

**AHMED KHALIL BOULAHIA**

**M1 BIG DATA**

**Ahmedboulahia@dauphine.tn**

[Projet Modèle Linéaire]

Projet véhicule

1. Objectifs du projet.
2. Description de la table de données.
3. Statistiques descriptives des données.
4. Détermination du meilleur modèle linéaire.
5. Objectif du projet :

On veut déterminer le meilleur modelé qui explique la consommation de carburant d’un véhicule en (km/litre) en fonction de :

-Nombre de cylindres du moteur,

-La cylindrée du véhicule,

-La puissance véhicule (nombre de chevaux),

-Le poids du véhicule (en livres),

-L'accélération du véhicule,

-L'âge et l'origine du véhicule.

1. Description de la table de données.

Pour expliquer la consommation des véhicules

conso ="Consommation du véhicule (km par litre)"

On a une table de données qui contient 392 véhicules et

10 variables explicatives dont 5 qualitatives :

nom ="Nom du véhicule"

nbcylindre ="Nombre de cylindres du moteurs"

poidcat ="Poids du véhicule (en livres)"

agecat ="Age du véhicule (en année)"

origine ="Origine du véhicule"

Et 5 variables quantitatives :

nbcylindre ="Nombre de cylindres du moteurs"

puiss ="La puissance véhicule (nombre de chevaux)"

poid ="Poids du véhicule (en livres)"

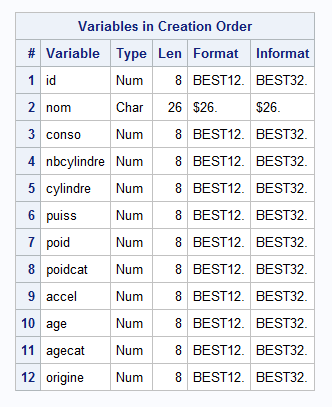
accel ="Accélération du véhicule"

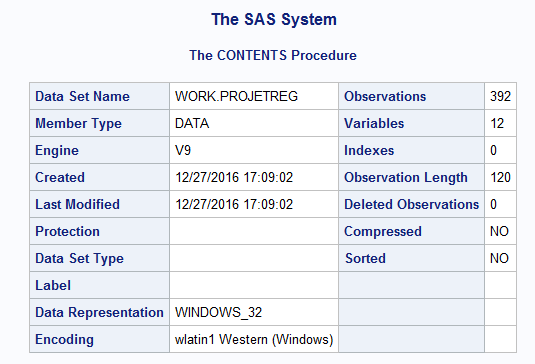
age ="Age du véhicule (en année)"

Pour une description totale du notre table de données on utilise la commande SAS :

**proc** **contents** data=projetreg varnum ;

**run** ;





Notre étape suivante est de crée des formats pour les variables qualitatives pour avoir une table de données bien lisible puis on affecte les formats on variables.

La création des formats se fait via cette commande Sas :

**proc** **format** library = PV;

value origine /\*Nom du format \*/

**1** = "Amerique du nord"

**2** = "Europe"

**3** = "Asie" ;

value f\_poidcat /\*Nom du format \*/

**1** = "[0-2200]"

**2** = "]2200-2600]"

**3** = "]2600-3100]"

**4** = "]3100-3800]"

**5** = ">3800";

value f\_nbcylindre /\*Nom du format \*/

**1** = "3 ou 4"

**2** = "5 ou 6"

**3** = "7 ou 8";

value f\_agecat /\*Nom du format \*/

**1** = "[1-4]"

**2** = "[5-9]"

**3** = "[10-13]";

**run**;

L’affectaion des formats aux variables se fait avec la commande Sas suivante :

**data** projetreg ;

set projetreg ;

format origine origine. ;

format poidcat f\_poidcat. ;

format nbcylindre f\_nbcylindre. ;

format agecat f\_agecat. ;

**run**;

On peut aussi labéliser nos variables avec cette commande Sas :

**data** PV.projetreg ;

set PV.projetreg ;

/\* labeliser les variables \*/

label id = "Identifiant du véhicule" ;

label nom ="Nom du véhicule";

label conso ="Consommation du véhicule (km par litre)";

label nbcylindre ="Nombre de cylindres du moteurs";

label cylindre ="La cylindrée du véhicule";

label puiss ="La puissance véhicule (nombre de chevaux)";

label poid ="Poids du véhicule (en livres)";

label poidcat ="Poids du véhicule (en livres)";

label accel ="Accélération du véhicule";

label age ="Age du véhicule (en année)";

label agecat ="Age du véhicule (en année)";

