## **Deuxième partie : Introduction**

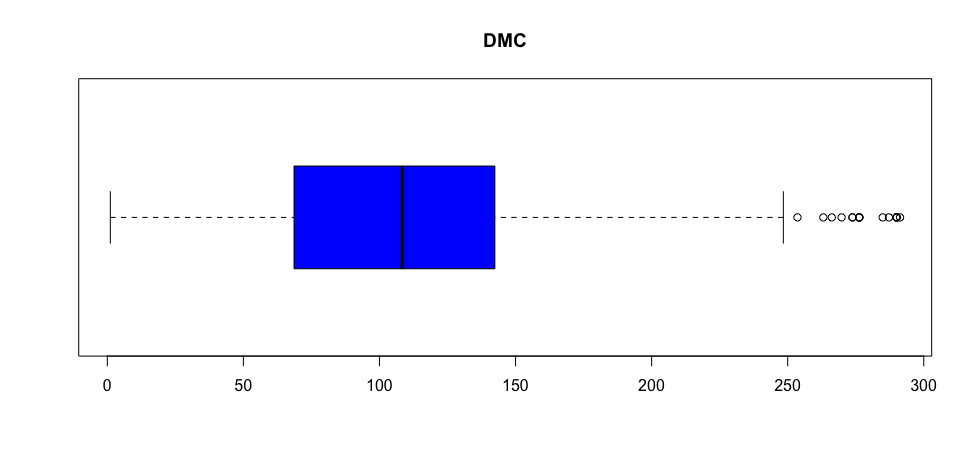
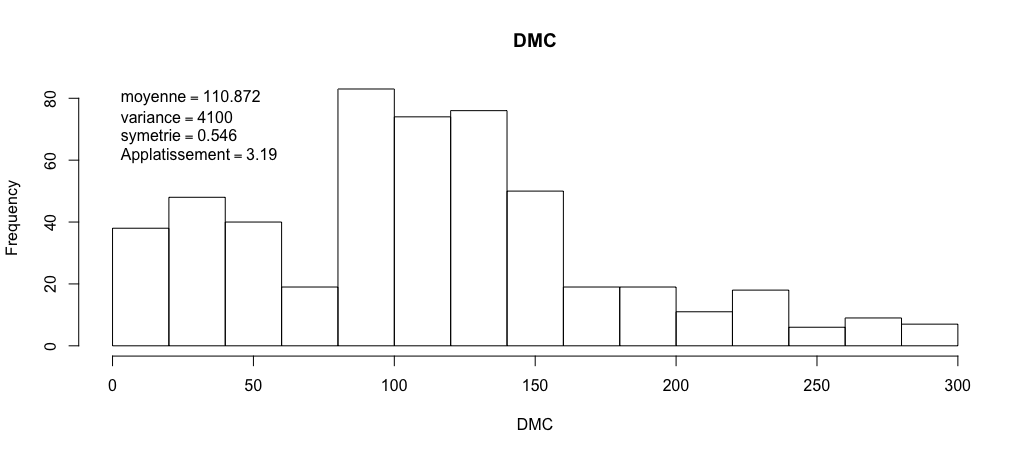
Dans la première partie de ce rapport nous avons explicité le contexte général des données sur lesquelles nous comptons effectuer nos analyses. En effet, nous y avons exposé en détails le jeu de données « *Forest Fires »*. Cette explication du contexte s’est conclue avec quelques questions ouvertes qui permettraient de mieux comprendre les feux de forêts qu’on peut observé dans le parc de Montesinho. Dans cette deuxième partie du rapport il sera question de soutirer de l’information de ces données afin de mieux comprendre ce qui caractérisent ces feux de forêts. Par la suite, nous utiliserons ces notions afin de répondre aux questionnements proposés dans la première partie du rapport. Afin de bien comprendre le phénomène à l’étude il est essentiel de débuter par une phase d’analyse statistique descriptive. En effet, les 13 variables qui composent notre jeu de donnée, ne sont pas intelligibles, c’est-à-dire qu’il est difficile de ressentir les tendances qui émanent de celle-ci lorsqu’on les visualise dans une matrice. Autrement dit, il nous faut décrire les données en les résumant et en les structurant [4 p.192]. À partir de ces descriptions il nous sera possible de remarquer certaines propriétés qui caractérisent une forêt favorisant les incendies. Ensuite, nous exposerons les potentielles distributions probabilistes sous-jacentes à certaines de ces variables. Pour ce faire, nous utiliserons d’abord notre instinct visuel. En effet, certaines de ces variables semblent d’emblée exprimée des signes de distribution normale. Nous utiliserons ensuite le test du chi carré et de afin de vérifier si il est probable ou non que ces variables soient effectivement issues d’une distribution normale. Ces vérifications seront fait sous la proposition d’une hypothèse nulle stipulant qu’elles suivent effectivement une loi normale. Comme nous le verrons, ces hypothèse s’avèreront improbable selon la *p-value* choisi. Afin de renforcir cette conclusion, nous utiliserons ensuite le teste de *Shapiro Wilk* selon la même p-value. Enfin, nous terminerons notre analyse statistique par une proposition d’un modèle de régression sur deux variables qui semblent partager une interaction en lien avec une question que nous avions laisser ouverte dans la première partie concernant le lien potentiel entre l’indice d’humidité du humus et la vitesse de propagation des feux. Finalement, nous conclurons en soulignant les résultats qui nous semblent pertinents et en posant quelques ouvertes.

## **1) Analyse descriptive des données**

Dans le but de bien cerner certaine tendance de l’état de la forêt lors d’incendie dans le parc de Montesinho, nous avons généré des histogrammes pour les variables qui composent le jeu de données. Nous avons joint à ces figures des légendes contenant certaines statistiques descriptives des variables à l’étude. Ces statistiques sont la moyenne, l’écart-type, le coefficient de symétrie et le coefficient d’aplatissement. Nous avons ignoré certaines variables comme le jour et les mois et les précipitations. Les deux premières car elle ne nous semblent d’aucun intérêt pour le moment et la dernière, car comme on s’en doute, les précipitation lors les incendies de forêts sont toujours nulle.

