Chapitre 4

Analyse de variance

4.1 Introduction

Les techniques dites d'analyse de variance sont des outils entrant dans le cadre général du modèle linéaire, où une *variable quantitative* est expliquée par une ou plusieurs *variables qualitatives*. Ici une variable qualitative va modéliser par exemple l'appartenance à un groupe, par exemple $T = 0$ signifie qu'un patient est sain, $T = 1$ signifie qu'il a une pathologie donnée, et $T = 2$ signifie qu'il a une autre pathologie.

L'objectif essentiel est alors de comparer les moyennes empiriques de la variable quantitative observées pour différentes catégories d'unités statistiques. Ces catégories sont définies par l'observation des variables qualitatives ou facteurs prenant différentes modalités ou encore de variables quantitatives découpées en classes ou niveaux.

Il s'agit donc de savoir si un facteur ou une combinaison de facteurs (interaction) a un effet sur la variable quantitative en vue, par exemple, de déterminer des conditions optimales de production ou de fabrication, une dose optimale de médicaments. Ces techniques apparaissent aussi comme des cas particuliers de la régression linéaire multiple en associant à chaque modalité une variable indicatrice (dummy variable) et en cherchant à expliquer une variable quantitative par ces variables indicatrices. L'appellation *analyse de variance* (ANOVA pour ANalysis Of Variance) vient de ce que les tests statistiques sont bâtis sur des comparaisons de sommes de carrés de variations.

Notons que l'analyse de variance avancée conduit à l'étude de plans d'experience. Ces derniers ne seront pas abordé dans ce cour mais sont un champs important de l'analyse statistique.

4.2 Modèle ANOVA à un facteur

Cette situation est un cas particulier d'étude de relations entre deux variables statistiques : une quantitative Y admettant une densité et une qualitative T ou facteur qui engendre une partition ou classification de l'échantillon en J groupes (ou cellules, classes, ...) indicées par j. L'objectif est de comparer les distributions de Y pour chacune des classes en particulier les valeurs des moyennes et variances.

Notons qu'avant de lancer une analyse avec les outils présentés ci-dessous, il est recommandé de réaliser un graphique constitué de boites à moustaches parallèles, une pour chaque modalité. Cette représentation donne une première appréciation de la comparaison des distributions (moyenne, variance) internes à chaque groupe.

4.2.1 Modèle

Présentation

On dispose de n observations comme dans les sections précédentes mais on ne considère que les y_i. Chaque niveau j de T avec $j = 1, \ldots, J$ et $J < n$ correspond à un groupe d'observations : Pour chaque j, on observe n_j valeurs $y_{1j}, \ldots, y_{1n_j}$ de la variable Y où $n = \sum_{j=1}^{J} n_j$. On suppose qu'à l'intérieur de chaque groupe, les observations sont indépendantes équidistribuées de moyenne μ_j et de variance homogène $\sigma_j^2 = \sigma^2$. Ceci s'écrit :

$$y_{ij} = \mu_j + \varepsilon_{ij}$$

où les ε_{ij} sont i.i.d. suivent une loi centrée de variance σ^2 qui sera supposée $\mathcal{N}(0, \sigma^2)$ pour la construction des tests. Les espérances μ_j ainsi que le paramètre de nuisance σ^2 sont les paramètres inconnus à estimer.

Estimation des paramètres

On note respectivement :

$$\bar{y}_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$$

$$s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{.j})^2$$

$$\bar{y}_{..} = \frac{1}{n} \sum_{j=1}^{J} \sum_{i=1}^{n_j} y_{ij}$$

les moyennes et variances empiriques de chaque groupe, et la moyenne générale de l'échantillon. Alors :

- Les paramètres μ_j sont estimés sans biais par les moyennes $\bar{y}_{.j}$.
- Comme $y_{ij} = \bar{y}_{.j} + (y_{ij} - \bar{y}_{.j})$, l'estimation des erreurs e_{ij} dans chaque groupe j est naturellement :

$$e_{ij} = (y_{ij} - \bar{y}_{.j}) .$$

- Les valeurs prédites dans chaque groupe j sont $\widehat{y}_{ij} = \bar{y}_{.j}$.

