SAYD Yassine

Université de Montpellier, Faculté des sciences

Master 2 MIND

HMMA 307

Département de Mathématiques

2020-2021

Analyse de la covariance

Sous la direction de :

Joseph SALMON

Table des matières

Introduction

L'analyse de la covariance (ancova) est utilisé lorsqu'on souhaite comparer deux ou plusieurs lignes de régression les unes aux autres ; ancova nous dira si les lignes de régression sont différentes les unes des autres dans la pente ou l'interception.

Cette approche est utilisée lorsqu'on a deux variables de mesure et une variable nominale. La variable nominale divise les régressions en deux ensembles ou plus. Les variables de mesure sont, comme son nom l'indique, des choses que vous pouvez mesurer. Une observation individuelle d'une variable de mesure est toujours un nombre. Tanids que les variables nominales classent les observations en catégories distinctes.

Le but d'ancova est de comparer deux ou plusieurs lignes de régression linéaires. C'est une façon de comparer la variable Y entre les groupes tout en contrôlant statistiquement la variation de Y causée par la variation de la variable X.

Walker (1962) a étudié les chants d'accouplement des grillons mâles. Chaque coup d'aile par un grillon produit une impulsion de chant, et les femelles peuvent utiliser le nombre de impulsions par seconde pour identifier les mâles de l'espèce correcte. Walker voulait savoir si les gazouillis des grillons Oecanthus exclamationis et Oecanthus niveus avaient des taux de pouls différents.Il a donc mesuré le pouls des grillons à une variété de températures. Dans ce projet, nous allons nous intéresser à l'analyse de la covariance sur le jeu de données Crickets de Walker.

Chapitre 1

Extraction des données et analyse descriptive

La figure 1.1, ci-dessous, présente le jeu de données de Walker :

O. exclamationis		***O. niveus***	
Température (C°)	**Impulsions par seconde**	**Température (C°)**	**Impulsions par seconde**
20.8	67.9	17.2	44.3
20.8	65.1	18.3	47.2
24.0	77.3	18.3	47.6
24.0	78.7	18.3	49.6
24.0	79.4	18.9	50.3
24.0	80.4	18.9	51.8
26.2	85.8	20.4	60.0
26.2	86.6	21.0	58.5
26.2	87.5	21.0	58.9
26.2	89.1	22.1	60.7
28.4	98.6	23.5	69.8
29.0	100.8	24.2	70.9
30.4	99.3	25.9	76.2
30.4	101.7	26.5	76.1
		26.5	77.0
		26.5	77.7
		28.6	84.7
Veux dire	85.6	Veux dire	62.4

Figure 1.1 – Jeu de données Crickets.

Dans un premier temps, on a procédé à une analyse descriptive pour mieux visualiser les données. Cette analyse comprend les histogrammes correspondant aux 4 variables, le calcul des valeurs remarquables (nombre total, moyenne, écart-type, min, 1er quartile, médiane, 3ème quartile, max), la comparaison de leurs distributions et les scattergraphs du nombre d'impulsions par seconde en fonction de la température chez les deux groupes.

On va se contenter de présenter uniquement les scattergraphs comme le montrent les figure 1.2 et 1.3 ci dessous :

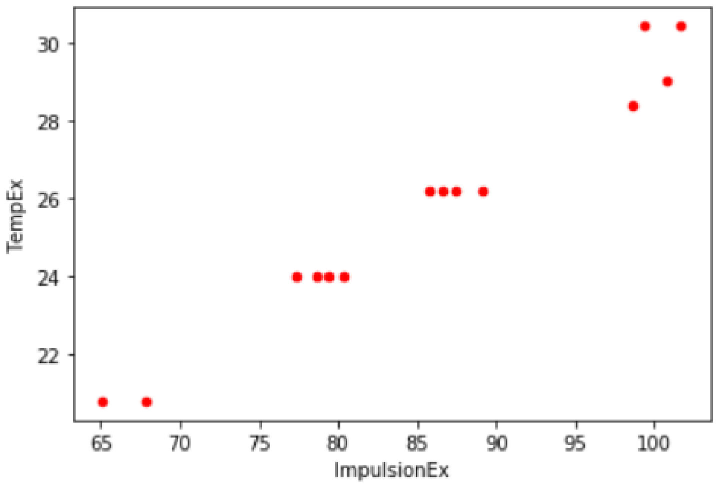

FIGURE 1.2 – Scattergraphs des Oecanthus exclamationis.

