3. Adéquation, Validation du Modèle

Au niveau de l'identité d'analyse de variance, l'usage de partitions pour tester officiellement l'absence de différences dans les traitements moyens exige que certaines suppositions soient satisfaites: observations suffisamment décrites par le modèle $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ et erreurs normales, $\varepsilon_{ij} \rightsquigarrow iid(0, \sigma^2)$. Dans le cas où ces suppositions sont valides, la procédure d'analyse de variance est telle qu'un test exact de l'hypothèse d'absence de différences dans les traitements moyens.

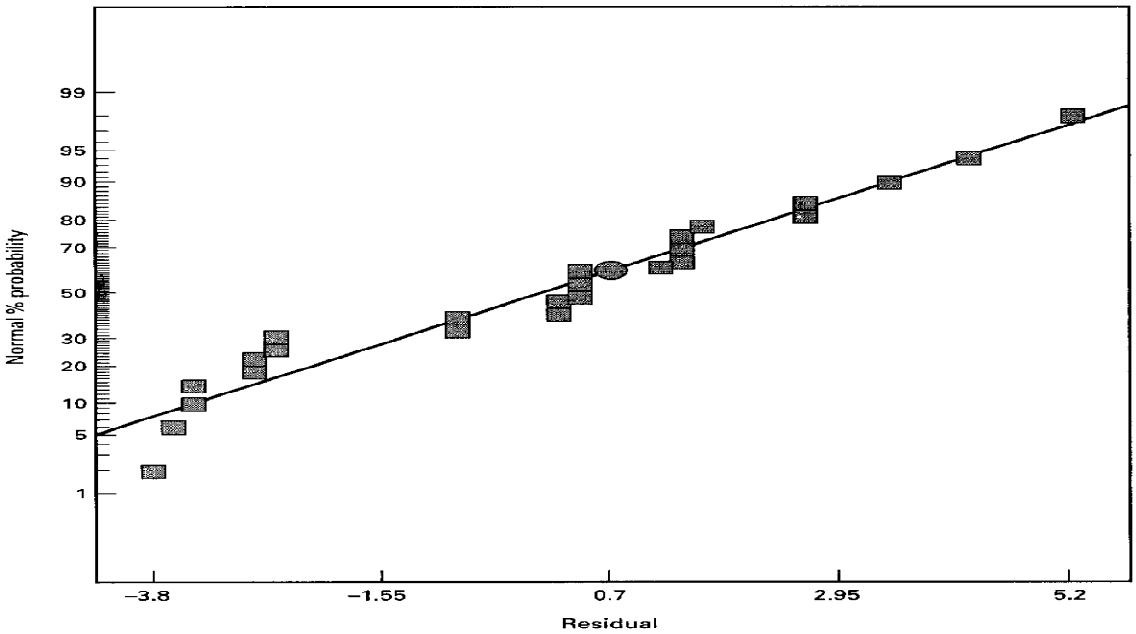
Cependant en pratique, ces hypothèses ne sont pas toujours vérifiées exactement. Violations des hypothèses fondamentales ainsi que la compétence du modèle sont vérifiées via l'analyse des résidus.

L'examen et l'étude des résidus doivent être considérés comme une tâche automatique et essentielle de toute analyse de variance: modèle satisfaisant pour des résidus moins constructifs, rapportant des schémas non évidents:

i. Hypothèse de Normalité:

L'examen de normalité peut être fait à partir de l'histogramme des résidus, ou plus précisément par le plot de probabilité normale, procédures extrêmement utiles pour l'analyse de la variance:

* Si la distribution des erreurs est normale alors le plot donne une ligne droite mettant l'accent davantage aux valeurs centrales que sur les extrêmes, voir graphique.



* Un déplacement modéré de la Normalité est "peu inquiétant" dans l'analyse de variance à effets fixes où le F-test est légèrement affecté: L'analyse de variance est bien robuste à