- Sous l'hypothèse d'homogénéité des variances, la meilleure estimation sans biais de σ^2 s'écrit donc comme une moyenne pondérée des variances

empiriques de chaque groupe :

$$s^2 = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_{.j})^2}{n - J}$$
$$= \frac{1}{n-J}((n_1 - 1)s_1^2 + \ldots + (n_J - 1)s_J^2)\,.$$

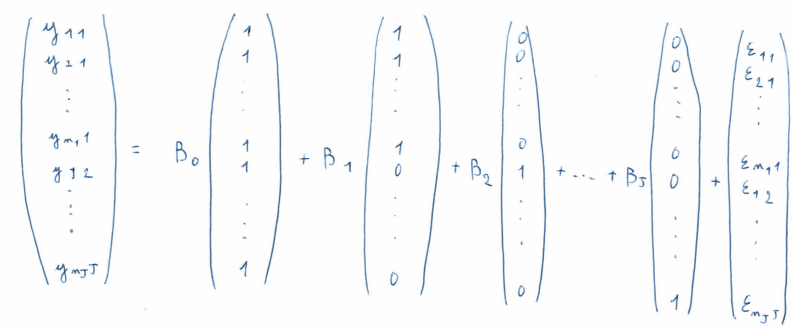
Ecriture sous forme vectorielle

On note :
- $\mathbf{y}$ le vecteur colonne de taille n des observations y_{ij} pour $i = 1, \ldots, n_j$ et $j = 1, \ldots, J$.
- $\mathbf{u}$ le vecteur colonne de taille n des erreurs ε_{ij} pour $i = 1, \ldots, n_j$ et $j = 1, \ldots, J$.
- $\mathbf{1}_j$ le vecteur colonne de taille n des indicatrices du fait d'être dans la classe j. Son $i$ème élément vaut 1 si la $i$ème observation est dans la classe j, et 0 sinon.

Comme dans le cas de la régression linéaire multiple, le modèle consiste à écrire que l'espérance de la variable Y appartient au sous-espace linéaire engendré par les variables explicatives, ici les variables indicatrices :

$$\mathbf{y} = \beta_0\mathbf{1} + \beta_1\mathbf{1}_1 + \ldots + \beta_J\mathbf{1}_J + \mathbf{u}$$

où $\mathbf{1}$ vaut 1 partout. Cette équation est détaillée ci-dessous :



La matrice $\mathbf{X}$, équivalente à celle du modèle linéaire multiple, peut être construite en agrégeant horizontalement les $\mathbf{1}_j$, $j = 1, \ldots, J$. La matrice $\mathbf{X}'\mathbf{X}$ n'est cependant pas inversible en général et le modèle admet une infinité de solutions. Nous disons alors que les paramètres β_j ne sont pas estimables ou identifiables. En revanche, certaines combinaisons linéaires de ces paramètres sont estimables et appelées contrastes. On estimera alors les paramètres comme dans la sous-section précédente.

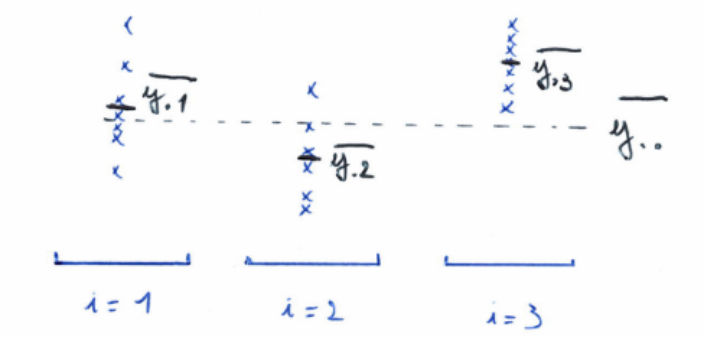
4.3 Test sur la moyenne

Test standard

On désigne les différentes sommes des carrés des variations par :