## **1.1) L’indice d’humidité de l’humus (*DMC)***

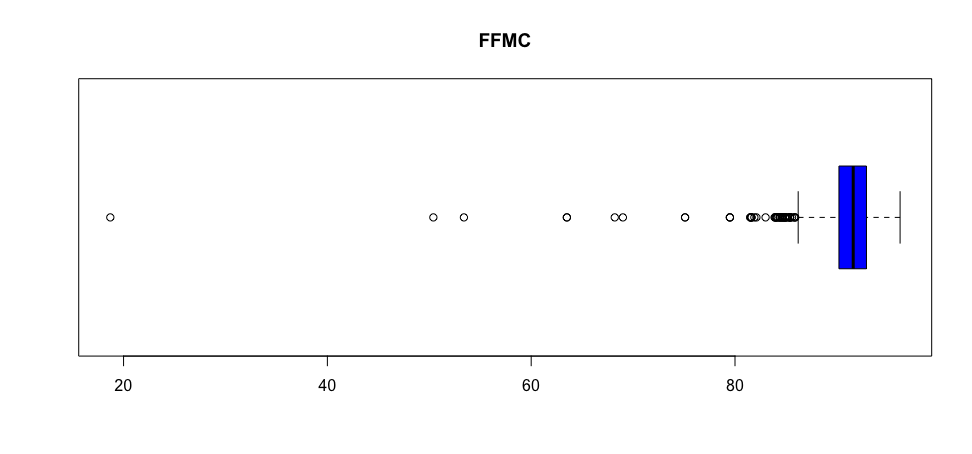
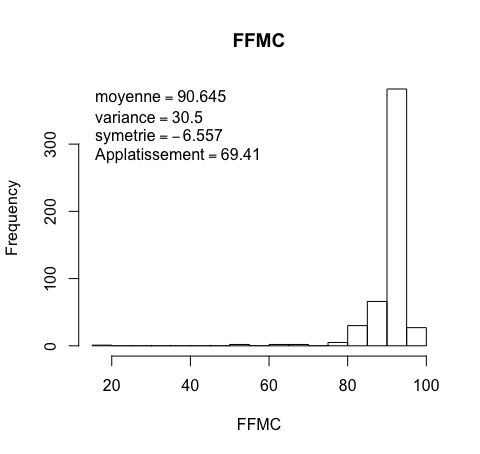
Comme mentionné dans la première partie, le *DMC* représente l’indice d’humidité du humus, c’est-à-dire le facteur de combustion de la couche organique situé entre 5cm et 10cm sous le sol. Il est à noter que cette section du sol n’est pas affectée par les pluies de moins de 20mm. Cette section du sol est aussi protégée des périodes de sècheresse rapide. On remarque sans surprise que les instances de feux de forêts sur un sol ayant un indice de *DMC* très élevé se font plus rare. Par contre, on remarque une certaine tendance : le plus grand nombre d’instance de feux on une métrique de *DMC* qui semblent au milieu de l’échelle. On remarque aussi un écart type assez grand par rapport à l’échelle en question. Les index d’humidité du humus sont donc plutôt dispersés dans un scénario de feux de forêts. Cet indice sera utilisé dans les sections qui suivent. En effet, comme mentionné dans à la fin de la section 1, nous aimerions savoir si il existe une relation entre le DMC d’une forêt et son indice ISI (voir page 2). L’histogramme de fréquence des métrique observées de cet indice en plus de sa boîte à moustache.

**

**Figure 1 Histogramme et boite à moustache de l'indice DMC**

## **1.2) Indice de combustible léger (*FFMC*)**

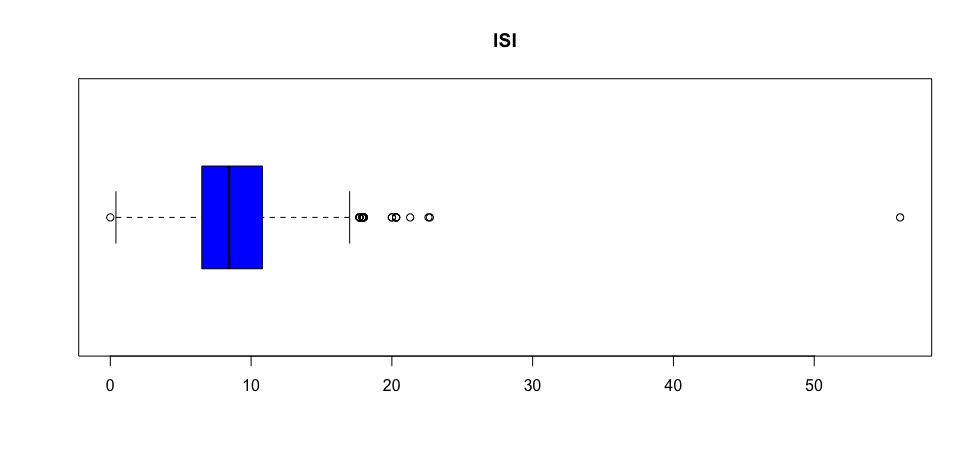
L’indice de combustible léger correspond au niveau d’humidité de végétation de moins de 1cm de diamètre (foin, herbes, petites branches). On s’intéresse alors ici à une couche de la végétation un petit peu plus en surface que celle précédemment étudié. Contrairement au *DMC,* le *FFMC* est beaucoup sensible au changement de température et aux précipitations du au fait que le matériel organique se retrouve en surface. On remarque dans l’histogramme du FFMC que les feux de forêts sont très similaires quant à leur indice de combustible léger. En effet, on remarque que l’écart-type est de 5.52 ce qui est très petit compte tenu de l’échelle en question. La boite à moustache confirme cette idée de par sa forme compacte.

