

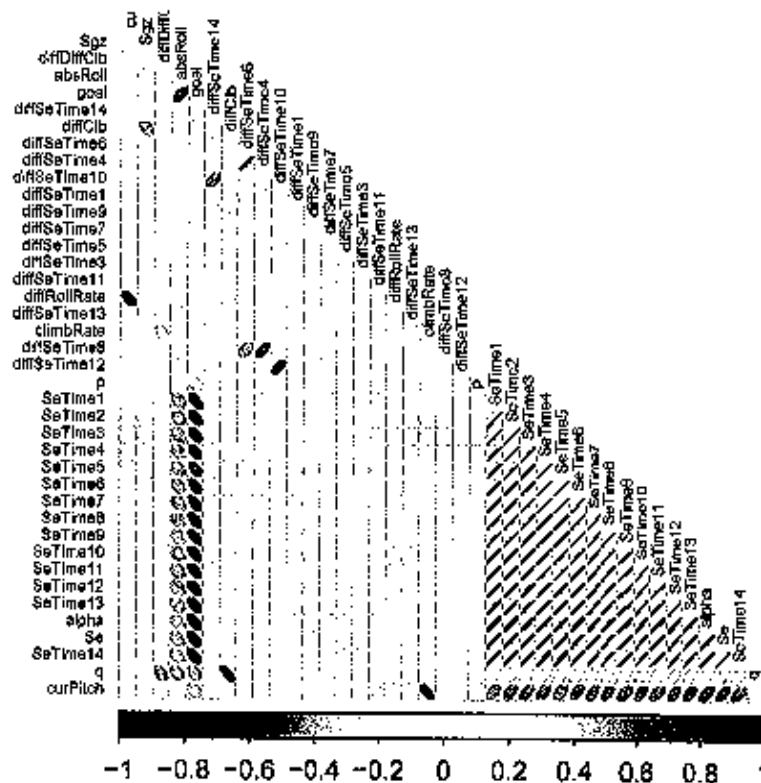


Nous allons travailler pour ce projet sur des données décrivant différents paramètres d'un avion F16. Le but est de décrire et prédire l'influence des ailerons sur le mouvement de l'appareil.

Le jeu de données comprend 41 variables, et 13750 observations. Parmi ces variables, ClimbRate décrit la vitesse verticale de l'avion. Sgz correspond à son accélération verticale. P désigne la vitesse angulaire de roulis, et Q celle de tangage. CurPitch donne l'ordre de profondeur. curRoll et absRoll permettent de décrire le roulis. diffClib et diffDiffClib décrivent le taux de montée, et diffRollRate celui de roulis. Les SeTime et diffSeTime sont des variables temporelles, décrivant des temps d'acquisition et de réponse. Enfin, alpha correspond à l'angle d'incidence de l'aéronef, et la variable cible est goal.

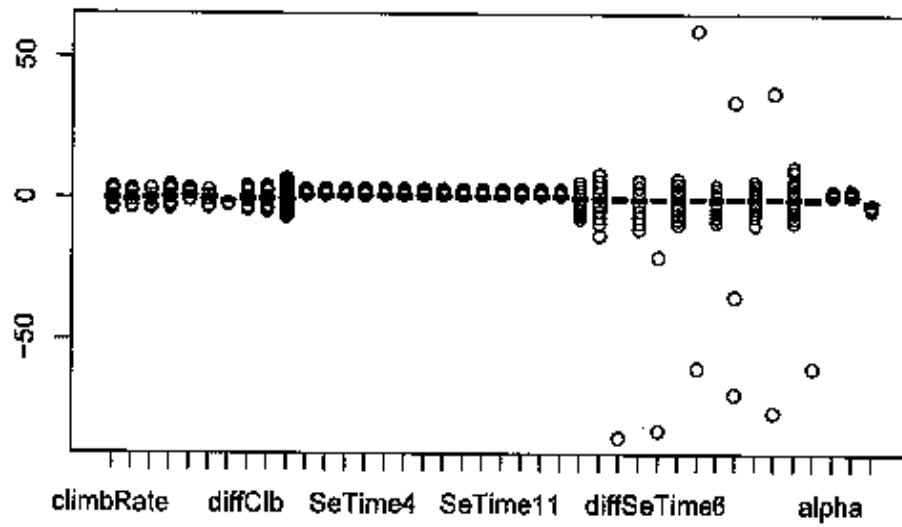
Si on calcule les corrélations entre les différentes variables (cf graphique ci-dessous), on s'aperçoit d'une part que les SeTime sont très corrélés entre eux, et d'autre part que goal est très corrélé aux SeTimes. On constate aussi une forte corrélation entre diffSeTime4 et diffSeTime5.

