatrice qui vaut 1 si la personne a eu une infection de l'oreille et 0 sinon. Nous avons 4 variables explicatives : Swim (nageur fréquent ou pas), Loc (nage à la mer ou pas), Age et Sex.

Analyser ces données en utilisant un modèle adéquat. Vos analyses doivent comporter les éléments suivants:

- Statistiques descriptives avec des graphiques.
- Définition claire du modèle.
- Sélection des variables: est-ce qu'on doit inclure l'âge comme une variable continue ou discrète ? Si l'âge est considéré comme une variable continue, est-ce qu'on doit inclure des termes carrés et cubiques de l'âge dans le modèle? est-ce qu'on doit inclure des termes d'interaction dans le modèle?
- Choix de la fonction de lien. Est-ce que le choix de la fonction de lien influence les résultats ?
- Interprétation du modèle obtenu.
- Validation du modèle: est-ce que le modèle retenu prédit bien l'infection de l'oreille? est-ce qu'il y a des observations influentes ou aberrantes ?