Résumé de cours

Cadre théorique du modèle linéaire gaussien : q prédicteurs X_j (non aléatoires ici), prédictand Y

Y vecteur de IR^n : $Y = T\beta + e$ avec $e \sim N_n(0,\sigma^2 I_n)$ $\quad (Y_i = \sum_{j=1}^{q} \beta_j X_{ij} + e_i)$

T matrice $n \times q$ de rang q (q = dimension du modèle)

→ 4 hypothèses portant sur les erreurs :
erreurs gaussiennes, centrées, indépendantes et de variabilité stable (homoscédasticité=variance cste),

ce qui implique pour le prédictand : $Y \sim N_n(T\beta, \sigma^2 I_n)$

Estimateur des moindres carrés ordinaires de β, vecteur gaussien : $\hat{\beta} = ({}^tTT)^{-1}\,{}^tTY$

$E[\hat{\beta}] = ({}^tTT)^{-1}\,{}^tTE[Y] = \beta$ $\qquad V[\hat{\beta}] = ({}^tTT)^{-1}\,{}^tTV[Y]^t(({}^tTT)^{-1}\,{}^tT) = \sigma^2({}^tTT)^{-1}$

Vecteur Y* des valeurs ajustées par le modèle :

$Y^* = T\hat{\beta} = \prod_Q Y$ $\qquad (Y_i^* = \sum_{j=1}^{q} \hat{\beta}_j X_{ij})$; Y* projeté orthogonal de Y sur Q=Im(T)

Vecteur des résidus estimés ε défini par : Y=Y*+ ε

$\varepsilon = Y - T\hat{\beta} = (I_n - \prod_Q)Y = \prod_{Q^\perp} Y$

ε, vecteur gaussien, projeté orthogonal de Y sur $Q^\perp$

$E[\varepsilon] = E[Y - T\hat{\beta}] = T\beta - TE[\hat{\beta}] = 0$

$V[\varepsilon] = V[\prod_{Q^\perp} Y] = \prod_{Q^\perp} V[Y]^t \prod_{Q^\perp} = \sigma^2 \prod_{Q^\perp}$

Estimateur sans biais de la variance de l'erreur e :

$$\hat{\sigma}^2 = \frac{\|\varepsilon\|^2}{n-q} = \frac{\|Y - T\hat{\beta}\|^2}{n-q}$$

car $E[\|\varepsilon\|^2] = E[tr({}^t\varepsilon\varepsilon)] = E[tr(\varepsilon^t\varepsilon)] = tr(E[\varepsilon^t\varepsilon]) = tr(V[\varepsilon]) = \sigma^2 tr(\prod_{Q^\perp}) = \sigma^2(n-q)$

Test de significativité des composantes du vecteur β :

Pour chaque paramètre β_k avec k=1,…,q on teste au niveau α l'hypothèse nulle H_0 : « $\beta_k = 0$ » contre l'hypothèse alternative H_1 : « $\beta_k \neq 0$ » (test bilatéral).

Sous H_0 , la variable $\frac{\hat{\beta}_k}{\hat{\sigma}_k}$ suit la loi de Student à n-q degrés de liberté, avec :

$$\hat{\sigma}_k = \sqrt{\hat{\sigma}^2 ({}^tTT)_{kk}} = \sqrt{\frac{\left\| Y - T\hat{\beta} \right\|^2}{n-q} ({}^tTT)_{kk}}$$

Règle de décision en exploitant la p-value du test :

Si p_value < α : rejet de H_0 au niveau α (on considère β_k non nul → prédicteur associé à conserver)

Si p_value > α : les données ne permettent pas le rejet de H_0 au niveau α (on considère β_k nul et le prédicteur est alors sans intérêt).

Le logiciel R fournit un tableau résumant le modèle estimé, le nombre de lignes correspondant à la dimension du modèle (dimension du vecteur β). Chaque ligne rassemble les résultats relatifs à l'estimation de chaque composante de β, le nom du prédicteur associé à chaque paramètre est mentionné en début de ligne (*intercept* pour le terme constant β_0).

La colonne *Estimate* contient les estimations des q composantes β_k de β.
La colonne *Std.Error* contient les écarts-types estimés des estimateurs des paramètres β_k.
La colonne *t.value* contient les valeurs prises par la statistique de test sur votre échantillon.
La colonne *Pr(>|t|)* contient les p-values du test de Student par défaut bilatéral réalisé.

