**RESUME SECTION 1.4 INF232**

**ANALYSE DE DONNEES**

**Membre du groupe :**

|  |  |  |  |
| --- | --- | --- | --- |
| **N°** | **Nom et Prénom** | **Matricule** | **Pourcentage Participation** |
| 1 | Douanla Sonhafo Champlain | 21T2657 | 100% |
| 2 | Kouam Noubissi Brice | 21T2432 | 100% |
| 3 | Meka Moise Christian Junior | 21T2561 | 100% |
| 4 | Rudy Tchamba | 21T2981 | 100% |

**Titre : CAS D’ERREURS GAUSSIENNES**

Mieux que les expressions des estimateurs et des variances pour la régression linéaire, la connaissance de lois permet de retrouver leur région de confiance et ainsi effectuer des tests d’hypothèses. Le but ici est de pouvoir généraliser les estimations faites sur notre échantillon à la population et ainsi voir si notre modèle de régression linéaire peut être extrapolé ou généralisé à la population.

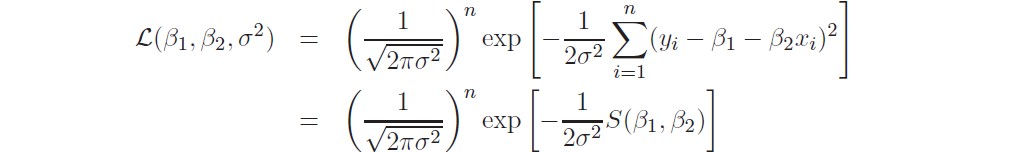
Tout d’abord faisons les hypothèses suivantes :

Le modèle de régression linéaire simple devient un modèle paramétrique à paramètre (). La loi de étant connue, les lois des deviennent et les sont mutuellement indépendantes.

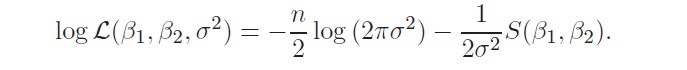
1. **Estimateur du maximum de vraisemblance**

Pour pouvoir () nous allons utiliser le maximum de vraisemblance et chercher les paramètres (, , ) qui maximise notre vraisemblance.

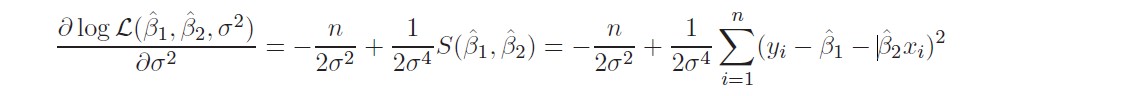
L’expression de la vraisemblance est :



Etant donné que manipuler une telle fonction semble complexe, nous calculons la log vraisemblance :



Nous voulons maximiser cette quantité pour (). Les paramètres appartiennent uniquement au terme -S qu’il faut minimiser (car est -S). Or on a déjà les quantités et des MCO qui minimisent ce terme. On cherche maintenant à maximiser log L () par rapport à .



On en déduit l’estimateur du maximum de vraisemblance de par :

=

Cet estimateur est biaisé car E (. On doit le rendre sans biais. Ainsi, =

Mais pour un nombre d’observation suffisamment grand, ce biais est négligeable.

1. **Lois usuelles**

Les lois d’usage constant en régression linéaire sont : **loi de khi-2, loi de Student et loi de Fisher.**

Dans cette partie, nous allons à tour de rôle décrire ces différentes lois et donner un cas d’utilisation pour chacune d’elles.

1. **Loi du khi-2 (ꭓ2)**

Soit x1, x2, …, xn des variables aléatoires iid (indépendantes et identiquement distribuées) telles que .

La loi de X= ꭓ2(n). Où ꭓ2(n) est la loi du ꭓ2 à n degré de liberté (ddl).

Cette loi est généralement utilisée pour tester l’indépendance de deux séries statistiques à valeur qualitative (exemple savoir si la région d’origine a une influence sur le statut social. Ici on essaie de mesurer la distance entre la valeur observée et la valeur théorique et on compare le résultat à la valeur théorique du khi-2 à **(c-1)(l-1) ddl** où c et l sont respectivement le nombre de catégories dans la variable région et dans la variables statut social).

1. **Loi de Student**

Soit Z une variable aléatoire telle que Z , soit X une variable aléatoire suivant le ꭓ2(n). La variable aléatoire T = est appelé loi de Student à n ddl et est noté T *T*n.