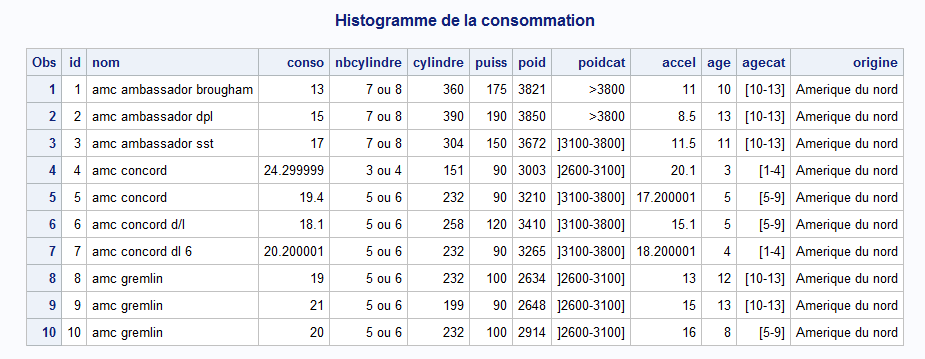
label origine ="Origine du véhicule";

**run**;

Finalement pour 10 observations on a cette table de données :

**proc** **print** data=PV.projetreg (obs=**10**) ;

**run**;

****

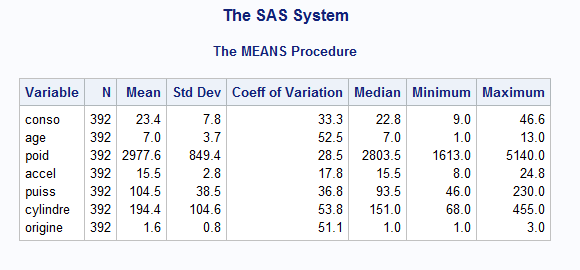
1. Statistiques descriptives des données.

Pour l’Analyse descriptive des variables quantitative on utilise la commande Sas suivante :

**proc** **means** data= PV.projetreg n mean std cv median min max maxdec=**1**;

var conso age poid accel puiss cylindre ;

**run**;

Ca va nous afficher un tableau qui contient la moyenne, le coefficient de variance, la médiane, le maximum et le minimum de chaque variable quantitative.

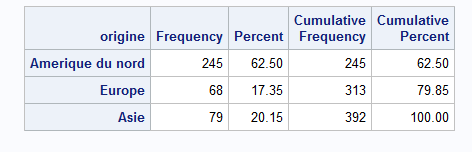
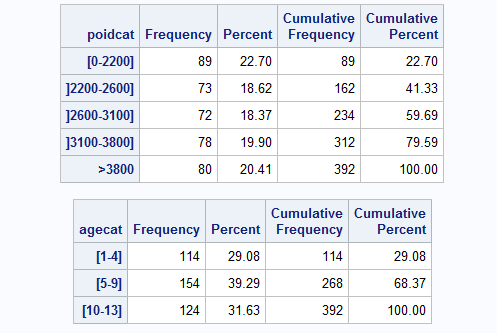
Et pour l’Analyse descriptive des variables qualitative on utilise la commande Sas suivante :

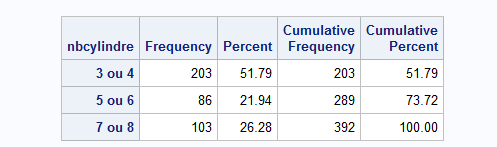
**proc** **freq** data=PV.projetreg;

table poidcat agecat nbcylindre origine ;

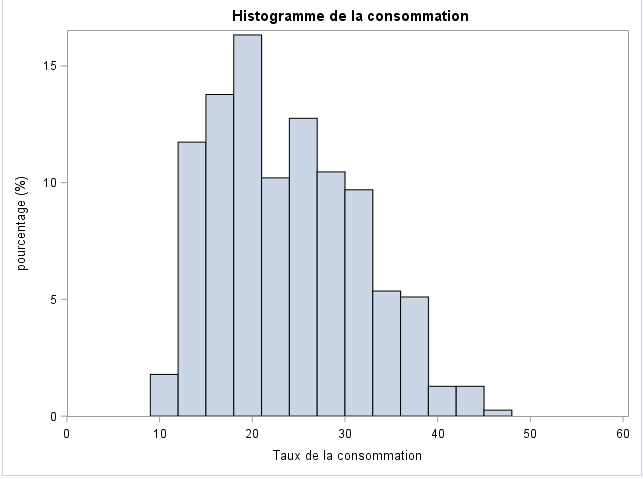
**run** ;