$$\begin{aligned}
\text{SST} &= \sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_{..})^2 = \sum_{j=1}^{J}\sum_{i=1}^{n_j} y_{ij}^2 - n\bar{y}_{..}^2, \\
\text{SSW} &= \sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_{.j})^2 = \sum_{j=1}^{J}\sum_{i=1}^{n_j} y_{ij}^2 - \sum_{j=1}^{J} n_j\bar{y}_{.j}^2, \\
\text{SSB} &= \sum_{j=1}^{J} n_j(\bar{y}_{.j}-\bar{y}_{..})^2 = \sum_{j=1}^{J} n_j\bar{y}_{.j}^2 - n\bar{y}_{..}^2,
\end{aligned}$$

où T signifie totale, W (within) intra ou résiduelle et B (between) inter ou expliquée par la partition. La figure ci-dessous permet d'illustrer ces notations :



On considère alors l'hypothèse à tester

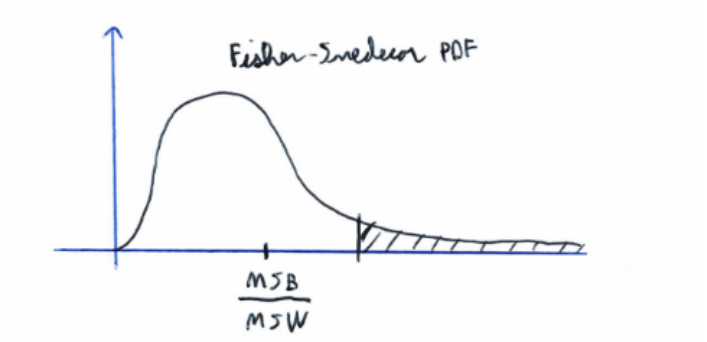
$$H_0 : \mu_1 = \ldots = \mu_J \,,$$

qui revient à dire que la moyenne est indépendante du groupe ou encore que le facteur n'a pas d'effet, contre l'hypothèse

$$H_1 : \exists(j,k) \text{ tel que } \mu_j \neq \mu_k \,,$$

qui revient à reconnaître un effet ou une influence du facteur sur la variable $\mathbf{Y}$.

L'étude de cette hypothèse revient à comparer par un test de Fisher d'égalité des variances un modèle complet (les moyennes sont différentes) avec un modèle réduit supposant la nullité des paramètres β_j. La statistique de test est alors MSB/MSW, où $MSB = SSB/(J-1)$ et $MSW = SSW/(n-J)$. Cette statistique suit une distribution de Fischer à $(J-1)$ et $(n-J)$ degrés de libertés (voir appendice A).



Notons qu'une hypothèse a été faite ici sur l'homogénéité de la variance dans toutes les classes. Le test de Barlett et celui de Levene permettent de tester cette hypothèse.

Tests non-paramétriques

On rappelle que l'on a supposé dans cette section que les ε_{ij} suivent une loi $\mathcal{N}(0, \sigma^2)$. Lorsque l'hypothèse de normalité n'est pas satisfaite et que la taille de l'échantillon est trop petite, et ne permet ainsi pas de supposer des propriétés asymptotiques, une procédure non-paramétrique peut encore être mise en œuvre. Elle est une alternatives intéressante au test de Fisher pour tester l'égalité des moyennes.

La procédure la plus utilisée est la construction du test de Kruskal-Wallis basée sur les rangs. Toutes les observations sont ordonnées selon les valeurs y_{ij} qui sont remplacées par leur rang r_{ij}, les ex-equo sont remplacés par leur rang moyen. On montre que la statistique de ce test, construite sur la somme des rangs à l'intérieur de chaque groupe, suit asymptotiquement une loi du χ^2 à $(J-1)$ degrés de liberté.

4.4 Recherche de moyennes significativement différentes

Si l'hypothèse nulle est rejetée, il est légitime de se demander quel sont les groupes qui possèdent des moyennes significativement différentes. De nombreux tests et procédures ont été proposés dans la littérature pour répondre à cette question.