**NB :** Elles sont au nombre de trois :

* Lorsque n=1, T est une loi de Cauchy et n’a ni espérance, ni variance ;
* Lorsque n=2, T est centrée mais de variance infinie ;
* Pour n ≥ 3, T est centrée et de variance ;
* Lorsque n devient grand, par la loi des grands nombres T et

La loi de Student est généralement utilisée lorsque l’on veut faire un test sur la moyenne. C’est-à-dire tester si la moyenne de la population prend une valeur particulière (Par exemple pour répondre à l’affirmation : en moyenne les étudiants de l’Université de Yaoundé 1 sont plus âgés que les étudiants de Université de Dschang. Il s’agira ici de comparer à travers un test d’hypothèse si les moyennes des deux populations sont égales ou si celle de UY1 est supérieure à celle de Uds. On calculera une statistique de test qui suivra une loi de **Student**. Puis de comparer la valeur calculée à la valeur théorique se trouvant dans la table de loi du **Student** pour le ddl correspondant).

1. **Loi de Fisher**

Soient U1 ꭓ2(n1) et U2 ꭓ2(n2) deux variables aléatoires indépendantes. La loi de F = est la loi de Fisher à (n1, n2) ddl. On note F .

NB : Pour n2≥2, l’espérance de est . Dans la suite, n2 sera grand de sorte qu’à nouveau la loi des grands nombres implique que 1. Dans ce cas, F peut être vu comme un ꭓ2 normalisé par son ddl : F ꭓ2n1/n1.

L’utilisation de la loi de Fisher se retrouve dans la comparaison des variances. C’est-à-dire tester si les variances de deux populations sont égales (Par exemple pour répondre à l’affirmation : la dispersion des âges des étudiants de l’Université de Yaoundé 1 est supérieure à celle des étudiants de Université de Dschang. Il s’agira ici de comparer à travers un test d’hypothèse si les variances des deux populations sont égales ou si celle de UY1 est supérieure à celle de Uds. On calculera une statistique de test qui suivra une loi de **Fisher**. Puis de comparer la valeur calculée à la valeur théorique se trouvant dans la table de loi du **Fisher** pour le ddl correspondant).

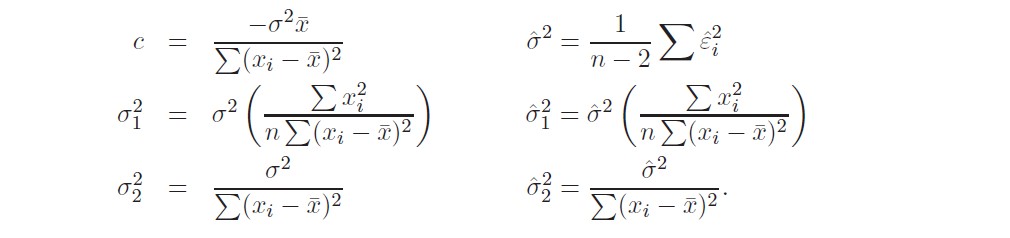
1. **Cas d’utilisation dans la régression linéaire**

Dans le cadre de la régression linéaire, ces trois lois sont utilisées pour pouvoir tester si notre modèle peut être généralisé à la population avec :

* Un test de Student sur paramètre (Ici on test si avec pour hypothèse nulle ). Cela permet de voir si effectivement la variable explicative X explique réellement la variable expliqué Y.
* Un test de Fisher qui permet de voir tout comme le test de Student si . Cela se fait par le calcul d’une statistique de test F (qui sera présenté plus bas) que l’on comparera à la valeur théorique de . Ici on utilise implicitement ꭓ2 pour ce qui est de la variance des erreurs du modèle.

1. **Lois des estimateurs et région de confiance**

Afin de faciliter la lecture, nous poserons :



Où C, , sont mes variances et covariance des estimateurs MCO. Et et sont les variances de et

Les lois des estimateurs sont :

1. Où et V=

=

1. ꭓ2n-2
2. et sont indépendants.

On aura alors :

1. *T*n-2
2. *T*n-2

Cette propriété nous permet de donner l’intervalle de confiance (IC) des estimateurs.

1. **Intervalles de confiance**

Ci-dessus nous avons fait une estimation ponctuelle des paramètres (). Or il est plus pratique de donner une plage de valeur dans laquelle la valeur réelle du paramètre pourrait s’y trouver. Cette plage de valeur est l’intervalle de confiance (IC). Ainsi pour nos paramètres () l’estimation par intervalle de confiance est :

1. IC ( = Où est le quantile de niveau de la loi de Student à n-2 ddl
2. IC ( = Où est le quantile de niveau de la loi de Student à n-2 ddl
3. IC ( = Où est le quantile de niveau de ꭓ2n-2.
4. **Validation du modèle de régression linéaire**

Les paramètres de notre modèle étant estimés il est intéressant de savoir si notre modèle estimé peut-être généralisé ou extrapolé dans la population. Pour ce faire nous avons le choix entre un test de significativité de et un test de comparaison de la somme des carrés des estimés (SCE) et la somme des carrés résiduels (SCR).