Cette commande Sas affiche une table pour chaque variable qualitative qui contient : la fréquence, le pourcentage, les fréquences cumulées et les pourcentages cumulés, les tables affichées en sortie du Sas sont les suivantes :





Pour des statistiques descriptives plus détailles on va utiliser les Histogrammes pour les variables qualitatives :



**proc** **sgplot** data=PV.projetreg;

histogram conso / scale = percent showbins ;

xaxis label="Taux de la consommation" min=**0** max=**60**;

yaxis label="pourcentage (%)" ;

title " Histogramme de la consommation";

**run** ;

On remarque qu’environ 15% des voitures consomment

un litre pas 20 km,

Et on remarque aussi qu’il y’a des voitures consomment

Un litre pas 45 km ce que je trouve un peu de la réalité

Je trouve que c’est intéressant aussi de voir le taux de consommation du carburant par origine de la voiture et par l’âge de la voiture aussi, Pour affiche l’Histogramme du taux de consommation en fonction de l’origine de la voiture on utilise cette commande Sas :

**proc** **sgpanel** data=PV.projetreg;

panelby origine / columns=**3** layout=panel uniscale=all ;

histogram conso / scale=percent ;

colaxis label ="Taux de consommation" min=**0** max=**60** ;

rowaxis label="pourcentage (%)";

**run**;

Et pour l’Histogramme du taux de consommation du carburant en fonction de l’âge de la voiture on utilise cette commande Sas :

**proc** **sgpanel** data=PV.projetreg;

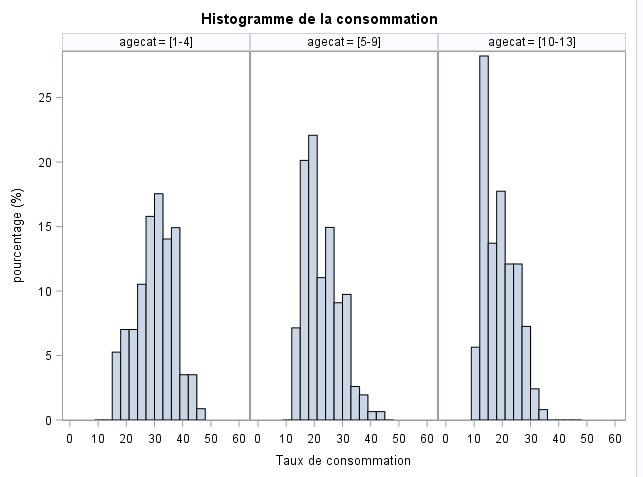
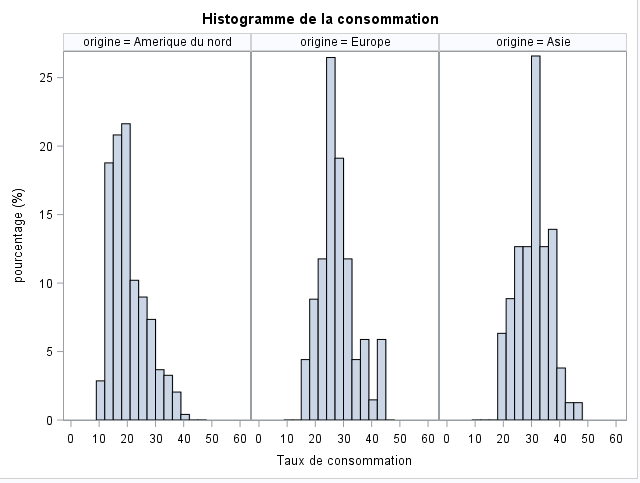
panelby agecat / columns=**3** layout=panel uniscale=all ;

histogram conso / scale=percent ;

colaxis label ="Taux de consommation" min=**0** max=**60** ;

rowaxis label="pourcentage (%)";

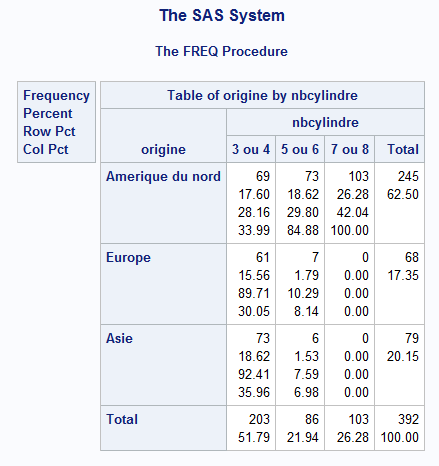
**run**;

Dans la sortie du Sas on récupère ces deux histogrammes :

La première interprétation qui saute au yeux est plus que 25% des voitures dont l’âge entre 10 et 13 ans consomment une litre chaque 15 km c’est-à-dire qu’ils représentent une grande source de pollution.