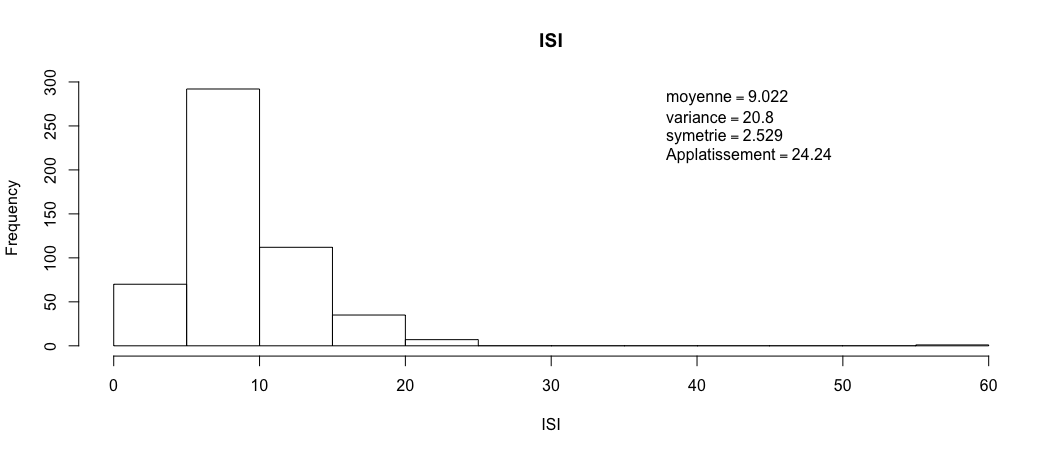


**Figure 2 Histogramme et boite à moustache du FFMC**

## **1.3) Indice de propagation initiale (*ISI)***

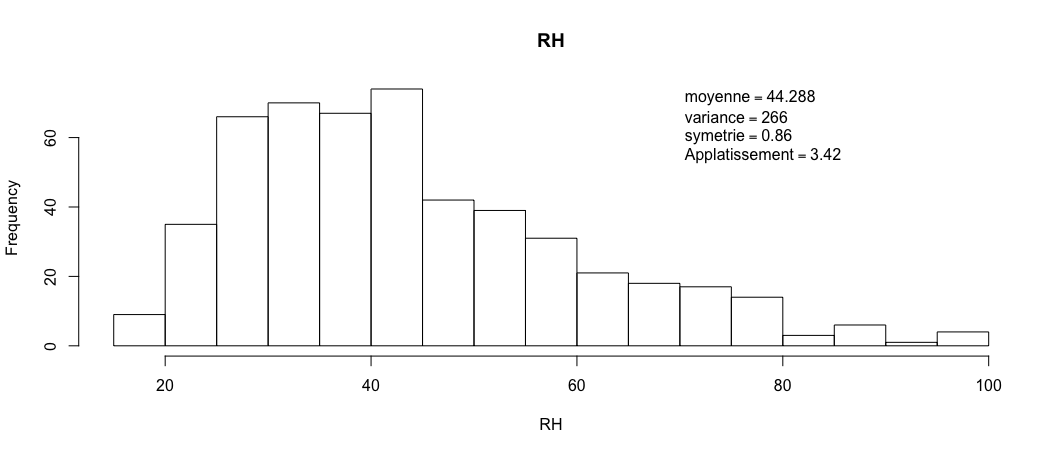
L’indice de propagation initiale (*ISI)* correspond à une métrique servant à évaluer à quel point un feu gagne en envergure par rapport au temps. Encore une fois, on remarque une certaine tendance dans début de l’histogramme. La boite a moustache montre bien qu’un feu typique semble avoir un ISI qui gravite autour de 10.

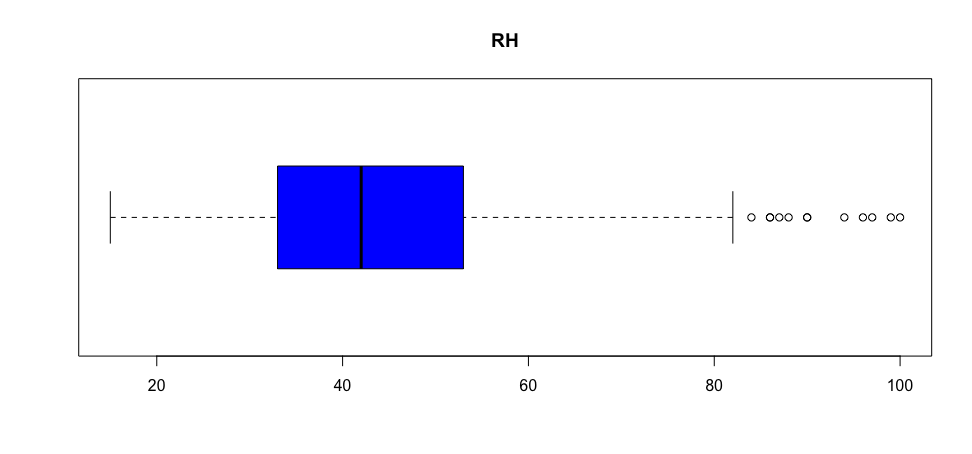




## **1.4) Humidité relative (*RH*)**

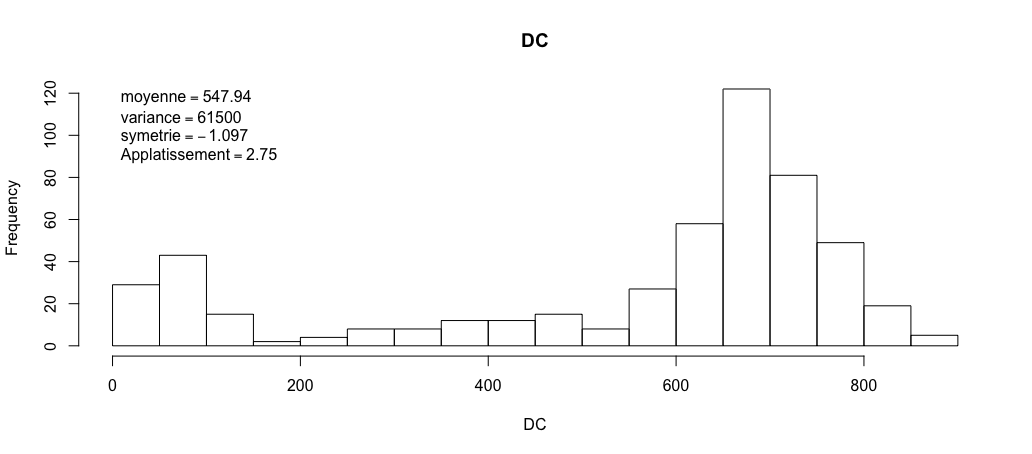
En ce qui concerne le niveau d’Humidité relative, on observe une progression vers le bas au niveau des fréquence plus le niveau d’humidité augmente. Il est intéressant de remarquer que les feux typiques ne profitent pas particulièrement d’un niveau d’humidité relative en dessous de 30. Ceci est relativement surprenant. En effet, on s’aurait attendu à ce que le plus grand nombre d’instance de feux de forêts se produisent lorsque l’indice d’humidité est au plus bas. La boite à moustache et l’histogramme témoigne tout les deux

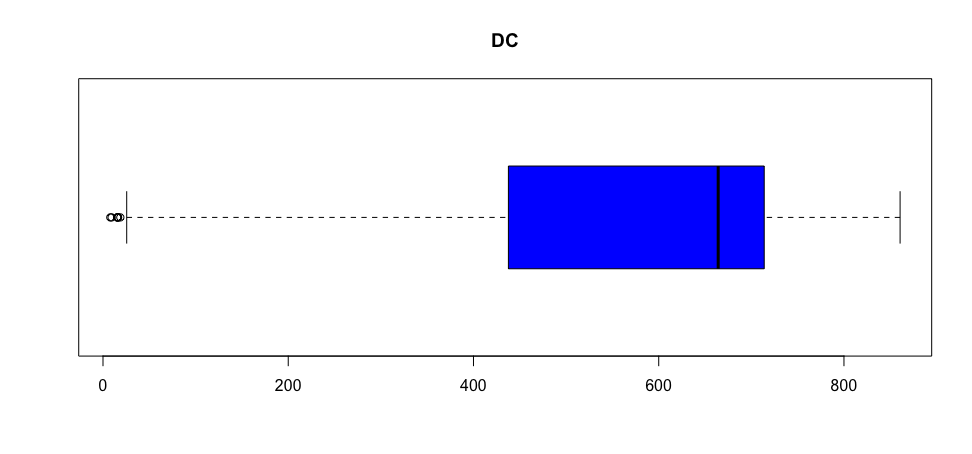




## **1.5) Indice de sècheresse (*DC*)**

L’indice de sècheresse correspond à un indice d’humidité du contenu en profondeur de la matière organique compacte [5]. Le DC est indicatif des périodes d’humidité du contenu organique sur de longues périodes étant donné que ce contenu est en profondeur. L’histogramme montre une tendance normale à la fin de la distribution ; ce à quoi on s’attendrait de façon intuitive. En effet, il fait du sens de penser qu’une forêt qui favorise les feux est généralement plus sèche. La boite à moustache confirme encore une fois cette idée et montre bien les quelques *outlier* au début de la distribution sont potentiellement des données aberrantes.