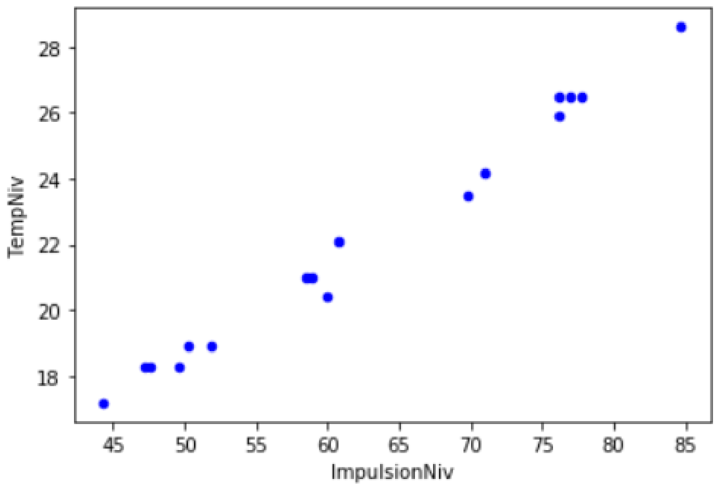

FIGURE 1.3 – Scattergraphs des Oecanthus niveus.

Chapitre 2

Test de Student

Dans un premier temps, on ignore les températures et on compare simplement les fréquences d'impulsion moyennes chez les 2 groupes.Oecanthus exclamationis a un taux plus élevé que Oecanthus niveus, et la différence des moyennes est très significatives (23.2).

Nous allons vérifier cela grâce au test de Student. On obtient une p-value de l'ordre de 2×10^5, donc on rejette l'hypothèse nulle d'égalité des moyennes.

Comme la température moyenne pour les mesures Oecanthus exclamationis était de 3.6 degrès Celsus plus élevé que pour Oecanthus niveus, il est insensé de négliger la température. Oecanthus exclamationis pourrait avoir un taux plus élevé que Oecanthus niveus à certaines températures, mais pas d'autres.

On peut contrôler la température avec l'ancova, qui nous dira si la ligne de régression pour Oecanthus exclamationis est plus élevée que la ligne pour Oecanthus niveus ; si c'est le cas, cela signifie que Oecanthus exclamationis aurait un taux de pouls plus élevé à n'importe quelle température.

Chapitre 3

Ancova : Analyse de la covariance

Nous testons deux hypothèses nulles dans une ancova. L'équation d'une ligne de régression prend la forme $Y = a + bX$, où a est l'interception Y et b est la pente. La première hypothèse nulle de l'ancova est que les pentes des lignes de régression (b) sont toutes égales ; en d'autres termes, que les lignes de régression sont parallèles les unes aux autres. Si on accepte l'hypothèse nulle selon laquelle les lignes de régression sont parallèles, on teste la deuxième hypothèse nulle : que les interceptions Y des lignes de régression (a) sont toutes les mêmes.

On a appliqué une régression OLS sur les 2 groupes, comme le montrent les figures 3.1 et 3.2 ci-dessous :