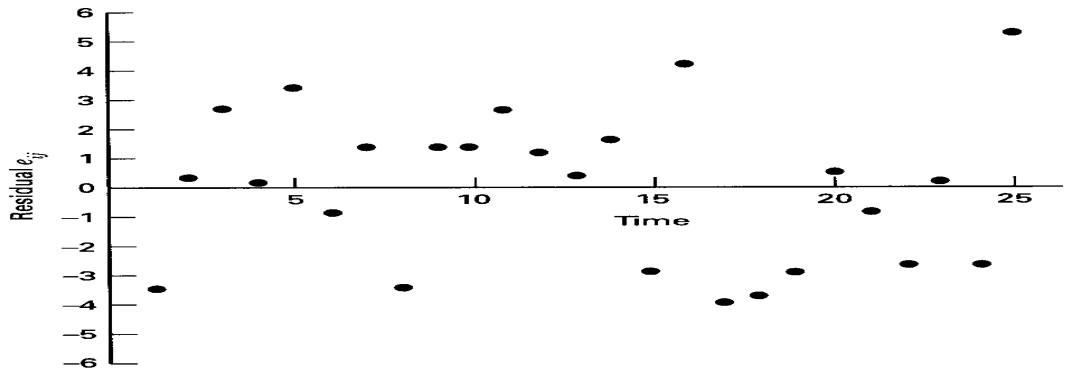
l'hypothèse de Normalité, hypothèse qui provoque le niveau réel de signification et le pouvoir de différer légèrement des valeurs.

* La présence d'outlier(s), résidu(s) plus large(s) que les autres, peut fausser cette analyse de variance.

ii. Analyse graphique des résidus dans l'ordre temporel de collecte des données aide à détecter la présence de possible corrélation entre résidus.

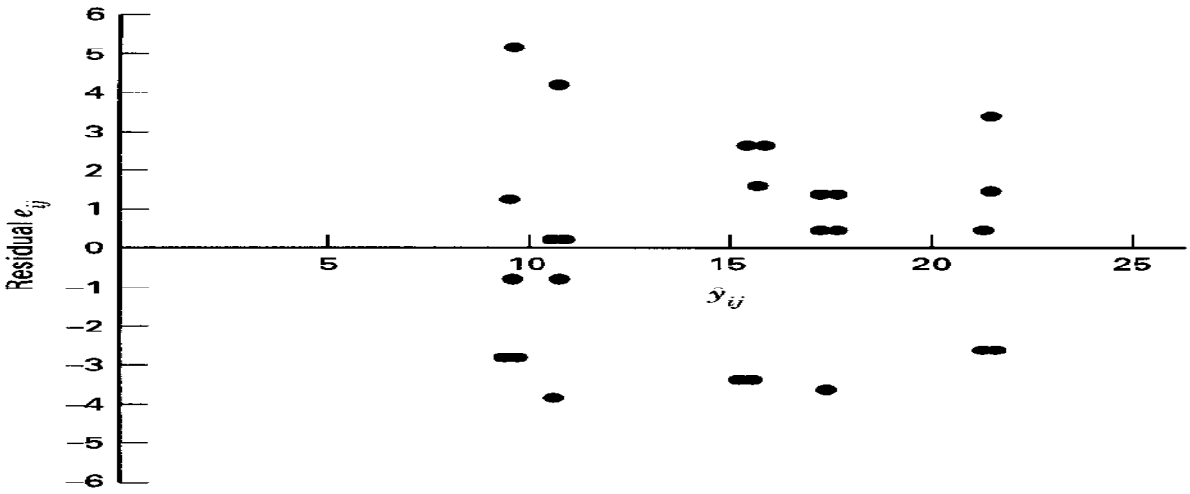
La tendance d'avoir des résultats résiduels positifs et négatifs indique une corrélation positive résiduelle, impliquant ainsi la violation de l'hypothèse fondamentale d'indépendance des erreurs: problème sérieux où il est difficile de le corriger. Il est donc important d'éviter si possible tout problème dans la collecte des données. Une bonne expérience aléatoire est considérée comme étape importante pour obtenir l'indépendance (résiduelle).

Parfois, la compétence de l'expérimentateur peut changer à mesure que l'expérience progresse ou bien le processus étudié peut dériver ou devenir plus erratique, instable; ce qui entraîne une modification de la variance de l'erreur au fil du temps; i.e. un plot résiduel dans l'ordre du temps plus dispersé et plus répandu à ces extrémités qu'un autre: Donc une variance résiduelle non constante dans le temps, considérée comme problème potentiellement sérieux. Voir plot résidus vs temps:



On observe une séquence résiduelle et temporelle de la collecte de données qui indique absence de soupçon sur une certaine violation de l'indépendance ou des hypothèses de stabilité de la variance.

iii. Analyse graphique des résidus versus la valeur estimée: Un modèle correcte et des hypothèses bien satisfaites génèrent des résidus sans structures, en particulier sans rapport avec d'autres variables, y compris la variable-réponse prédite. Le contrôle simple est fait par le graphe des résidus versus les valeurs estimées ($\hat{y}_{ij} = \bar{y}_{i.}$) qui ne doit pas révéler une certaine forme graphique évidente. Voir plot résidus versus valeurs estimées:



On note qu'aucune structure inhabituelle n'est apparente.