1. Validation à partir de

Dans ce cas, un modèle de régression linéaire simple est valide lorsque la variable explicative X explique effectivement la variable expliquée Y. Cela revient à avoir **.** On fera donc un test bilatéral de comparaison :

Dont la statistique de test est : t*T*n-2

Notre modèle sera valide si et seulement si la valeur calculée de la statistique de test en valeur absolue est supérieure à .

1. Validation à partir de la comparaison de SCE et SCR

Pour valider notre modèle de régression linéaire simple en utilisant la comparaison de SCE et SCR, nous calculons le test statistique :

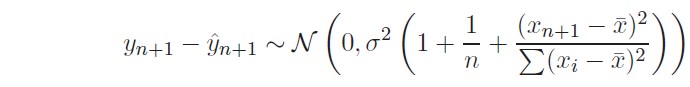
↝

Ensuite nous comparons cette valeur calculée au quantile à 1 et n-2 ddl à 1-α de confiance. Le modèle sera valide lorsque la valeur calculée sera supérieure à la valeur théorique se trouvant dans la table de loi de Fisher.

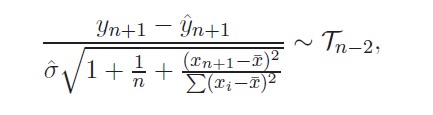
1. **Prévision**

Le but principal de la modélisation étant de faire une représentation simplifiée la réalité d’une part, et de faire de la prévision d’autres part ; notre modèle de régression linéaire validé doit nous permettre de prédire une nouvelle valeur pour un donné.

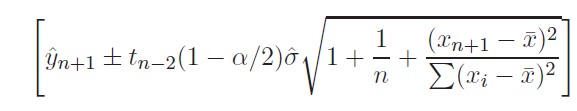
Étant linéaire en et , on peut préciser sa loi :



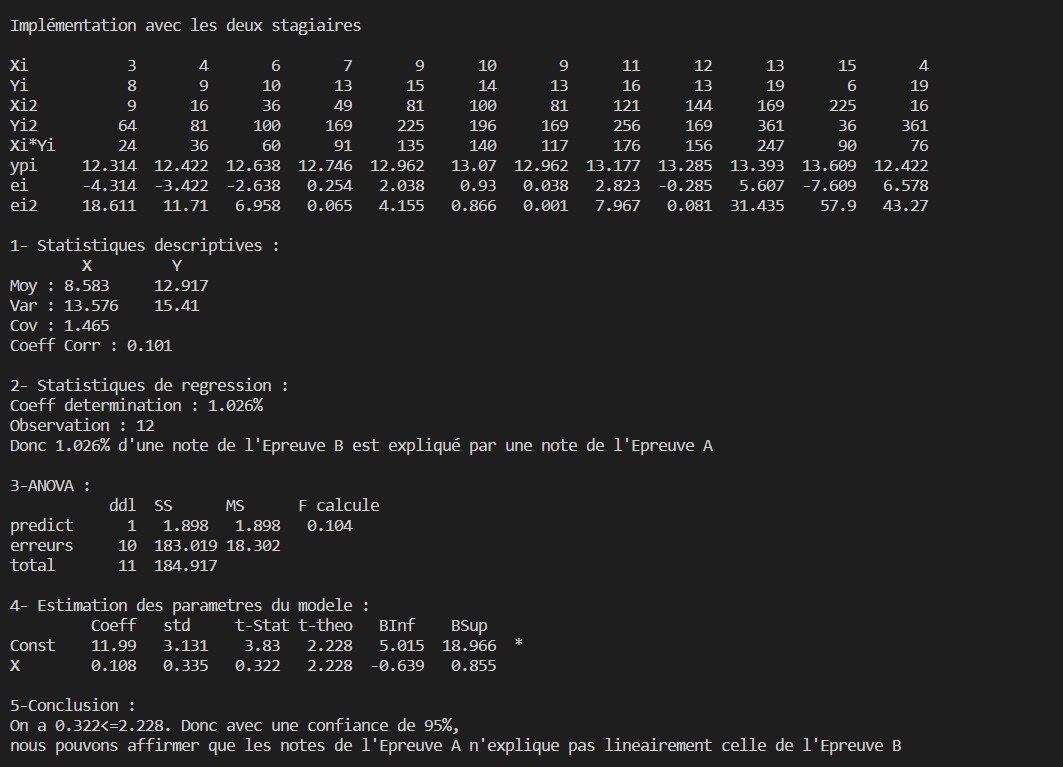
On ne connait pas . On l’estime avec et on peut alors donner l’intervalle de confiance de en utilisant la loi :



Et on obtient :



1. **Cas pratique (Implémentation en Python)**

****