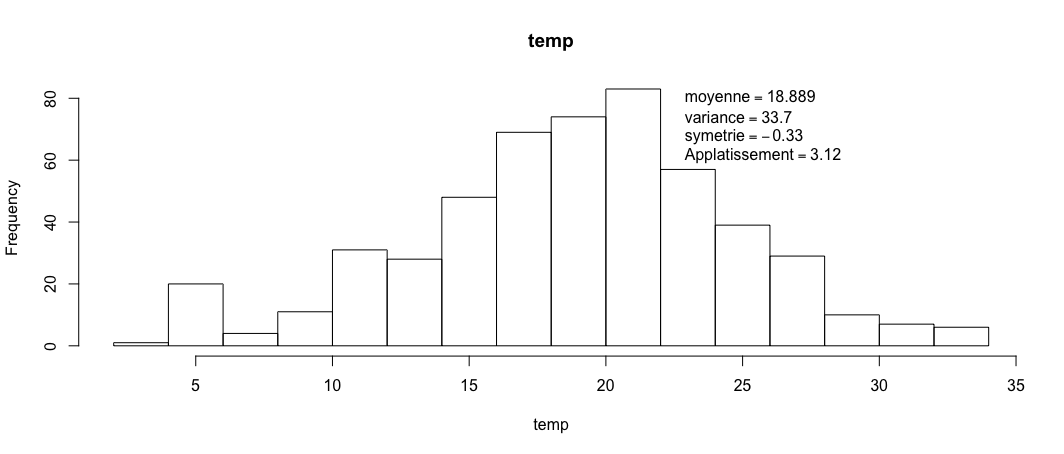
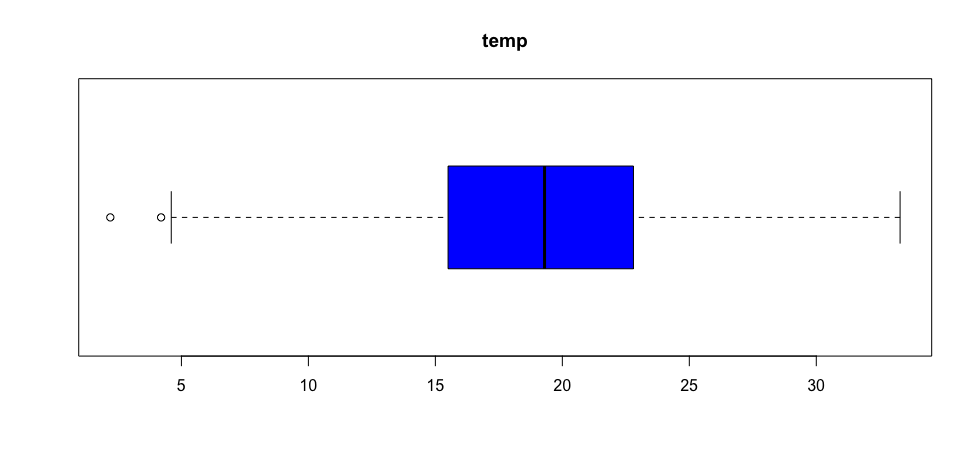




**Figure 4Histogramme et boite à moustache de la métrique *DC***

## **1.6) Indice de température**

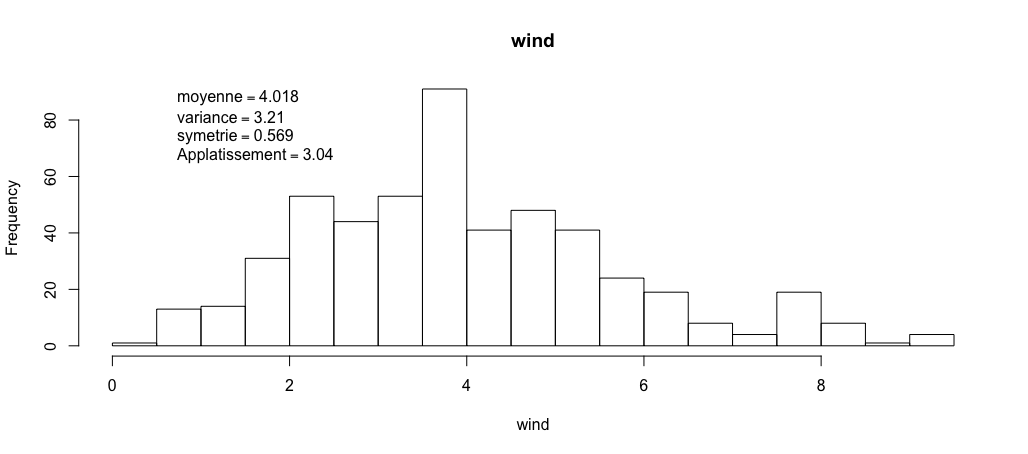
À l’œil, l’indice de température nous suggère une distribution normale. En effet, l’histogramme dénote des signes de distribution plutôt uniforme. Nous étudierons d’ailleurs cette hypothèse dans la prochaine section. Qui plus est, on remarque que le début de la distribution parait un peu curieux pour une distribution normale.