-- Remarque: Le défaut qui peut apparaître occasionnellement sur ce type de plot est l'instabilité de la variance: en effet, la variance des observations peut augmenter proportionnellement à l'augmentation de la grandeur de l'observation. Le cas serait si l'erreur de l'expérience était un pourcentage, fonction de la taille de l'observation; dans ce cas, les résidus augmentent plus au fur et à mesure que y_{ij} devient plus importante (en valeur) et le plot des résidus par rapport à $\hat{y}_{ij}$ ressemblerait à un entonnoir à ouverture vers l'extérieur. Aussi, le problème de variance non constante se pose dans le cas où les données suivent une distribution skewned, non normale.

Dans le cas où l'hypothèse d'homogénéité des variances est violée, le F-test est légèrement affecté dans le modèle à effets fixes balancé. Cependant, le problème est plus sérieux au niveau du modèle non balancé ou dans le cas où une variance d'erreur est beaucoup plus large que les autres variances. Aussi, dans un modèle balancé, l'inégalité des variance d'erreur peut perturber significativement les inférences sur les composantes de la variance.

L'approche habituelle pour traiter une variance non constante est d'appliquer une méthode de Transformation de stabilisation de la variance puis de mener l'analyse de variance sur les données transformées et les conclusions de l'analyse s'appliquent aux populations transformées. Exemple de méthodes de transformations, on trouve la transformation racine carrée, $y_{ij}^{*} = \sqrt{y_{ij}}$, pour des observations qui suivent la loi de poisson, la transformation logarithmique, $y_{ij}^{*} = \log y_{ij}$, pour des données de distribution log-normale, etc. Pour plus de discussions sur les transformations, on peut se référer à Bartlett, 1947- Box & Cox 1964 - Dolby 1963 et Draper & Hunter 1969.

** Tests statistiques d'égalité de variance: Bien que les plots des résidus sont utilisés fréquemment pour diagnostiquer l'inégalité des variances, plusieurs tests statistiques sont proposés aussi. Ce sont des tests formels, définis par le corps d'hypothèses suivant:

$$H_0: \sigma_1^2 = \cdots = \sigma_a^2$$
$$H_a: non\ vraie\ pour\ au\ moins\ un\ \sigma_i^2$$

La procédure utilisée est le $Bartlett - test$. Sous l'hypothèse d'une population indépendante, normale, la statistique du test est définie par:

$$\chi_0^2 = 2.3026\frac{q}{c} \rightsquigarrow \chi^2(a-1)$$

où

$$q = (N-a)\log_{10} S_p^2 - \sum_{i=1}^{a}(n_i - 1)\log_{10} S_i^2$$

$$c = 1 + \frac{1}{3(a-1)}\left[\sum_{i=1}^{a}(n_i-1)^{-1} - (N-a)^{-1}\right]$$

$$S_p^2 = \frac{\sum_{i=1}^{a}(n_i-1)S_i^2}{N-a}$$

$S_i^2\ est\ la\ variance\ d'échantillon\ du\ ième\ traitement;$

$La\ quantité\ \ q\ est\ large\ lorsque\ S_i^2\ diffère\ énormement,$

$q = 0\ lorsque\ toutes\ les\ S_i^2 sont\ égales;$

$On\ rejette\ H_0\ \ pour\ \chi_0^2 > \chi_{\alpha,a-1}^2.$

$La\ P - value\ peut\ être\ aussi\ utilisée.$

$Le\ Bartlett'test\ est\ trés\ sensible\ à\ l'hypothèse\ de\ normalité;\ dans\ le\ cas\ de\ doute\ sur\ la\ validité\ \ de\ l'hypothèse,\ on\ utilise\ le\ test\ suivant:$

La procédure du test de Levene modifié (1960), procédure robuste à toute violation ou manquement de la normalité. Pour tester l'hypothèse d'égalité de variances dans tous les traitements, le $Levene\ modifié - test$ utilise l'écart, la déviation absolu(e) des observations dans chaque traitement de la médiane de traitement, notée $\tilde{y}_i$. Ces écart, déviations sont définis par

$$d_{ij} = |y_{ij} - \tilde{y}_i|$$
$$i = 1, \cdots, a \quad et\ \ j = 1, \cdots, n_i$$

Le $Levene\ modifié - test$ évalue la moyenne de ces écarts: si elle est égale ou non pour tous les traitements. Il s'avère que si les déviations moyennes sont égales alors les variances de toutes les observations dans tous les traitements sont égales aussi. La statistique du $Levene\ modifié - test$ est simplement la statistique $F - ANOVA$ standard, habituelle pour tester l'égalité des moyennes, appliqués aux écarts absolus.