Pratique de la régression linéaire

Frédéric Lavancier

Objectif : Expliquer au mieux la variable quantitative Y comme une fonction affine des p variables explicatives $X_1, \dots, X_p$, toutes supposées ici quantitatives (voir l'ANCOVA si une est qualitative).

Procédure d'estimation et de validation d'un modèle :

1. Observer le lien entre Y et chacune des variables explicatives : par des nuages de points et en calculant les corrélations linéaires. Vérifier qu'un lien linéaire semble raisonnable.
   À défaut : Transformer la variable explicative pour rendre le lien linéaire avec Y plus convaincant.
2. Ajuster le modèle aux données (fonction `lm` sous `R`).
3. Vérifier la significativité des variables explicatives (tests de Student).
   Modifier le modèle en conséquence (i.e. supprimer les variables non significatives).
4. Vérifier qu'il n'y a pas d'instabilité due à une multicolinéarité : calcul des VIF.
   Si problème : enlever de l'analyse une variable explicative trop corrélée à une autre, ou effectuer une régression robuste (sur composantes principales, PLS, Ridge ou Lasso).
5. Analyser les résidus $\hat{\epsilon}$ en vérifiant :
   - la non corrélation de $\hat{\epsilon}$ et $\hat{Y}$: analyse graphique du nuage de points $(\hat{\epsilon}, \hat{Y})$.
     Si problème : le lien linéaire initial peut être remis en question.
   - l'homoscédasticité : analyse graphique des résidus studentisés, test de Breusch-Pagan (`bptest` dans la libraire `lmtest`).
     Si problème : une transformation de la variable Y peut s'avérer utile ; ou si les différentes variances sont estimables, ajuster le modèle par les Moindres Carrés Généralisés (MCG).
   - si les observations sont échantillonnées dans le temps, la non-corrélation temporelle des résidus : analyse graphique, test de Durbin-Watson (`dwtest`), test de Breush-Godfrey (`bgtest`).
     Si problème : on pourra inclure le passé de Y dans les variables explicatives, ou modéliser la dépendance temporelle des résidus pour utiliser les MCG.
   - si les observations sont peu nombreuses, la normalité des résidus (à l'aide d'un qq-plot).
6. Analyser l'impact de chaque observation : en calculant leur distance de Cook (`cooks.distance`).
   Ecarter éventuellement de l'analyse les individus trop atypiques.

Choisir le meilleur modèle issu des p variables explicatives $X_1, \dots, X_p$:

- Procédure automatique :
  — *exhaustive* : `regsubsets` dans la libraire `leaps`, puis `plot.regsubsets` ;
  — ou *pas à pas* : fonction `step` (ou `regsubsets` en changeant les options).
  Le critère de sélection dans ces algorithmes peut être au choix AIC, BIC, C_p ou R_a^2.
  Lorsque le nombre d'observations est important (i.e. $n \to \infty$), seul BIC choisit le bon modèle, les autres peuvent sélectionner un modèle trop gros (mais pas trop petit).
  <u>Attention</u> : quel que soit le modèle final retenu, on doit vérifier les points 4., 5., 6. ci-dessus.
- Comparaison entre 2 modèles :
  test de Fisher (`anova`) s'ils sont emboités, sinon en comparant leur AIC, BIC, C_p ou R_a^2.

Utiliser le modèle :

- Test de contraintes linéaires sur les paramètres : soit à la main en comparant les SCR du modèle contraint et du modèle complet, soit en utilisant `linearHypothesis` dans la librairie `car`.
- Prévision : `predict.lm`

Références : - "Régression. Théorie et applications" de P.-A. Cornillon et E. Matzner-Løber
- "Le modèle linéaire par l'exemple" de J.-M. Azais et J.-M. Bardet.