On remarque aussi que presque 60% des voitures d’origine aérique du nord consomment une litre du carburant par 20km contrairement au voitures d’origines asiatique qui consomment environ une litre par 35km.

On peut faire la même étude pour les variables qualitatives, par exemple on veut voir quelle est l’origine des voitures qui ont un nombre de cylindre élevé, cette étude se fait via cette commande Sas :



/\* croisement des 2 var qualitatives\*/

**proc** **freq** data=PV.projetreg;

table origine \* nbcylindre ;

**run** ;

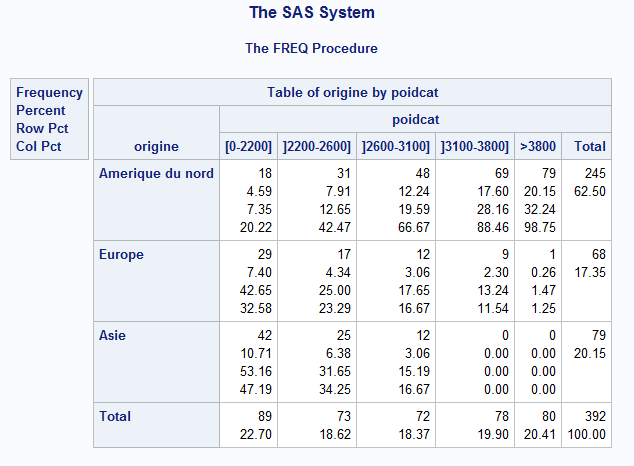
En regardant la table affichée en sortie Sas on trouve

Que seule les voitures de l’Amerique du nord contient

7 ou 8 cylindre ce qui justifie bien le taux de consommation

élevé du carburant.

On peut aussi étudier la relation entre l’origine des voitures et leurs poids via cette commande :



**proc** **freq** data=PV.projetreg;

table origine \* poidcat ;

**run** ;

Une seule voiture européenne dépasse 3800 kg,

Aucune voiture asiatique ne dépasse 3100 kg

par contre on a 79 voitures de l’Amérique du nord qui de passe 3800 kg.

4-Determination du meilleur modèle linéaire.

Dans cette partie du projet on va déterminer le meilleur modelé qui explique le taux de consommation du carburant en fonction des divers caractéristiques.

Comme première étape on va étudier la corrélation entre les variables quantitatives ( age, cylindre, poid, accel et puiss), en utilisant les coefficients de corrélation de **Pearson** et de **Spearman.**

Coefficient de corrélation de **Pearson** : évalue la relation linéaire entre deux variables continues. Une relation est dite linéaire lorsqu'une modification de l'une des variables est associée à une modification proportionnelle de l'autre variable.

Coefficient de corrélation de **Spearman**: évalue la relation monotone entre deux variables continues. Dans une relation monotone, les variables ont tendance à changer ensemble, mais pas forcément à une vitesse constante.

Le calcule des matrices de corrélation de **Pearson** et de **Spearman** entre les variables quantitatives se fait via cette commande Sas :

**proc** **corr** data = PV.projetreg pearson spearman

plots = matrix(histogram) ;

var cylindre accel puiss poid age ;

**run** ;

On obtient les résultats suivants :

|  |  |
| --- | --- |
| *5 Variables:* | cylindre accel puiss poid age |

| *Simple Statistics* | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| *Variable* | *N* | *Mean* | *Std Dev* | *Median* | *Minimum* | *Maximum* |
| *cylindre* | 392 | 194.41199 | 104.64400 | 151.00000 | 68.00000 | 455.00000 |
| *accel* | 392 | 15.54133 | 2.75886 | 15.50000 | 8.00000 | 24.80000 |
| *puiss* | 392 | 104.46939 | 38.49116 | 93.50000 | 46.00000 | 230.00000 |
| *poid* | 392 | 2978 | 849.40256 | 2804 | 1613 | 5140 |
| *age* | 392 | 7.02041 | 3.68374 | 7.00000 | 1.00000 | 13.00000 |

