Université du Québec à Montréal (UQAM)
Faculté des sciences

ACT3035– Examen Intra 2
Laboratoire d'actuariat

Enseignant: Noureddine Meraihi

2021/07/12

Nom: _____

Code permanent: _____

Signature: _____

Cet examen contient 4 pages (incluant la page couverture) et 5 questions sur un total de 25 points.

Instructions

- L'examen commence à 17:30 pour une durée de 120 minutes;
- L'examen compte pour 35% de la note finale;
- Vous serez informés par courriel/Discord lorsque l'examen sera corrigé.

1. (2 points) Vous obtenez les résultats suivants à partir d'un modèle de régression.

i	y_i	$\hat{f}(x_i)$
1	2	1
2	5	3
3	6	9
4	8	3
5	4	6

Calculer l'erreur quadratique moyenne, ou *Mean Squared Error* (MSE) et encerclez la bonne réponse.

A. -35
B. -5
C. 5
D. 43
E. 46

2. (2 points) Lisez attentivement les affirmations suivantes

I Tout modèle de régression linéaire doit être linéaire en variables indépendantes.
II Dans la régression linéaire multiple, une ligne droite est ajustée aux données.
III Dans la régression linéaire simple, on peut ajuster une parabole (polynôme dans une variable indépendante) à des points de données.
IV On peut estimer les paramètres d'un modèle de régression linéaire en utilisant l'estimation du maximum de vraisemblance.

Encerclez les affirmations correctes

A. Affirmations I et II uniquement
B. Affirmations I et III uniquement
C. Affirmations I et IV uniquement
D. Affirmations II et III uniquement
E. Affirmations III et IV uniquement

3. (2 points) Dans une régression linéaire simple, l'approche des moindres carrés choisit β_0 et β_1 pour minimiser:

   A. valeur absolue de la somme des erreurs

   B. la somme de valeur absolue des erreurs

   C. la somme des erreurs au carré

   D. somme des résidus au carré.

4. (4 points) Le tableau (1) contient les informations disponibles pour une petite base de données d'assurance incendie. Les variables sont:

   - **Numéro contrat**: le numéro du contrat;
   - **Incendie**: variable indicatrice d'un incendie (1) ou non (0);
   - **Age**: l'âge de la propriété, en années; et
   - **Type**: le type de construction: bois, béton ou autre.

Numéro contrat	Incendie	Age	Type
1	0	40	Bois
2	0	10	Autre
3	1	20	Bois
4	1	110	Autre
5	0	22	Béton

Tableau 1: Base de données.

   Ajustez un modèle binaire sur ces données.

5. Vous travaillez chez un assureur qui vous demande de modéliser les frais médicaux facturés à vos clients. Pour ce faire, vous avez un jeu de données d'assurance qui est divisée en deux parties;

   - Les données `train.csv` sur lesquels vous ajustez votre modèle préféré
   - Les données `test.csv` sur lesquels vous testez la justesse de votre modèle

   Dans vos deux jeux de données, vous retrouvez exactement les mêmes variables explicatives ainsi que votre variable réponse `charges` que vous voulez estimer. Voici une brève description de ces variables:

   - age : âge du contractant d'assurance, en années
   - sex : sexe du contractant d'assurance, [femme, homme].
   - bmi : indice de masse corporelle, permettant de comprendre le corps, les poids qui sont relativement élevés ou faibles par rapport à la taille, indice objectif du poids corporel (kg / m^2)

- children : nombre d'enfants couverts par l'assurance maladie / nombre de personnes à charge
- smoker : fumeur, [oui, non].
- region : zone de résidence du bénéficiaire aux Etats-Unis, [nord-est, sud-est, sud-ouest, nord-ouest].
- charges : Frais médicaux individuels facturés par l'assurance maladie,

(a) (3 points) Créer un premier modèle linéaire afin d'estimer les valeurs de la variable `charges` en utilisant toutes les variables explicatives de votre jeu de données d'entraînement.

(b) (2 points) Quelle est la valeur du coefficient de détermination linéaire de Pearson de ce modèle

(c) (2 points) Calculer la valeur prédite des frais médicaux individuels du jeu de données test.

(d) (2 points) Calculer la différence entre frais médicaux individuels observés et les frais médicaux individuels prédits $(y_i - \hat{y}_i)$. Où y_i sont les valeurs observées et $\hat{y}_i$ sont les valeurs prédites.

(e) (2 points) Calculer Racine de l'erreur quadratique moyenne:

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{T} (\hat{y}_t - y_t)^2}{T}}$$

(f) (3 points) Calculer le RMSE d'un autre modèle de régression linaire qui exclut cette fois la variable `sex`.

(g) (1 point) En comparant les valeurs de ces deux RMSE (le premier modèle qui inclut la variable `sex` et le deuxième modèle qui exclut cette variable du calcul), que pouvez-vous dire sur la performance du deuxième modèle? Répondre à cette question en 2 lignes de texte.

Fin de l'examen