Procédure naïve

Une procédure naïve consiste à exprimer, pour chaque paire j et l de groupes, un intervalle de confiance au niveau $100(1-\alpha)\%$ de la différence $(\mu_j - \mu_l)$, avec α typiquement égal à 0.05 :

$$(\bar{\mu}_{.j} - \bar{\mu}_{.l}) \pm t_{n-J}(\alpha/2) s \left(\frac{1}{n_j} + \frac{1}{n_l}\right)^{1/2}.$$

où $t_{n-J}(.)$ est donné appendice A. Si 0 est inclu dans cet intervale pour un couple de groupes (j, l), leur moyennes ne sont pas jugées significativement différentes au niveau α.

L'orthogonalité des facteurs rendent les tests indépendants mais elle ne peut être systématisée si J est grand. Dans ce cas, il y a en effet un total de $J(J-1)/2$ comparaisons à tester ce qui peut être long. De manière plus critique encore, on peut s'attendre à ce que sur le simple fait du hasard $\alpha \times J(J-1)/2$ paires de groupes soient jugés de moyennes significativement différentes même si le test global accepte l'égalité des moyennes. Par définition même du niveau d'incertitude, le test se trompe en effet environ $\alpha \times 100\%$ des fois.

Procédure de Bonferroni

D'autres procédures visent à corriger cette démarche afin de contrôler globalement le niveau des comparaisons. La plus standard est la procédure de Bonferroni qui propose des intervalles plus conservatifs (plus grands) en ajustant le niveau $\alpha' < \alpha$ définissant les valeurs critiques $t_{n-J}(\alpha'/2)$ dans la loi de Student avec :

$$\alpha' = \frac{2\alpha}{J(J-1)}$$

Cette procedure est tout de même plus conservative que la procédure naïve et a ainsi la propriété d'augmenter le nombre de faux positifs *i.e.* de moyennes significativement différentes non détectées. D'autres méthodes comme celle de Scheffe (encore plus conservative) ou bien d'autres basées sur des intervalles studentisés avec des valeurs critiques spécifiques existent. La recherche est encore active dans ce domaine.

4.5 Extension à deux facteurs

Introduction

La considération de deux (ou plus) facteurs explicatifs, dans un modèle d'analyse de variance, engendre plusieurs complications. Nous aborderons ici celles qui concernent l'interaction entre variables explicatives. Cette section décrit alors le cas de deux facteurs explicatifs croisés c'est-à-dire dont les niveaux d'un facteur ne sont pas conditionnés par ceux de l'autre. On note :

- Les niveaux du 1er facteur sont notés par l'indice j variant de 1 à J.
- Les niveaux du 2eme facteur sont notés par l'indice k variant de 1 à K.
- Pour chaque combinaison (j,k), on dispose de n_{jk} observations y_{ijk}, $i = 1,\ldots,n_{jk}$.

Un plan d'experience peut être soit *équilibré* (ou équirépété) soit *déséquilibré*. Un plan sera dit équilibré si pour chaque combinaison (j,k), on dispose du même nombre d'observations : $n_{jk} = c$, $\forall (j,k)$. Ce cas introduit des simplifications importantes dans l'estimations des paramètres ainsi que dans la décomposition des variances. On se placera dans le cas équilibré dans la suite de cette sous-section.