| *Pearson Correlation Coefficients, N = 392 Prob > |r| under H0: Rho=0* | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | *cylindre* | *accel* | *puiss* | *poid* | *age* |
| *cylindre* | 1.00000 | -0.54380 <.0001 | 0.89726 <.0001 | 0.93299 <.0001 | 0.36986 <.0001 |
| *accel* | -0.54380 <.0001 | 1.00000 | -0.68920 <.0001 | -0.41684 <.0001 | -0.29032 <.0001 |
| *puiss* | 0.89726 <.0001 | -0.68920 <.0001 | 1.00000 | 0.86454 <.0001 | 0.41636 <.0001 |
| *poid* | 0.93299 <.0001 | -0.41684 <.0001 | 0.86454 <.0001 | 1.00000 | 0.30912 <.0001 |
| *age* | 0.36986 <.0001 | -0.29032 <.0001 | 0.41636 <.0001 | 0.30912 <.0001 | 1.00000 |

| *Spearman Correlation Coefficients, N = 392 Prob > |r| under H0: Rho=0* | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | *cylindre* | *accel* | *puiss* | *poid* | *age* |
| *cylindre* | 1.00000 | -0.49940 <.0001 | 0.87617 <.0001 | 0.94563 <.0001 | 0.30658 <.0001 |
| *accel* | -0.49940 <.0001 | 1.00000 | -0.65814 <.0001 | -0.40511 <.0001 | -0.27831 <.0001 |
| *puiss* | 0.87617 <.0001 | -0.65814 <.0001 | 1.00000 | 0.87882 <.0001 | 0.38950 <.0001 |
| *poid* | 0.94563 <.0001 | -0.40511 <.0001 | 0.87882 <.0001 | 1.00000 | 0.28098 <.0001 |
| *age* | 0.30658 <.0001 | -0.27831 <.0001 | 0.38950 <.0001 | 0.28098 <.0001 | 1.00000 |

INTERPRETATION DES RESULTATS :

On remarque que tous les coefficients de corrélation de Pearson et de Spearman ont un p-valeur <0.0001 , donc on conclut que les variables ( age , puiss , cylindre , poid et accel ) sont 2 a 2 décorrèles.

Comme deuxième étape on va calculer les coefficients de corrélation de **Pearson** et de **Spearman** du taux de consommation du carburant avec les variables quantitatives ( age, cylindre, poid, accel et puiss) cet étude se fait a travers cette commande Sas :

**proc** **corr** data = PV.projetreg pearson spearman

plots = matrix(histogram) ;

var cylindre accel puiss poid age ;

with conso ;

**run** ;

On obtient les résultats suivants :

|  |  |
| --- | --- |
| *1 With Variables:* | conso |
| *5 Variables:* | cylindre accel puiss poid age |

| *Simple Statistics* | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| *Variable* | *N* | *Mean* | *Std Dev* | *Median* | *Minimum* | *Maximum* |
| *conso* | 392 | 23.44592 | 7.80501 | 22.75000 | 9.00000 | 46.60000 |
| *cylindre* | 392 | 194.41199 | 104.64400 | 151.00000 | 68.00000 | 455.00000 |
| *accel* | 392 | 15.54133 | 2.75886 | 15.50000 | 8.00000 | 24.80000 |
| *puiss* | 392 | 104.46939 | 38.49116 | 93.50000 | 46.00000 | 230.00000 |
| *poid* | 392 | 2978 | 849.40256 | 2804 | 1613 | 5140 |
| *age* | 392 | 7.02041 | 3.68374 | 7.00000 | 1.00000 | 13.00000 |

| *Pearson Correlation Coefficients, N = 392 Prob > |r| under H0: Rho=0* | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | *cylindre* | *accel* | *puiss* | *poid* | *age* |
| *conso* | -0.80513 <.0001 | 0.42333 <.0001 | -0.77843 <.0001 | -0.83224 <.0001 | -0.58054 <.0001 |

| *Spearman Correlation Coefficients, N = 392 Prob > |r| under H0: Rho=0* | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | *cylindre* | *accel* | *puiss* | *poid* | *age* |
| *conso* | -0.85523 <.0001 | 0.44154 <.0001 | -0.85362 <.0001 | -0.87559 <.0001 | -0.57484 <.0001 |



INTERPRETATION DES RESULTATS :

En regardant les graphs Scatter Plot Matrix on remarque que la courbe de corrélation entre la conso et poid a la même tendance que la courbe de corrélation entre conso et puiss.

