6.1 On s'intéresse à la satisfaction de clients par rapport à un produit. Cette dernière est mesurée à l'aide d'une échelle de Likert, allant de très insatisfait (1) à très satisfait (5). Le jeu de données `multinom.sas7bdat` contient les variables suivantes :

- `y` : score de satisfaction
- `sexe` : sexe de l'individu, soit homme (`0`), soit femme (`1`)
- `educ` : niveau d'éducation le plus élevé complété, soit secondaire (`sec`), soit collégial (`cegep`), soit universitaire (`uni`)
- `revenu` : variable catégorielle indiquant le revenu, soit faible (`1`), moyen (`2`) ou élevé (`3`).
- `age` : âge de l'individu (en années).

Modélisez la satisfaction des clients, `y`, en fonction de l'âge, du niveau d'éducation, du sexe et du niveau de revenu.

(a) Est-ce que le modèle de régression multinomiale ordinale à cote proportionnelles est une simplification adéquate du modèle de régression multinomiale logistique? Si oui, utilisez ce modèle pour la suite. Si non, ajustez le modèle de régression multinomiale logistique avec `1` comme catégorie de référence pour `y`, `1` pour revenu et `sec` pour éducation et utilisez ce dernier pour répondre aux autres questions.

Solution

On peut regarder directement la sortie du test du score pour l'hypothèse des cotes proportionnelles; le modèle ordinal a 10 paramètres, contre 28 pour le modèle multinomial logistique. La valeur-p du test du score est inférieure à 10^{-4}. On peut également faire un test du rapport de vraisemblance en comparant la différence des log-vraisemblance des deux modèles emboîtés. La valeur-p estimée est 0,00088. Le modèle à cote proportionnelle n'est pas adéquat.

(b) Interprétez l'effet des variables éducation et sexe pour la catégorie 2.

Solution

- La cote pour les femmes pour insatisfait par rapport à très insatisfait est 42,4% plus élevée que pour les hommes, toute chose étant égale par ailleurs.
- La cote pour les individus qui ont un diplôme collégial pour insatisfait par rapport à très insatisfait est 2,5% plus basse que pour ceux qui on un diplôme secondaire, toute chose étant égale par ailleurs.
- La cote pour les individus qui ont un diplôme universitaire pour insatisfait par rapport à très insatisfait est 16,2% plus élevée que pour ceux qui on un diplôme secondaire, toute chose étant égale par ailleurs.

(c) Est-ce que le modèle avec une probabilité constante pour chaque item est adéquat lorsque comparé au modèle qui inclut toutes les covariables?

Solution

La statistique pour le test du rapport de vraisemblance que tous les coefficients associés aux covariables sont nuls (24 paramètres) est 55.41, et si le modèle sans covariable était vrai, cette statistique serait approximativement χ^2_{24}. La valeur-p est 0,0003 et on conclut qu'au moins une covariable est utile pour prédire une cote par rapport au modèle avec une probabilité constante.

(d) Est-ce que l'effet de la variable âge est globalement significatif?

Solution

Puisqu'on modélise quatre rapport de cotes à l'aide d'un modèle logistique, on regarde ici le test de Wald (on pourrait manuellement ajuster un modèle sans âge pour faire le test du rapport de vraisemblance). La statistique de Wald vaut 21.69 et la probabilité d'obtenir un tel résultat si $\beta_{\texttt{age}_2} = \beta_{\texttt{age}_3} = \beta_{\texttt{age}_4} = \beta_{\texttt{age}_5} = 0$ est approximativement 0,0002. On conclut que l'âge impacte la probabilité des différents items de satisfaction.

(e) Fournissez un intervalle de confiance à niveau 95% pour l'effet de la variable âge pour chacune des cote par rapport à très insatisfait (`1`). Que concluez-vous sur l'effet de âge pour les réponses `2`, ..., `5` par rapport à `1`?

Compiled 23/03/2021 at 11:09

Solution

Les intervalles de confiance sont [0,975;1,011] pour $\beta_{\texttt{age}_{2|1}}$ (pas significatif), [0,871;0,963] pour $\beta_{\texttt{age}_{3|1}}$ (significatif), [0,941;0,989] pour $\beta_{\texttt{age}_{4|1}}$ (significatif) et [0,988;1,021] pour $\beta_{\texttt{age}_{5|1}}$ (pas significatif)

(f) Écrivez l'équation de la cote ajustée pour satisfait (4) par rapport à très insatisfait (1).

Solution

Pour obtenir l'équation ajustée, on utilise uniquement les coefficients pour y=4 dans le tableau des coefficients.

$$\frac{\mathsf{P}(Y=4 \mid \boldsymbol{X})}{\mathsf{P}(Y=1 \mid \boldsymbol{X})} = \exp(-0{,}114 - 0{,}035\texttt{age} + 0{,}184\texttt{cegep} + 0{,}308\texttt{uni} - 0{,}028\texttt{revenu}_2 + 0{,}018\texttt{revenu}_3 + 0{,}613\texttt{sexe})$$

(g) Prédisez la probabilité qu'un homme de 30 ans qui a un diplôme collégial et qui fait partie de la classe moyenne sélectionne une catégorie donnée. Quelle modalité est la plus susceptible?

Solution

Les probabilités pour (1, 2, 3, 4, 5) sont (0,321 ; 0,222 ; 0,039 ; 0.127 ; 0.292). La modalité la plus susceptible est donc très insatisfait (1).

Compiled 23/03/2021 at 11:09