Modèle complet

On écrit un modèle de variance à un facteur présentant $J \times K$ niveaux (j,k) :

$$y_{ijk} = \mu_{jk} + \varepsilon_{ijk}$$

avec $j = 1, \ldots, J$, $k = 1, \ldots, K$ et $i = 1, \ldots, n_{jk}$. On se place dans le cas équilibré avec $n_{jk} = c$, $\forall (j,k)$. En supposant que les termes d'erreur ε_{ijk} sont mutuellement indépendants et de même loi. Chacun des paramètres μ_{jk} est estimé sans biais par la moyenne

$$\bar{y}_{.jk} = \frac{1}{c} \sum_{i=1}^{c} y_{ijk} \, .$$

On défini de même les moyennes suivantes :

$$\begin{aligned}
\bar{y}_{.j.} &= \frac{1}{K} \sum_{i=1}^{K} \bar{y}_{.jk} \\
\bar{y}_{..k} &= \frac{1}{J} \sum_{i=1}^{J} \bar{y}_{.jk} \\
\bar{y}_{...} &= \frac{1}{J} \sum_{i=1}^{J} \bar{y}_{.j.} = \frac{1}{K} \sum_{i=1}^{K} \bar{y}_{..k}
\end{aligned}$$

qui n'ont de sens que dans le cas équilibré. La même convention du point en indice est également utilisée pour exprimer les moyennes des paramètres μ_{ijk}.

On estime alors différents termes avec :

— L'effet général $\mu_{..}$ est estimée avec $\bar{y}_{...}$.

— L'effet différentiel du niveau j du 1er facteur $\alpha_j = \mu_{j.} - \mu_{..}$ est estimé avec $\bar{y}_{.j.} - \bar{y}_{...}$.

— L'effet différentiel du niveau k du 2eme facteur $\beta_k = \mu_{.k} - \mu_{..}$ est estimé avec $\bar{y}_{..k} - \bar{y}_{...}$.

— L'effet de l'interaction des niveaux j et k $\gamma_{jk} = \mu_{jk} - \mu_{j.} - \mu_{.k} + \mu_{..}$ est estimé avec $\bar{y}_{.jk} - \bar{y}_{.j.} - \bar{y}_{..k} + \bar{y}_{...}$.

Un modèle d'analyse de variance à deux facteurs s'écrit alors :

$$y_{ijk} = \mu_{..} + \alpha_j + \beta_k + \gamma_{jk} + \epsilon_{ijk}$$

avec les contraintes :

$$\begin{aligned}
\sum_{j=1}^{J} \alpha_j &= 0 \\
\sum_{k=1}^{K} \beta_k &= 0 \\
\sum_{j=1}^{J} \gamma_{jk} &= 0 \, , \, \forall k \\
\sum_{k=1}^{K} \gamma_{jk} &= 0 \, , \, \forall j
\end{aligned}$$

qui découlent de la définition des effets et assurent l'unicité de la solution.

Modèles de régression

Comme dans le cas du modèle à un facteur, l'analyse d'un plan à deux facteurs se ramène à l'estimation et l'étude de modèles de régression sur variables indicatrices. En plus de celles des niveaux des deux facteurs $\mathbf{1}_1^1, \ldots, \mathbf{1}_J^1$, et $\mathbf{1}_1^2, \ldots, \mathbf{1}_K^2$, la prise en compte de l'interaction nécessite de considérer les indicatrices de chaque cellule ou traitement obtenues par produit des indicatrices des niveaux associés $\mathbf{1}_{jk}^{1\times 2} = \mathbf{1}_j^1 \times \mathbf{1}_k^2$. On peut alors écrire un modèle de régression comme dans le cas a un facteur (sous-section 4.2.1) en considérant toutes les combinaisons de (j,k) possibles. Il conduit aussi dans le cas général à inverser une matrice $\mathbf{X}'\mathbf{X}$ non inversible. Il est alors usuel de supprimer des indicatrices.

Stratégie de test

On considère les sommes de carrés spécifiques au cas équirépété :