On décide alors d’éliminer la variable poid de notre modèle .

DETERMINATION DU MEILLEUR MODELE

Détermination du meilleur modèle qui explique le taux de consommation du carburant en fonction de divers caractéristique se fait via cette commande Sas :

**proc** **reg** data = PV.projetreg ;

model conso = cylindre puiss poid age accel / clb;

**run** ;

**quit**;

On obtient :

|  |  |
| --- | --- |
| *Number of Observations Read* | 392 |
| *Number of Observations Used* | 392 |

| *Analysis of Variance* | | | | | |
| --- | --- | --- | --- | --- | --- |
| *Source* | *DF* | *Sum of Squares* | *Mean Square* | *F Value* | *Pr > F* |
| *Model* | 5 | 19264 | 3852.80085 | 326.50 | <.0001 |
| *Error* | 386 | 4554.98881 | 11.80049 |  |  |
| *Corrected Total* | 391 | 23819 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| *Root MSE* | 3.43518 | *R-Square* | 0.8088 |
| *Dependent Mean* | 23.44592 | *Adj R-Sq* | 0.8063 |
| *Coeff Var* | 14.65152 |  |  |

| *Parameter Estimates* | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Variable* | *DF* | *Parameter Estimate* | *Standard Error* | *t Value* | *Pr > |t|* | *95% Confidence Limits* | |
| *Intercept* | 1 | 47.15626 | 1.99114 | 23.68 | <.0001 | 43.24142 | 51.07110 |
| *cylindre* | 1 | 0.00278 | 0.00546 | 0.51 | 0.6108 | -0.00796 | 0.01352 |
| *puiss* | 1 | 0.00102 | 0.01376 | 0.07 | 0.9410 | -0.02604 | 0.02808 |
| *poid* | 1 | -0.00687 | 0.00066525 | -10.33 | <.0001 | -0.00818 | -0.00557 |
| *age* | 1 | -0.75412 | 0.05261 | -14.33 | <.0001 | -0.85756 | -0.65067 |
| *accel* | 1 | 0.09032 | 0.10191 | 0.89 | 0.3760 | -0.11004 | 0.29069 |

TEST DU MODELE GLOBALE DE LA REGRESSION.

On remarque que la p-valeur du test du modèle globale est <0.0001

Donc on rejette H0 du notre test :

H0: β1, …., βp = 0

contre

H1: Il existe au moins un βj parmi β1, …, βp non égal à 0.

(p=5, n=392)

Et on considère le modèle avec au moins un covariable non nulle .

TEST DE NULLITE DES VARABLES.

On remarque que la p-valeur du test de nullité des variable cylindre , puiss et accel a une p-valeur <0.0001

On conclue alors que ces trois variables n’ont pas un effet sur notre modèle

Donc on accepte H0 du notre teste et on considère ces trois variables nulles

H0: βk = 0

contre

H1: βk diffèrent de 0 (1<= k <=5)

Modelé Finale :

|  |  |
| --- | --- |
| *Number of Observations Read* | 392 |
| *Number of Observations Used* | 392 |

| *Analysis of Variance* | | | | | |
| --- | --- | --- | --- | --- | --- |
| *Source* | *DF* | *Sum of Squares* | *Mean Square* | *F Value* | *Pr > F* |
| *Model* | 2 | 19250 | 9625.02057 | 819.47 | <.0001 |
| *Error* | 389 | 4568.95193 | 11.74538 |  |  |
| *Corrected Total* | 391 | 23819 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| *Root MSE* | 3.42715 | *R-Square* | 0.8082 |
| *Dependent Mean* | 23.44592 | *Adj R-Sq* | 0.8072 |
| *Coeff Var* | 14.61727 |  |  |

| *Parameter Estimates* | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Variable* | *DF* | *Parameter Estimate* | *Standard Error* | *t Value* | *Pr > |t|* | *95% Confidence Limits* | |
| *Intercept* | 1 | 48.51016 | 0.64927 | 74.71 | <.0001 | 47.23364 | 49.78668 |
| *poid* | 1 | -0.00663 | 0.00021456 | -30.91 | <.0001 | -0.00705 | -0.00621 |
| *age* | 1 | -0.75732 | 0.04947 | -15.31 | <.0001 | -0.85459 | -0.66005 |





Finalement : d’après notre base de données on conclue que seul L’âge et le poids influes sur le taux de consommation de carburant .