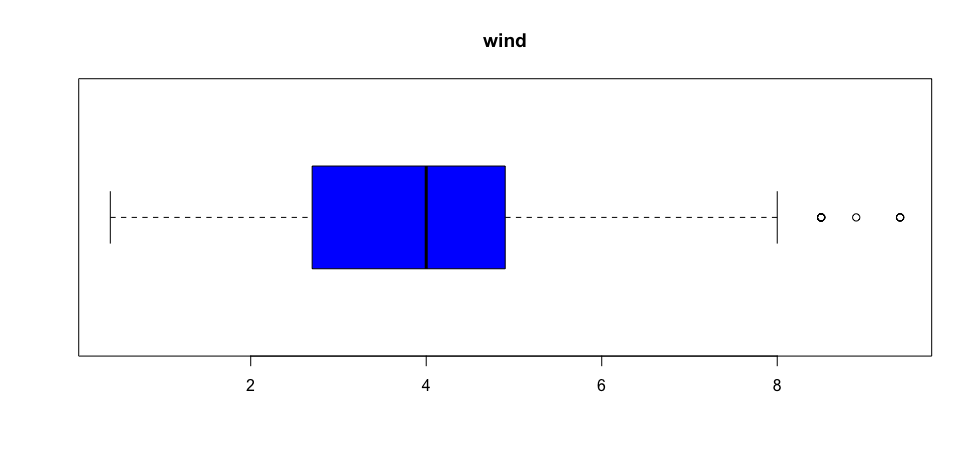


**Figure 5 Histogramme et boite à moustache de la température**

## **1.6) Indice du vent**

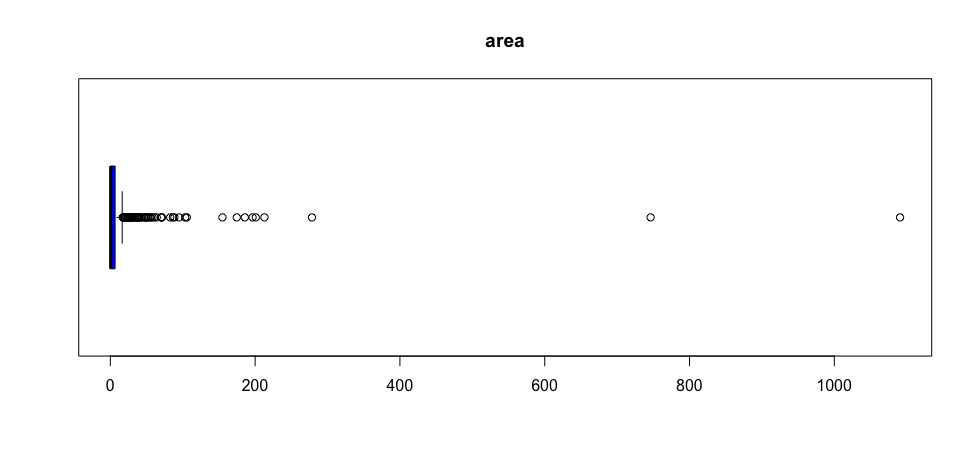
L’indice du vent dénote lui aussi certain signe de distribution normale. Ceci semble dire qu’il existe une vitesse du vent qui favorise les feux forestiers. Cependant, encore une fois, une analyse visuelle un peu plus rigoureuse fait remarquée que la fin de la distribution est une peu curieuse pour une distribution normale.





## **1.7) Surface incendié**

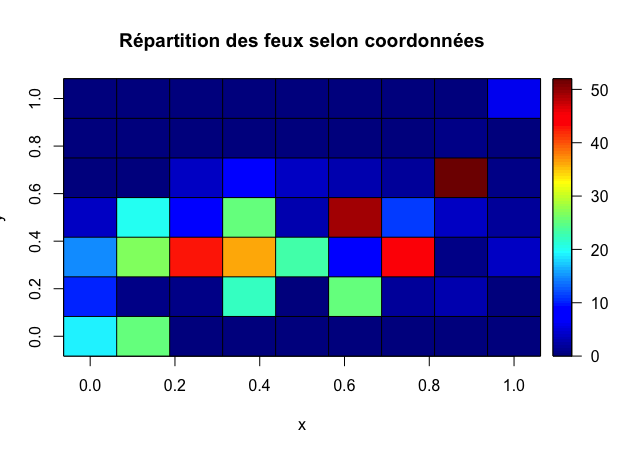
La surface incendiée n’est pas intéressante à visualiser à l’aide d’un histogramme. Cependant, on peut quand même utiliser la boite à moustache afin de bien comprendre que les feux qui attaquent la forêt de Montesinho sont généralement très petits. En effet, la majorité d’entre eux font en dessous de 100m carré [2].



**Figure 6 Boite à moustache représentant la surface incendiée**

## **1.8) Position géographique des feux**

Le jeux de données que nous avons utiliser comporter les coordonnées en X et Y de la géolocalisation de chaque instance de feux. La figure ci-dessous montre justement où l’on retrouve le plus grand nombre de d’incendie forestier selon ces coordonnées. On remarque qu’il y a clairement une tendance des endroits plus à risque.



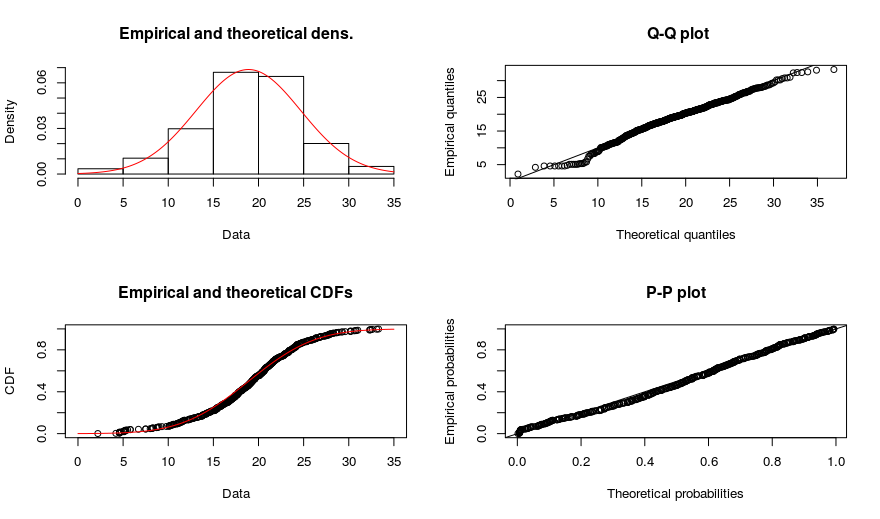
**Figure 7 Répartition des feux selon les coordonné X et Y (en nombre d’instance)**

## **2) Analyse de distributions sous-jacentes à certaine variables**

Dans le domaine des statistiques, il est souvent très pratique de savoir selon quelle type de distribution notre échantillon se comporte. En effet, cette information nous permet ensuite de pouvoir générer une fouler d’hypothèse, en particulier l’estimation de paramètres dans le cas d’une distribution normale. Nous expliquerons dans cette section de quelle manière nous avons procédé afin de d’accepter ou rejeter l’hypothèse que certaines de nos variables sous tendent une distribution normale. Les deux variables que nous avons testés sont le vent et la température qui caractérisent les feux forestiers du parc Montesinho.

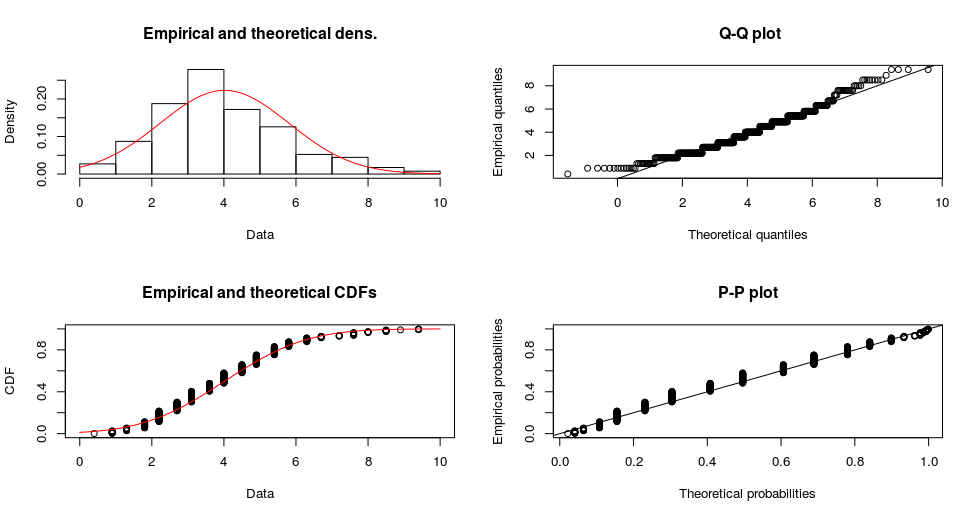
## **2.1) Analyse visuelle et sélection des variables**

Nous avons observé dans la section de l’analyse descriptive des données que deux des variables du jeu de données semblent, à vue d’œil, être distribué normalement, c’est-à-dire la température et le vent. Avant d’étudier cette hypothèse de façon plus rigoureuse, nous allons d’abord tenter de visualiser ce comportement normal à l’aide de différents outils de visualisation disponible dans R. Plus précisément, les figures qui que nous allons étudier présentent différente façon de visualiser si une distribution semble normale ou non. La première de ces quatre figures représente simplement la superposition d’une distribution normale de même moyenne et variance que notre distribution. La deuxième figure (dans le coin supérieur droit) représente le Q-Q plot. Cette figure représente les valeurs des quantiles équivalent théorique et empirique. Ainsi, la progression des valeurs quantile devrait être proportionnelle linéairement si les deux la même distribution. Les figures du bas sont basées sur le même principe. Celle de droite représente la fonction de densité empirique superposée à celle théorique. Enfin, la dernière représentation compare les probabilités théoriques à celle empiriques. Voici donc les résultat pour les variables « Temperature » et « Vent ».



