La régression en général, la régression linéaire en particulier

Présentation pour le cours SOL6210
Analyse quantitative avancée

© Claire Durand, 2024

1

La régression linéaire...

- Vise à prédire la valeur d'une variable dépendante de type métrique (continu), ordinal ou pseudo-métrique (0,1)
- Par un certain nombre de variables qui sont également métriques, ordinales, ou pseudo-métriques (0,1).
- Le processus devrait permettre également d'expliquer comment les variables indépendantes se combinent pour prédire la variable dépendante. On parle d'effets de
  - Médiation: une variable en influence une autre qui à son tour influence la variable dépendante.
  - Modération: l'influence d'une variable sur la variable dépendante varie selon les niveaux d'une troisième variable.

2

La régression

Deux notions importantes

- On regarde la "quantité" de prédiction
  - Soit la proportion de variance expliquée
    - ☞ R^2, ΔR^2
- Mais également la "qualité" de la prédiction
  - Soit les résidus qui nous indiquent
    - Si la prédiction est aussi bonne quelles que soient les valeurs de $x_1, x_2, \ldots x_n$
  - On cherche la meilleure prédiction au moindre coût soit,
    - Le moins de variables possibles, les plus indépendantes possible entre elles
    - D'où la nécessité d'une réduction préalable.

3

Les modes d'entrée des variables

- ☺ Standard: on entre toutes les variables indépendantes en même temps.
- ☺ Hiérarchique: on entre les variables par bloc et on regarde l'impact de l'entrée de nouvelles variables sur celles qui sont déjà dans l'équation.
- ☹ Statistique: pas-à-pas, entrée descendante, ascendante. Ce sont des critères statistiques qui décident des variables qui demeureront dans l'analyse.
- "Setwise": peu disponible - recherche statistique du meilleur ensemble de variables pour bien prédire.

4

Les informations fournies...

À interpréter!

- La proportion de variance expliquée (r^2):
  - ☞ r^2 = SC reg/SC totale où SC est la somme des carrés des écarts à la moyenne.
- Test de signification de r^2:
  - F=CM reg/CM intra où CM est le carré moyen, la Somme de Carrés divisée par le nombre d'unités moins 1.
  - Un test F significatif indique que la proportion de variance expliquée est différente de zéro **dans la population**. Le "p" est la probabilité que le r^2 soit égal à zéro dans la population.

5

Les informations fournies...

À interpréter!

- Les coefficients de régression "b" :
  - $Y=a + b_1x_1 + b_2x_2 + b_3x_3 + ...+b_nx_n +e$
  - Indique par combien la valeur de Y – la variable dépendante – augmente ou diminue **en moyenne** pour une augmentation de 1 de la valeur de x_i – une variable indépendante.
- Le test de signification des coefficients:
  - Test t = b/erreur-type, significatif si t >1,96.
  - Indique la probabilité que le coefficient soit différent de zéro dans la population.

6

Les informations fournies...

À interpréter!

- Les coefficients de régression standardisés (β):
  - Permettent de comparer l'impact de chaque variable sur une même échelle (standardisée).
- Les résidus (e) :
  - Permettent d'évaluer jusqu'à quel point l'équation de régression prédit bien la valeur de Y quelle que soit la valeur des x_i pour chaque cas en examinant:
    - La présence de valeurs aberrantes.
    - Les patrons de résidus.
    - Les graphiques qui permettent d'évaluer la normalité et l'homoscédasticité.

7

La régression standard

- Permet de savoir si l'ensemble de variables choisi apporte une contribution significative à la prédiction:
  - $H_0 : R=0$; $H_1 : R \neq 0$.
- Permet d'identifier la contribution unique de chaque variable, une fois la contribution de toutes les autres variables prise en compte:
  - Corrélation semi-partielle.
- Permet de savoir si chacune des variables ajoute une contribution significative unique à la prédiction:
  - $H_{0i} : b_i=0$; $H_{1i} : b_i \neq 0$.
- On n'utilise pas la régression standard pour la qualité de la prédiction puisque plusieurs variables peuvent avoir une contribution nulle (non significative). Cela ajoute du "bruit", ce qui fait que l'examen des résidus est non pertinent.
- Idéalement: 20 cas par variable.

8

La régression avec entrée hiérarchique

- **On émet des hypothèses et on les vérifie.**
- Permet de vérifier des modèles de relations, ***la présence de médiations*** entre les variables.
- Permet de connaître la contribution ajoutée d'une bloc de variables ou d'une variable d'intérêt:
  - $H_0 : \Delta R^2=0$; $H_1 : \Delta R^2 \neq 0$.
- Idéalement, 20 cas par variable.

9

La régression statistique

- Permet “plus ou moins” de déterminer le meilleur ensemble de prédicteurs en ne gardant que les prédicteurs significatifs, toutes choses égales par ailleurs.
- Le β (*beta*, coefficient standardisé) est le critère utilisé pour choisir les variables.
- Selon les procédures, le processus peut être itératif (méthode pas à pas).
- Idéalement, 40 cas par variable parce que la procédure est très dépendante de l’échantillon.
- Procédure décriée maintenant par tous les statisticiens, avec raison!

10

Note

Quelle que soit la méthode utilisée pour entrer les variables...

- Si les mêmes variables demeurent dans l’équation finale, les coefficients seront les mêmes. Ils ont une valeur unique pour un même ensemble de prédicteurs.
- L’équation de prédiction est une addition...
- C’est le chercheur qui décide quelle est la variable dépendante...
- L’analyse des résidus est essentielle.

11

Le processus

- Faire une régression standard pour examiner la contribution unique de chaque variable et la variance partagée.
- Faire une régression avec entrée hiérarchique en fonction du modèle postulé.
  - **Il faut avoir élaboré un modèle au départ...**
- Faire une régression parcimonieuse où seules les variables significatives sont gardées.
- Examiner les résidus.

12

Indices diagnostiques

- Multicollinéarité
  - Indices diagnostiques:
    - Tolérance doit être le plus élevée possible
    - VIF (facteur d'inflation de la variance) indique un problème si plus grande que 10.
    - Une valeur propre égale à zéro indique un problème de multicollinéarité.
    - Un index conditionnel plus grand que 30 indique un problème.
  - L'instabilité des coefficients indique la présence d'un problème de multicollinéarité.
- Qualité de la prédiction:
  - Résidus doivent être distribués normalement.
  - Être plus petits que 3,16 (valeur standardisée à 99%) quand grande taille de l'échantillon (>1000).