Corrélations entre les variables

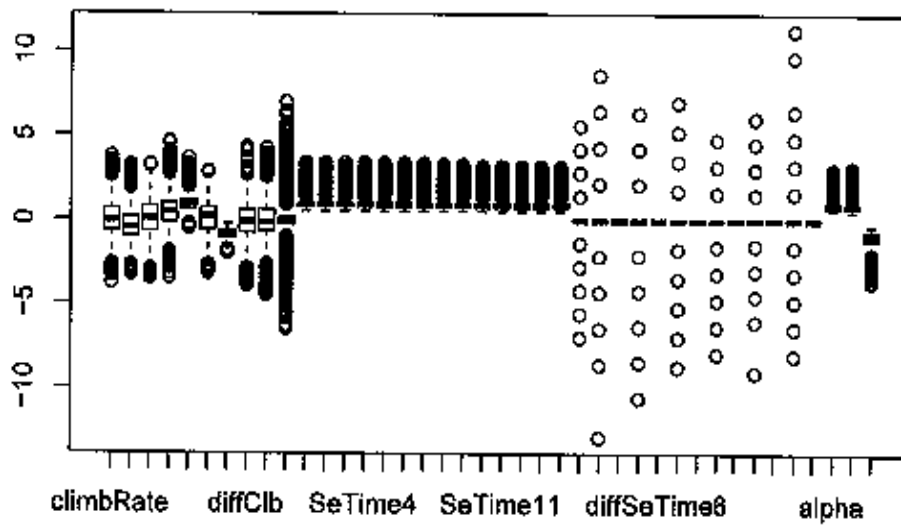


Il peut être intéressant de faire un boxplot des valeurs de chaque variable, après avoir mis à l'échelle les données, afin de voir si il y a des valeurs qui semblent aberrantes. C'est en effet le cas: certains diffSeTime semblent avoir des valeurs bien trop écartées de la normale. Après écrêtage des données, on obtient un résultat plus satisfaisant.