**Figure 8 Visualisation de la potentiel normalité de la variable "Température"**

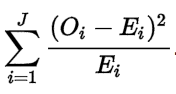




**Figure 9 Visualisation de la potentiel normalité de la variable "Vent"**

## **2.3) Analyse des hypothèse à l’aide du test du Khi-carré et de Shapiro–Wilk**

Les analyses précédentes nous donne de bonne raisons d’investiguer plus en profondeur la véracité de nos deux hypothèse nulle qui stipule que ces deux variables suivent une distribution normale. Afin de confirmer ou réfuter ces hypothèse avec un niveau de confiance de 95%, nous nous devons de faires passer à celles-ci des tests plus rigoureux. Le premier des tests que nous avons utilisé est celui du Khi-carré. Ce test consiste en une somation de l’erreur des observations par rapport aux valeurs attendues au carré, le tout divisé par la valeur espéré.

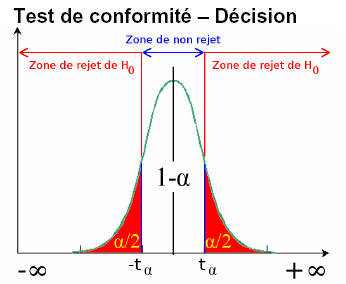


**Figure 10 Formule du test khi-carré**

Bien que ce tests peut être utilisé afin de vérifier une fouler de type d’hypothèse, nous l’utiliserons ici afin de vérifier si nos distributions suivent effectivement des distributions normales. Voici donc nos hypothèses :

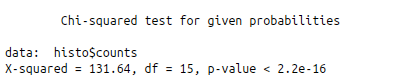
H0 : La distribution suit une loi normale.   
H1 : La distribution ne suit pas une loi normale.

Évidemment ce test existe dans le logiciel R. Nous avons donc utilisé la fonction « chisq.test ». Cependant, cette fonction nécessite un vecteur de probabilité théorique. Nous avons donc généré un vecteur de probabilité basé sur les mêmes classes que nos histogrammes. Ce vecteur de probabilité à été généré avec la fonction « dnorm » qui génère une fonction de densité normal de même moyenne et variance que notre distribution. Nous avons ensuite fais la différence de probabilité en entre les bornes supérieures de nos classes et leurs bornes inférieurs selon cette distributions. Puis, nous avons passé en paramètre à la fonction « chisq.test » nos données et le vecteur de probabilité. On se rappelle qu’ici l’objectif est de retenir l’hypothèse nulle qui stipule que nos distributions suivent effectivement une loi normale. Le schéma ici-bas montre bien ce qu’on doit obtenir pour retenir cette hypothèse.

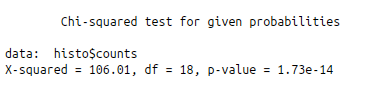


**Figure 11 Critère de non rejet de l'hypothèse nulle**

Et voici les output de R par concernant ce test :

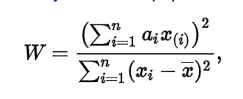


**Figure 12 khi-carré pour la température**

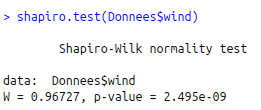


**Figure 13 Khi-carré pour le vent**

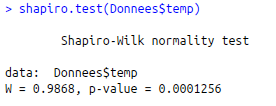
La p-value nous indique que la probabilité que ces deux distributions soient issues d’une loi normale est fortement improbable. Afin de confirmer ces résultats nous allons utiliser le test de Shapiro–Wilk. Ce test fait intervenir l’ordre statistique. Voici la formule:



Le résultat du test nous donnes des résultats très similaire à celui du khi-carré à savoir qu’il est fortement improbable que ces deux variables soient issue de distribution normale :



**Figure 14 Résultat shapiro test sur la variable vent**



**Figure 15 Résultat du test shapiro sur la variable vent**

À la lumière de ces tests on doit rejeter l’hypothèse nulle stipulant que ces distributions sont issue d’une distribution normal.

## **3. Régression**

Pour le test de régression, nous avons débutés par examiner visuellement les relations entre deux variables distinctes à l’aide de graphiques de nuages de points. Nous avons ensuite utilisé une boucle pour cycler sur toutes les paires possible et nous avons calculé l’indice de corrélation sur les éléments de la paire en question. Ces valeurs nous ont permis de filtrer toutes les variables qui n’avaient pas une corrélation supérieure à 0.4, à l'exception d’une paire que nous voulions examiner pour répondre aux questions ouvertes de la première partie. Après avoir soutiré les paires pertinentes, nous avons créés des modèles linéaires avec ces couples pour examiner s’ils admettaient effectivement une relation linéaire ou non. Nous avons examiné la régression linéaire sur deux couples de variables dans notre échantillon, le FFMC avec la température puis le ISI avec le DMC.

**3.1 Estimation de la régression**

Une fois qu’un couple potentiel à été identifié, nous utilisons la fonction lm() (*linear model*) pour créer un modèle linéaire avec deux variables. Cette fonction nous donne presque toutes les informations nécessaires pour estimer et déterminer la validité du modèle. La méthode summary(un.modele) nous affiche ces informations :

|  |
| --- |
| Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 82.89560 0.74672 111.01 <2e-16 \*\*\*  temp 0.41024 0.03779 10.86 <2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 4.985 on 515 degrees of freedom  Multiple R-squared: 0.1862, Adjusted R-squared: 0.1846  F-statistic: 117.8 on 1 and 515 DF, p-value: < 2.2e-16 |

Figure 18 : Exemple de sortie de la méthode summary()

Ce sont les résultats qui nous sont affiché avec un modèle linéaire construit avec les variables FFMC (indice de combustible léger) et température. Premièrement, nous avons directement les estimations des coefficients  et , respectivement 82.89560 et 0.41024. On obtient donc l’estimation de la régression. Nous avons les résultats du test de Student (*t value*) qui est utile pour calculer la valeur de p, mais nous avons déjà cette information sous la colonne *Pr(>|t|)*. La dernière variable intéressante est la valeur R carré. Ici il s'agit de *multiple R-squared*. Dans cet exemple nous obtenons la valeur 0.1862. R carré est la proportion de la variable y qui semble être attribuable à la relation linéaire avec x.[6, p.416] Donc ici on peut dire que 18.62% du comportement du FFMC est probablement attribuable à une relation linéaire avec la température. Cette proportion n’est pas énorme mais influence tout de même le FFMC.