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            ImpulsionEx   R-squared:                       0.972
Model:                            OLS   Adj. R-squared:                  0.970
Method:                 Least Squares   F-statistic:                     416.6
Date:                Sun, 08 Nov 2020   Prob (F-statistic):           1.10e-10
Time:                        15:11:19   Log-Likelihood:                -28.751
No. Observations:                  14   AIC:                             61.50
Df Residuals:                      12   BIC:                             62.78
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -11.0408      4.765     -2.317      0.039     -21.424      -0.658
TempEx         3.7514      0.184     20.410      0.000       3.351       4.152
==============================================================================
Omnibus:                        0.039   Durbin-Watson:                   1.533
Prob(Omnibus):                  0.981   Jarque-Bera (JB):                0.268
Skew:                           0.002   Prob(JB):                        0.874
Kurtosis:                       2.322   Cond. No.                         227.
==============================================================================
```

FIGURE 3.1 – Résultats de la régression OLS chez les Oecanthus exclamationis.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           ImpulsionNiv   R-squared:                       0.987
Model:                            OLS   Adj. R-squared:                  0.986
Method:                 Least Squares   F-statistic:                     1128.
Date:                Sun, 08 Nov 2020   Prob (F-statistic):           1.57e-15
Time:                        14:44:15   Log-Likelihood:                -30.321
No. Observations:                  17   AIC:                             64.64
Df Residuals:                      15   BIC:                             66.31
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -15.3893      2.347     -6.558      0.000     -20.391     -10.387
TempNiv        3.5175      0.105     33.585      0.000       3.294       3.741
==============================================================================
Omnibus:                        4.526   Durbin-Watson:                   1.743
Prob(Omnibus):                  0.104   Jarque-Bera (JB):                2.535
Skew:                           0.924   Prob(JB):                        0.282
Kurtosis:                       3.402   Cond. No.                         142.
==============================================================================
```

FIGURE 3.2 – Résultats de la régression OLS chez les Oecanthus niveus.

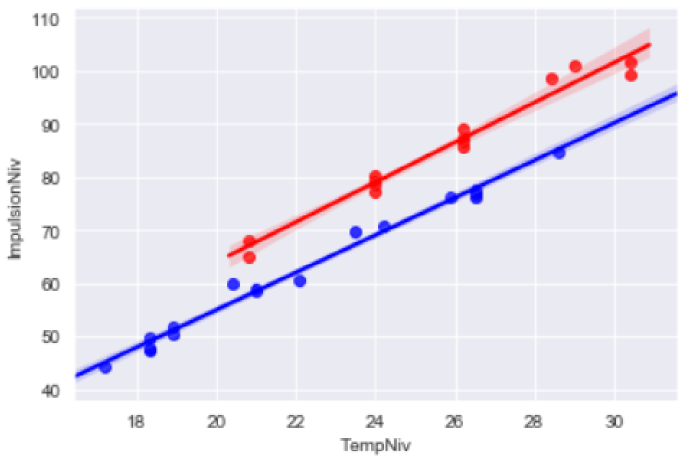

FIGURE 3.3 – Lignes de régression chez les 2 groupes.

Nous avons également superposer les lignes de régression chez les Oecanthus exclamationis (rouge) et chez les Oecanthus niveus (bleu) comme le montre la figure 3.3 ci-dessus. La ligne de régression pour Oecanthus exclamationis est $Y = 3,75X - 11,0$ et la ligne pour Oecanthus niveus est $Y = 3,52X - 15,4$. La ligne de régression pour Oecanthus exclamationis est plus élevée que la ligne pour Oecanthus niveus ; cela signifie que Oecanthus exclamationis aurait un taux de pouls plus élevé à n'importe quelle température.
La première hypothèse nulle de l'ancova est que les pentes des lignes de régression sont toutes égales ; en d'autres termes, que les lignes de régression sont parallèles les unes aux autres. On va acceptez l'hypothèse nulle selon laquelle les lignes de régression sont parallèles et nous testerons la deuxième hypothèse nulle : que les interceptions des lignes de régression sont toutes les mêmes.

Les pentes ne sont pas significativement différentes (P=0,25) ; la pente commune est de 3,60 ce qui se trouve entre les pentes pour les lignes séparées (3,52 et 3,75). Le dernier test dans l'ancova est de tester l'hypothèse nulle que toutes les interceptions Y des lignes de régression avec une pente commune sont les mêmes. Parce que les lignes sont parallèles, en disant qu'ils sont significativement différents à un moment donné (l'interception Y) signifie que les lignes sont différentes à tout moment. Sur cette partie-là, je n'ai pas réussi à tester cette hypothèse afin de vérifier la 2ème hypothèse nulle d'ancova.