Valeurs des variables



Valeurs des variables après écrêtage



MRR Projet : Etude des ailerons



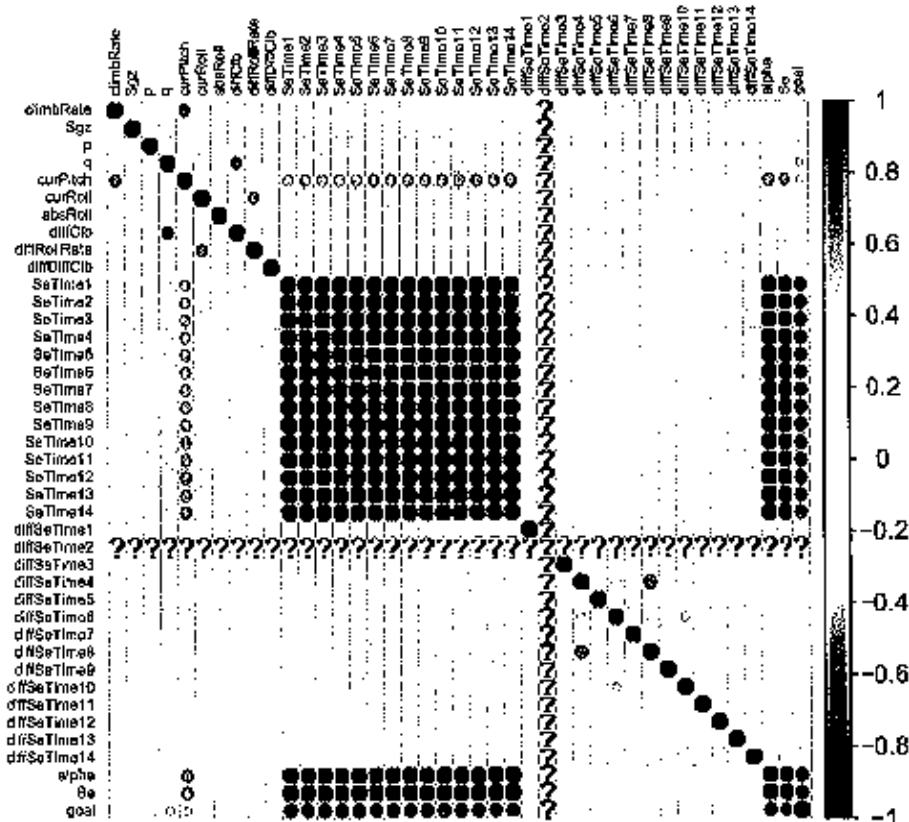
Nature générale du problème : Nous allons étudier un ensemble de données représentant les observations effectuées sur les attributs d'un avion F16. L'objectif mené par Rui Camacho est de prédire les actions que le pilote doit effectuer sur les ailerons de l'avion, dans le but de l'aider et de le guider dans ses commandes dans le suivi d'un plan de vol. Pour cela on dispose de plus variables de position de l'avion.

Les variables explicatives : Nous disposons de 40 attributs pour expliquer le comportement de l'aileron de l'avion, ces covariables (et notre goal) sont toutes continues. On peut citer entre autre climbRate (vitesse verticale de l'avion), Sgz (accélération verticale), curPitch (rotation latérale), curRoll et absRoll (respectivement la rotation longitudinale et sa valeur absolue), diffCib (différentiel de vitesse verticale), SeTime(X) (temps que met le pilote pour effectuer l'action X), diffSeTime(X) (délai d'acquisition des actions du pilote) ou encore Se (maximum de déviation autorisée). La variable cible (goal) représente l'écart entre la situation actuelle de l'avion et le plan de vol à suivre.

Pour étudier plus en détails les données, nous allons calculer la matrice de corrélation des données pour étudier si elles sont reliées entre elles. Voici la matrice de corrélation :

```
corrplot(cor(tab,use="pairwise.complete.obs"), tl.cex = 0.5)
```

```
## Warning in cor(tab, use = "pairwise.complete.obs"): l'écart type est nulle
```