$$\begin{aligned}
\text{SST} &= \sum_{i=1}^{c}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk}-\bar{y}_{...})^2 &&= \sum_{i=1}^{c}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk}^2 - cJK\bar{y}_{...}^2,\\
\text{SS1} &= cK\sum_{j=1}^{J}(\bar{y}_{.j.}-\bar{y}_{...})^2 &&= cK\sum_{j=1}^{J} y_{.j.}^2 - cJK\bar{y}_{...}^2,\\
\text{SS2} &= cJ\sum_{k=1}^{K}(\bar{y}_{..k}-\bar{y}_{...})^2 &&= cJ\sum_{k=1}^{K} y_{..k}^2 - cJK\bar{y}_{...}^2,\\
\text{SSI} &= c\sum_{j=1}^{J}\sum_{k=1}^{K}(\bar{y}_{.jk}-\bar{y}_{.j.}-\bar{y}_{..k}+\bar{y}_{...})^2 &&= c\sum_{j=1}^{J}\sum_{k=1}^{K}\bar{y}_{.jk}^2 - cK\sum_{j=1}^{J} y_{.j.}^2 -\\
& && \quad -cJ\sum_{k=1}^{K} y_{..k}^2 + cJK\bar{y}_{...}^2,\\
\text{SSE} &= \sum_{i=1}^{c}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk}-\bar{y}_{.jk})^2 &&= \sum_{i=1}^{c}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk}^2 - c\sum_{j=1}^{J}\sum_{k=1}^{K}\bar{y}_{.jk}^2.
\end{aligned}$$

Dans le cas équirépété, il est facile de montrer que tous les doubles produits des décompositions s'annulent (théorème de Pythagore) et que

$$SST = SS1 + SS2 + SSI + SSE\,.$$

On parle alors de plans orthogonaux et les trois hypothèses suivantes peuvent être considérées de façon indépendante :

- $H_{03} : \gamma_{11} = \ldots = \gamma_{JK}$, *i.e.* pas d'effet d'interaction. Hypothèse testée par un test de Fisher avec la statistique MSI/MSE où $MSI = SSI/((J-1)(K-1))$ et $MSE = SSE/(JK(c-1))$.
- $H_{02} : \beta_1 = \ldots = \beta_K$ et H_{03} *i.e.* pas d'effet du 2ème facteur. Hypothèse testée par un test de Fisher avec la statistique $MS2/MSE$ où $MS2 = SS2/(K-1)$.
- $H_{01} : \alpha_1 = \ldots = \alpha_J$ et H_{03} *i.e.* pas d'effet du 1er facteur. Hypothèse testée par un test de Fisher avec la statistique $MS1/MSE$ où $MS1 = SS1/(J-1)$.

En pratique, des questions supplémentaires se posent pour des plans désiquilibrés ou incomplets. Il existe en fait toute une littérature liée au plans d'experience. De même l'analyse de covariance peut être effectuée pour analyser l'interaction entre les variables mais déborde du cadre de ce cours.

4.6 Analyse de covariance

L'analyse de covariance se situe dans un contexte où une variable quantitative Y est expliquée par plusieurs variables à la fois quantitatives et qualitatives. Le principe général est alors d'estimer des modèles *intra-groupes* et de tester des effets différentiels *inter-groupes* des paramètres des régressions.

Nous nous intéressons ici au cas le plus simple où seulement une variable X parmi les explicatives est quantitative. Nous sommes alors amenés à tester l'hétérogénéité des constantes et celle des pentes (interaction) entre différents modèles de régression linéaire.

Modèle

On considère une variable quantitative Y expliquée par une variable qualitative T à J niveaux et une variable quantitative (appelée encore covariable) X. Pour chaque niveau j de T , on observe n_j valeurs $x_{1j}, \ldots, x_{n_j j}$ de X et n_j valeurs $y_{1j}, \ldots, y_{n_j j}$ de Y. La taille de l'échantillon est alors $n = \sum_{j=1}^{J} n_j$.