Bien que l'utilisation la plus courante de l'ancova soit de comparer deux lignes de régression, il est possible de comparer trois régressions ou plus. Si leurs pentes sont toutes les mêmes, vous pouvez tester chaque paire de lignes pour voir quelles paires ont des interceptions Y significativement différentes, en utilisant une modification du test Tukey-Kramer. On a procédé à un test de Tukey sous l'hypothèse que leurs pentes sont toutes les mêmes, une partie des résultats sont données dans les figures 3.4 et 3.5 ci-dessous :

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=================================================
group1 group2 meandiff p-adj  lower upper reject
-------------------------------------------------
  20.8   24.0    12.45 0.5566   nan   nan  False
  20.8   26.2    20.75 0.5566   nan   nan  False
  20.8   28.4     32.1 0.5566   nan   nan  False
  20.8   29.0     34.3 0.5566   nan   nan  False
  20.8   30.4     34.0 0.5566   nan   nan  False
  20.8    nan      nan 0.5566   nan   nan  False
  20.8    nan      nan 0.5566   nan   nan  False
  20.8    nan      nan 0.5566   nan   nan  False
  24.0   26.2      8.3 0.5566   nan   nan  False
  24.0   28.4    19.65 0.5566   nan   nan  False
```

FIGURE 3.4 – Résultats du test de Tukey chez les Oecanthus exclamationis.

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
group1 group2 meandiff p-adj   lower   upper  reject
-----------------------------------------------------
  17.2   18.3   3.8333 0.1844 -1.5031  9.1697  False
  17.2   18.9     6.75 0.0221  1.0899 12.4101   True
  17.2   20.4     15.7  0.001  9.1643 22.2357   True
  17.2   21.0     14.4  0.001  8.7399 20.0601   True
  17.2   22.1     16.4  0.001  9.8643 22.9357   True
  17.2   23.5     25.5  0.001 18.9643 32.0357   True
  17.2   24.2     26.6  0.001 20.0643 33.1357   True
  17.2   25.9     31.9  0.001 25.3643 38.4357   True
  17.2   26.5  32.6333  0.001 27.2969 37.9697   True
  17.2   28.6     40.4  0.001 33.8643 46.9357   True
```

FIGURE 3.5 – Résultats du test de Tukey chez les Oecanthus niveus.

Chez les Oecanthus exclamationis le test de Tukey ne rejette pas l'hypothèse nulle chez les Oecanthus exclamationis. Donc pas de différences significatives au niveau des intercepts entre chaque paire de lignes.

Chez les Oecanthus niveus le test de Tukey rejette quasiment à chaque fois l'hypothèse nulle chez les Oecanthus niveus. Donc il y a des différences significatives au niveau des intercepts entre la plupart des paires de lignes.

Chapitre 4

Ancova : Analyse de puissance

Afin d'analyser la puissance, j'ai utilisé la feuille de calcul en utilisant la méthode de Borm et al. (2007). Il ne fonctionne que pour ancova avec deux groupes, et il suppose que chaque groupe a le même écart type et le même r^2. Pour l'utiliser, on a besoin de : la taille de l'effet ou la différence dans les interceptions Y qu'on espère détecter, l'écart type : il s'agit de l'écart type de toutes les valeurs Y de chaque groupe (sans contrôle de la variable X), alpha ou le niveau d'importance (habituellement 0,05) ; la puissance i.e la probabilité de rejeter l'hypothèse nulle lorsque la taille de l'effet donné est la vraie différence (0,80 ou 0,90 sont des valeurs communes),le r^2 au sein des groupes. J'ai pris comme référence les valeurs des Oecanthus niveus. Les résultats sont donnés dans la figure 4.1 ci-dessous :

effect size:	0,2
standard deviation:	2,347
alpha:	0,05
power:	0,8
r-squared:	0,987
sample size needed:	29

Figure 4.1 – Résultats de la méthode de Borm et al. (2007).

Cela signifie qu'on aura besoin de 29 grillons Oecanthus exclamationis et de 29 grillons Oecanthus niveus.