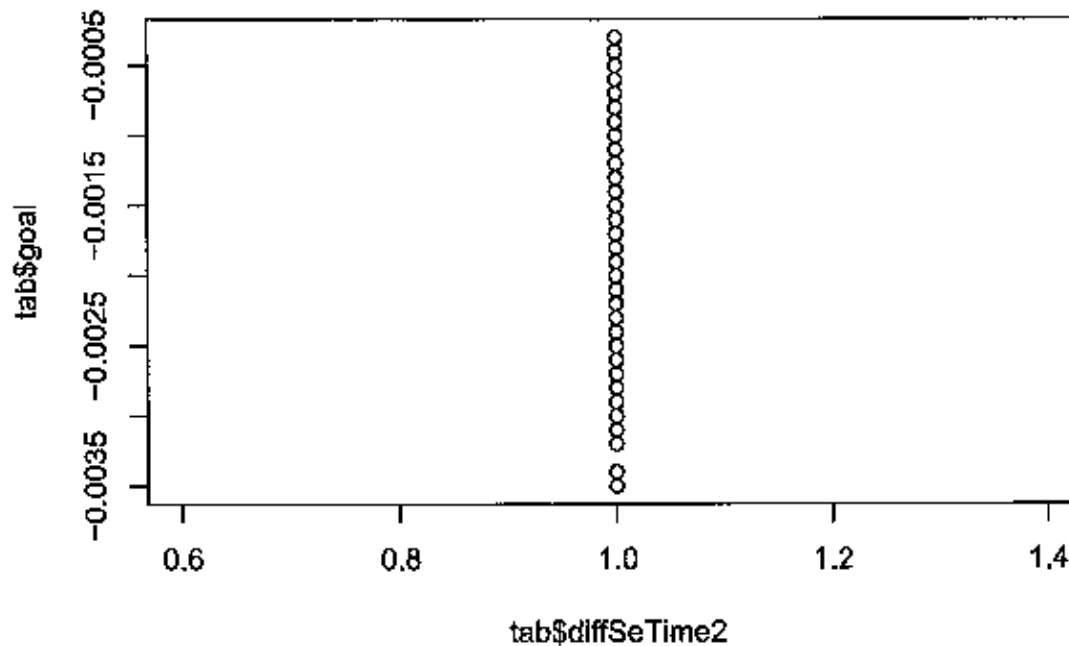


On peut voir sur la matrice de corrélation que les variables SeTime sont très fortement corrélées entre elles.

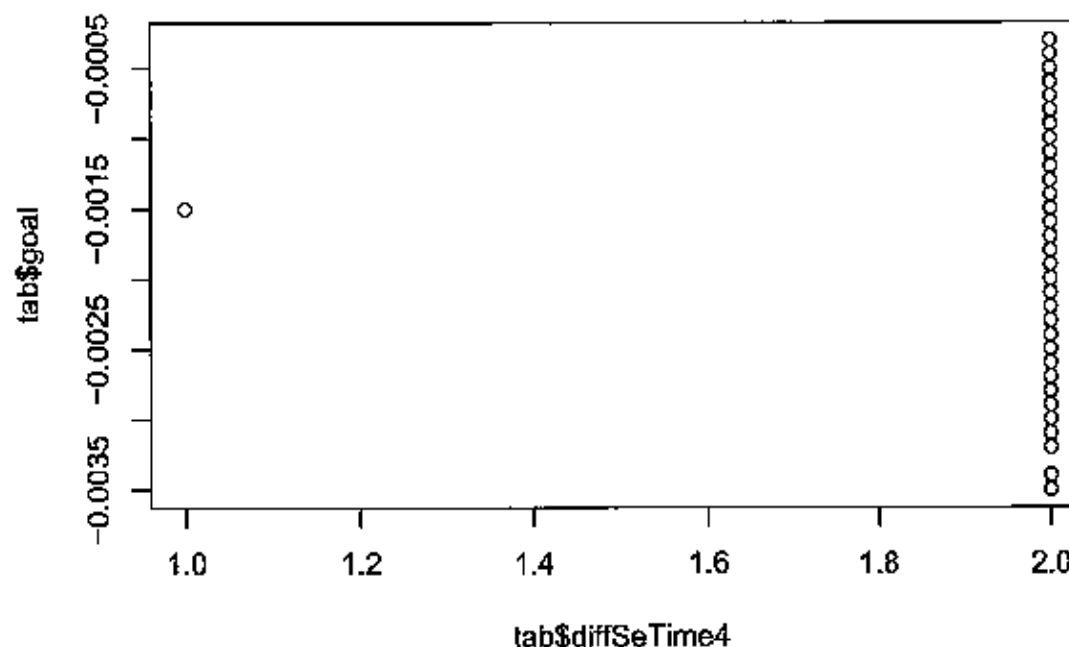
On peut voir également que la co-variable Se est fortement corrélée avec toutes les variables SeTime.