## **3.2 Graphique de résidus**

Si effectivement il existe une relation linéaire entre les deux variables, les erreurs sur l’estimation de la droite de régression devraient eux aussi suivre cette tendance. Pour examiner si effectivement ce comportement est respecté, nous avons utilisé le graphique de résidus. [7, p.411] Ce nuage de points montre la dispersion des erreurs en fonction de la variable variables x. Logiquement, cette dispersion de points ne devrait suivre aucune droite, à l’exception de la pente nulle. Premièrement, nous avons effectué le test avec le couple de FFMC et température.

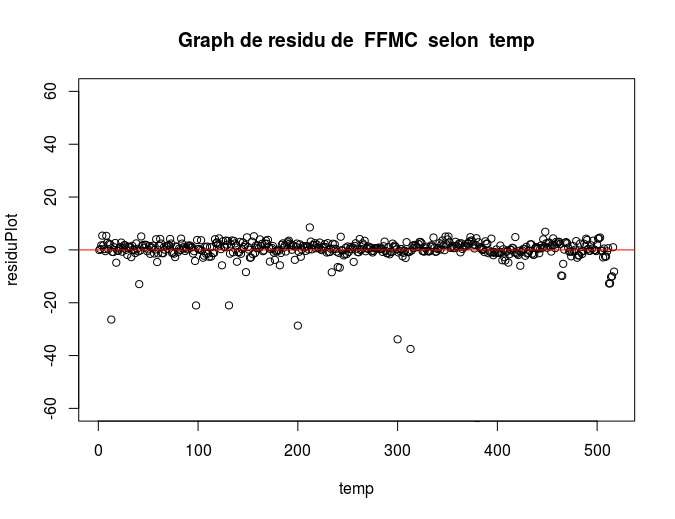


Figure 19 : Graphique de résidu de FFMC selon la température

On remarque ici qu’effectivement, la tendance des erreurs avec la droite de régression estimée suit toujours les mêmes valeurs, à l’exception de quelques données. La paire FFMC selon température passe ce test. Nous avons effectué le même test avec le couple ISI avec DMC.

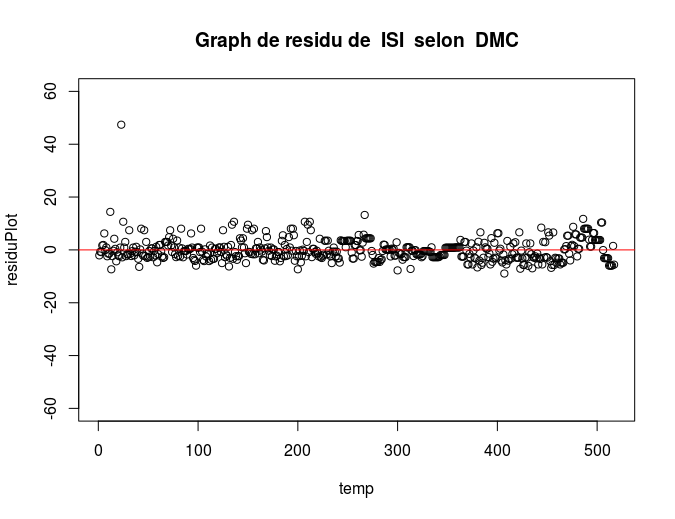


Figure 20 : Graphique de résidu de ISI selon DMC

Encore une fois, on remarque que les erreurs semblent toujours égales peu importe la valeur du DMC.

## **3.3 Graphiques quantile-quantile des erreurs**

La distribution des erreurs devrait approximer une distribution normale. [8, p.411] Une simple vérification s’en suit. Nous avons fait des graphiques quantile-quantile des erreurs avec l'estimation de la régression avec une distribution normale.

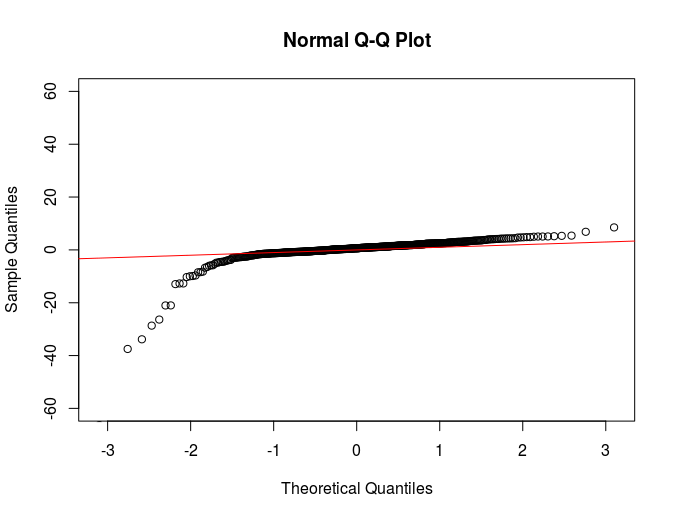


Figure 21 : Graphique quantile-quantile des erreurs de la paire FFMC et température

A l’exception des premières valeurs, la distribution de l’erreur créée par la droite estimée suit approximativement une normale. Cette estimation passe ce test.

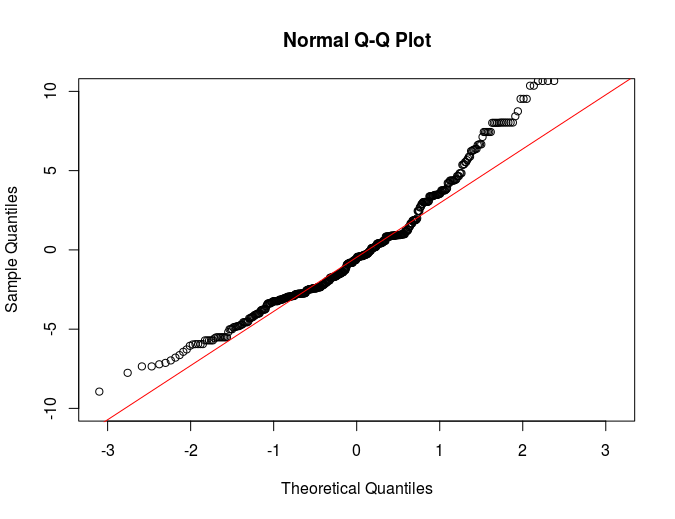


Figure 22 : Graphique quantile-quantile des erreurs de la paire ISI et DMC

La paire ISI et DMC ne correspond pas parfaitement la distribution normale mais est assez bonne. Cette paire passe donc ce test

## **3.4 Test d’hypothèse nulle**

Avec les informations recueillies jusqu’à maintenant, les deux couples sont de bons candidats de régression linéaire. Le dernier test à effectuer est le test d’hypothèse. [9, p.400] Comme hypothèse nulle nous avons :

: Il n’existe pas de relation linéaire entre les deux variables 

: Il existe une relation linéaire entre les deux variables

Pour examiner ces hypothèses, il suffit d’effectuer le test Student en comparant l’estimation de la régression obtenue avec l’échantillon avec celle de l’hypothèse nulle. On obtient cette équation :

Heureusement, la méthode summary() expliqué plus haut nous donne toutes les valeurs nécessaire afin de compléter ce calcul. Nous avons même déjà le résultat de cette équation dans la colonne.