On a donc : **t.value=Estimate/Std.Error**

Exploitation de prédicteurs qualitatifs :

L'introduction dans un modèle à terme constant de prédicteurs qualitatifs (variables à modalités) pose un problème technique, la matrice T n'étant alors plus de plein rang. Les vecteurs colonnes de T étant liés, ils ne constituent plus une base du sous-espace Q.
Le vecteur Y*, projeté orthogonal de Y sur Q existe et est unique mais son expression en fonction des vecteurs colonnes de T ne l'est plus (on peut en effet désormais trouver plusieurs jeux de paramètres β_k satisfaisant au problème). On dit que le modèle n'est plus identifiable.
Imposer une contrainte sur les paramètres de chaque prédicteur qualitatif va permettre de fixer un jeu de coefficients et d'identifier au final le modèle. **Le choix de la contrainte n'a aucun impact sur le modèle et ses prévisions (Y* est unique) mais il en aura un sur l'interprétation des paramètres finalement estimés**.

Pour simplifier, un seul prédicteur qualitatif Z à 2 modalités A et B est considéré dans la suite (les explications sont généralisables aux prédicteurs à plus de 2 modalités et à l'exploitation de plusieurs prédicteurs qualitatifs dans le même modèle).

L'impact du prédicteur Z sur le prédictand est modélisé en introduisant dans le modèle un paramètre par modalité, représentant l'effet moyen de la modalité ; soit ici α_A et α_B ces 2 paramètres.

$$Y_i = \beta_0 + \sum_{j=1}^{q} \beta_j X_{ij} + \alpha_i + e_i$$

Avec $\alpha_i = \alpha_A$ ou α_B selon l'occurrence observée de la modalité.

a)

Contrainte notée « $\alpha_A=0$ » dans la littérature : la modalité A de la variable Z est alors prise comme référence, **c'est la contrainte imposée par défaut par le logiciel R**.

Le modèle est reconsidéré ainsi : $Y_i = (\beta_0 + \alpha_A) + \sum_{j=1}^{q} \beta_j X_{ij} + (\alpha_i - \alpha_A) + e_i$

$$Y_i = \beta_0' + \sum_{j=1}^{q} \beta_j X_{ij} + \alpha_i' + e_i$$

Avec $\beta_0' = \beta_0 + \alpha_A$ et $\alpha_i' = \alpha_i - \alpha_A$, on a alors :

$\alpha_A' = \alpha_A - \alpha_A = 0$ et $\alpha_B' = \alpha_B - \alpha_A$ = effet différentiel de la modalité B par rapport à la modalité A.

Il n'y a plus ainsi qu'un paramètre à estimer α_B' concernant la variable Z, la matrice T perdant une dimension est de nouveau de plein rang et les calculs sont réalisables. Mais les paramètres estimés ne sont pas les paramètres initiaux du modèle, il faut donc faire très attention lors des interprétations.
Avec cette contrainte, les paramètres estimés sont les effets différentiels de chaque modalité par rapport à celle qui a été choisie comme référence.

b)

On peut utiliser d'autres contraintes, comme par exemple la contrainte notée « $\alpha_A + \alpha_B = 0$ », la référence prise alors étant alors la demi-somme des impacts moyens des 2 modalités :

$$Y_i = (\beta_0 + \frac{\alpha_A + \alpha_B}{2}) + \sum_{j=1}^{q} \beta_j X_{ij} + (\alpha_i - \frac{\alpha_A + \alpha_B}{2}) + e_i$$

$$Y_i = \beta_0' + \sum_{j=1}^{q} \beta_j X_{ij} + \alpha_i' + e_i$$

Avec $\beta_0' = \beta_0 + (\alpha_A + \alpha_B)/2$ et $\alpha_i' = \alpha_i - (\alpha_A + \alpha_B)/2$, on a alors :

$\alpha_A' = (\alpha_A - \alpha_B)/2$ et $\alpha_B' = (\alpha_B - \alpha_A)/2 = -\alpha_A'$

Là encore, un seul paramètre est à estimer (demi-effet différentiel), le second étant son opposé. La matrice T perd une dimension et les calculs peuvent être menés.