Pour chaque $j = 1, \ldots, J$ et $i = 1, \ldots, n_j$, on suppose alors que :

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij}$$

où mes ε_{ij} sont i.i.d. de loi centrée et de variance σ^2. Pour la construction de tests, on suppose que $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ La résolution simultanée des J modèles de régression est alors obtenue en considérant globalement le modèle :

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$$

où les matrices et vecteurs sont construits comme ci-dessous :

$$\begin{pmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{m_1 1} \\ y_{12} \\ \vdots \\ y_{m_J J} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & 0 & 0 & \cdots & 0 & 0 \\ 1 & x_{21} & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & x_{m_1 1} & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & x_{12} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & x_{m_J J} \end{pmatrix} \begin{pmatrix} \beta_{01} \\ \beta_{11} \\ \beta_{02} \\ \beta_{12} \\ \vdots \\ \beta_{0J} \\ \beta_{1J} \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{m_1 1} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{m_J J} \end{pmatrix}$$

On dénote alors $\mathbf{1_j}$ et $\mathbf{x}.\mathbf{1_j}$ les colonnes $2(j-1)+1$ et $2(j-1)+2$ de la matrice $\mathbf{X}$ qui contiennent respectivement des 1 et les valeurs de x_{ij} pour le $j$ème groupe de T. L'estimation de ce modèle global conduit à estimer les modèles de régression par bloc, dans chacune des cellules.

Tests

Afin d'obtenir directement les bonnes hypothèses dans les tests, il est standard de reparamétriser les β_{ij} de manière à ce qu'ils expriment une différence avec un autre paramètre. Typiquement des β_{0J} et β_{1J} sont estimés sur tout $\mathbf{x}$ et les effets différentiels consistent à estimer des $\beta_{ij}^{'} = (\beta_{ij} - \beta_{iJ})$, $i \in \{0, 1\}$ et $j = 1, \ldots, J-1$. Le modèle complet est alors :

$$\mathbf{y} = \beta_{0J}\mathbf{1} + \beta_{1J}\mathbf{x} + \sum_{j=1}^{J-1} \left(\beta_{0j}^{'}\mathbf{1_j} + \beta_{1j}^{'}\mathbf{x}.\mathbf{1_j}\right) + \mathbf{u}$$

Différentes hypothèses peuvent alors être testées en comparant ce modèle à un des modèles réduits suivants :

$$(i)\, \mathbf{y} = \beta_{0J}\mathbf{1} + \beta_{1J}\mathbf{x} + \sum_{j=1}^{J-1} \left(\beta_{0j}^{'}\mathbf{1_j}\right) + \mathbf{u}$$

$$(ii)\, \mathbf{y} = \beta_{0J}\mathbf{1} + \sum_{j=1}^{J-1} \left(\beta_{0j}^{'}\mathbf{1_j} + \beta_{1j}^{'}\mathbf{x}.\mathbf{1_j}\right) + \mathbf{u}$$

$$(iii)\, \mathbf{y} = \beta_{0J}\mathbf{1} + \beta_{1J}\mathbf{x} + \sum_{j=1}^{J-1} \left(\beta_{1j}^{'}\mathbf{x}.\mathbf{1_j}\right) + \mathbf{u}$$

On n'a pas de $\beta_{1j}^{'}\mathbf{x}.\mathbf{1_j}$ avec l'hypothèse (i), pas de $\beta_{1J}\mathbf{x}$ avec l'hypothèse (ii) et pas de $\beta_{0j}^{'}\mathbf{1_j}$ avec l'hypothèse (iii). Un test de Fisher teste alors l'égalité des variances des résidus entre les modèles comparées après estimations des β_{ij} optimums. En comparant le modèle complet à (i), (ii) ou (iii) on teste alors respectivement :

— $H_0^{(i)}$: pas d'interaction, *i.e.* $\beta_{11} = \ldots = \beta_{1J}$. Les droites partagent la même pente que $\beta_{1,J}$.

— $H_0^{(ii)}$: $\beta_{1,J} = 0$

— $H_0^{(iii)}$: $\beta_{01} = \ldots = \beta_{0J}$. Les droites partagent la même origine que $\beta_{0,J}$.

La démarche à suivre pour analyser un jeu de données avec ce modèle est alors : On commence donc par évaluer (i). Si le test n'est pas significatif, on regarde (ii) qui, s'il n'est pas non plus significatif, conduit à l'absence d'effet de la variable X. De même, toujours si (i) n'est pas significatif, on s'intéresse à (iii) pour juger de l'effet éventuel du facteur T.