Cette valeur nous donne l’information nécessaire pour obtenir la valeur de p, et encore une fois, on obtient cette valeur directement dans le tableau sous la colonne Pr(>|t|). Dans notre cas, nous avons. Plus cette valeur est petite, plus le comportement linéaire de la variable y est expliqué par la relation linéaire entre les deux variables, et non par chance. C’est-à-dire, il existe effectivement une relation linéaire qui agit sur y en fonction de x, respectivement FFMC et temp. La valeur obtenue, étant très petite, nous fait donc rejeter l’hypothèse nulle.

Pour le couple ISI et DMC, la méthode summary() retourne cette sortie :

|  |
| --- |
| Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 6.613285 0.382420 17.293 < 2e-16 \*\*\* DMC 0.021722 0.002987 7.271 1.33e-12 \*\*\* --- Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 4.346 on 515 degrees of freedom Multiple R-squared: 0.0931, Adjusted R-squared: 0.09134  F-statistic: 52.87 on 1 and 515 DF, p-value: 1.333e-12 |

Figure 23 : Exemple de sortie de la méthode summary() avec modèle ISI selon DMC

Ce modèle à passer tous les tests, graphiques et valeur p, mais sa valeur de R carré est basse, 0.0931. Alors nous pouvons expliquer que 9.31% du comportement linéaire du ISI est expliqué par sa relation linéaire avec le DMC. Cette proportion n’est pas très élevée, mais l’autre test confirme qu’une relation existe entre ces deux variables. Nous pouvons conclure qu’il existe une régression linéaire entre le ISI et le DMC.

## **3.5 Interprétation des résultats**

Il est intéressant d'utiliser la régression linéaire pour estimer certaines variables qui peuvent être plus difficile à examiner que d’autres. Nous avons prouvé une régression linéaire pour deux paire dans nos 13 variables, une avec le FFMC selon la température, une autre avec le ISI selon le DMC. Si on regarde le premier cas, la température est une mesure sensiblement facile à calculer et à prévoir. Avec ce modèle, il est possible de prévoir le comportement du FFMC. Le FFMC est l’indice combustible léger, une valeur numérique qui donne une indication d'inflammabilité du combustible léger. Ces deux variables ne sont pas directement liées, les combustibles léger est toute la broussaille et substances sèche qui est susceptible de prendre feu et il est faux de dire que plus la température est élevée, plus l’environnement est sec. Il est important de comprendre qu’avec la valeur du R carré obtenue, 0.1862, il ne s’agit pas d’une corrélation forte, mais nous pouvons conclure que plus la température est élevée, plus le FFMC tend à augmenter selon la courbe .

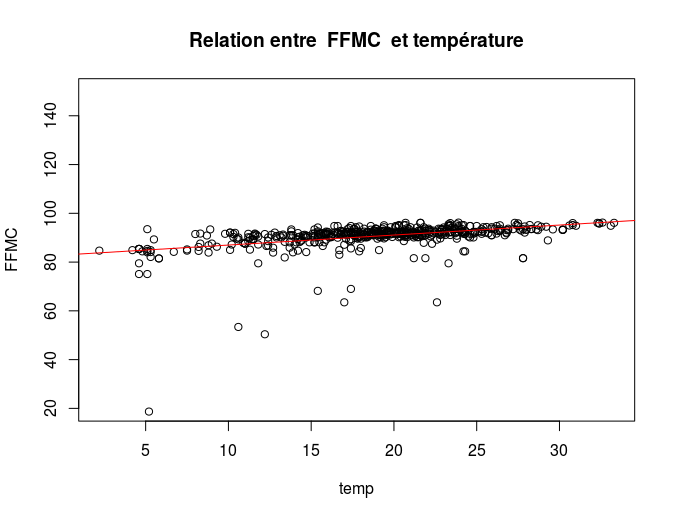


Figure 24 : Relation entre le FFMC et la température

Nous avons étudié le second modèle, ISI selon le DMC, afin de répondre à une question ouverte de la première partie. L’ISI est l’indice de propagation initial d’un feu. Nous voulions savoir si le DMC, un facteur de combustion des combustibles, avait une influence sur le ISI. Tous les tests nous indique qu’une régression linéaire entre ce deux variable existe, mais la valeur du R carré est basse, ce qui nous indique que la relation linéaire est faible. On peut donc conclure que oui il y a régression, mais ces deux variables n . La régression linéaire estimée est .

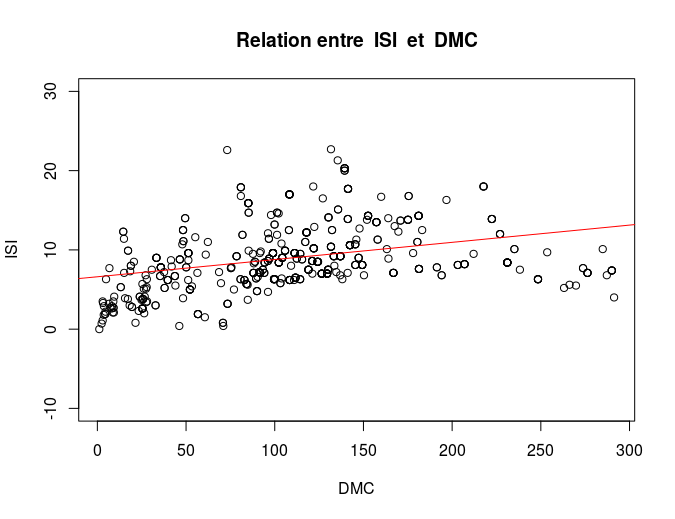


Figure 25 : Relation entre le ISI et le DMC

Conclusion

[4] W.W. Hines, D.C. Montgomery, D.M Goldsman, C. M.

Borror, Probability and Statistic pour ingénieurs, 3ème éd. Montréal, Canada, Chanelière-Éducation, 2017.

[5] W.J. De Groot, Interpreting the Canadian Forest Fire Weather Iinde (FWI) System[EN LIGNE]. Disponible: http://www.dnr.state.mi.us/WWW/FMD/WEATHER/Reference/FWI\_Background.pdf

[6] W.W. Hines, D.C. Montgomery, D.M Goldsman, C. M. Borror, Probability and Statistic pour ingénieurs, 3ème éd. Montréal, Canada, Chanelière-Éducation, 2017.

[7] W.W. Hines, D.C. Montgomery, D.M Goldsman, C. M. Borror, Probability and Statistic pour ingénieurs, 3ème éd. Montréal, Canada, Chanelière-Éducation, 2017.

[8] W.W. Hines, D.C. Montgomery, D.M Goldsman, C. M. Borror, Probability and Statistic pour ingénieurs, 3ème éd. Montréal, Canada, Chanelière-Éducation, 2017.

[9] W.W. Hines, D.C. Montgomery, D.M Goldsman, C. M. Borror, Probability and Statistic pour ingénieurs, 3ème éd. Montréal, Canada, Chanelière-Éducation, 2017.