De plus on peut voir que les co-variables de type diffSeTime semblent assez décorrélées de la variable cible goal comme ces graphiques semblent l'indiquer

```
plot(tab$diffSeTime2,tab$goal,main=names(tab$diffSeTime2))
```



```
plot(tab$diffSeTime4,tab$goal,main=names(tab$diffSeTime4))
```



Maintenant que nous avons une meilleure vue d'ensemble des données de notre problème nous allons pouvoir plus efficacement déterminer quelles co-variables conserver ou éliminer de notre modèle.

Projet de MRR : Ailerons

Le but de l'étude

Rui Camacho a constitué un jeu de données sur la dynamique des ailerons sur les avions (des F-16) de 7154 échantillons concernant 41 variables. Aucune donnée n'est manquante.

L'objectif est d'expliquer la variable 'goal' en fonction des 40 autres variables. Il y a 14 variables pour SeTime, de même pour diffSeTime.

Les variables

Variables	Signification
climbRate	vitesse verticale
Sgz	accélération verticale
p	vitesse angulaire du roulis
q	vitesse angulaire du tangage
curpitch	décrit le tangage
curRoll	décrit le roulis
absRoll	décrit le roulis
diffCib	décrit le taux de montée
diffDiffRate	décrit le taux de montée
diffRollRate	décrit le taux de roulis (rotation longitudinale)
Les SeTime	durée
Les diffSeTime	durée
alpha	angle d'attaque de l'avion
goal	variable cible : contrôle de l'avion par les ailerons

Figure 1. Tableau d'explication des variables du jeu de données de Rui Camacho

On peut analyser rapidement les données pour savoir si des variables sont corrélées. Dans le schéma ci-dessous, plus une case est colorée (bleue ou rouge), plus les 2 variables concernées sont corrélées.

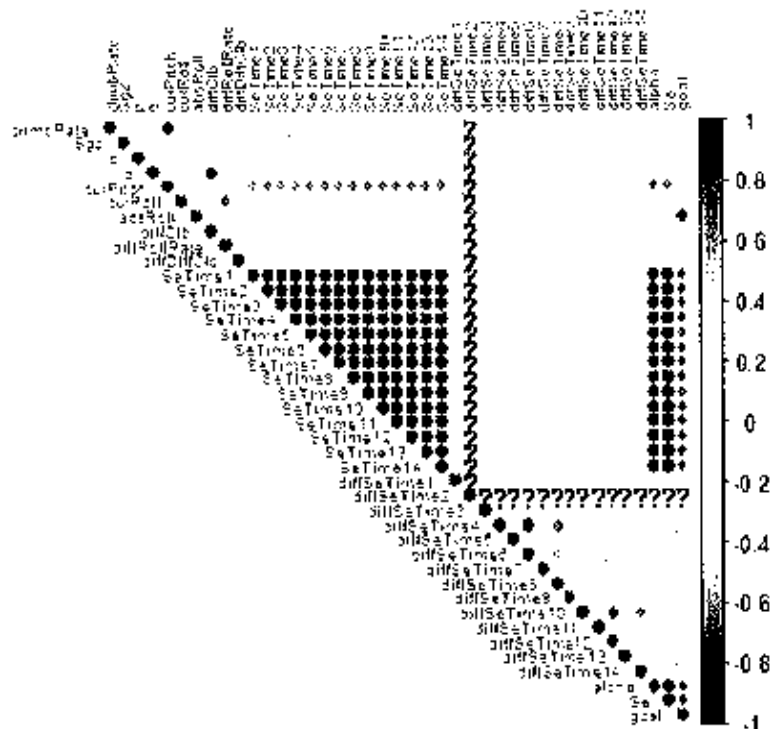
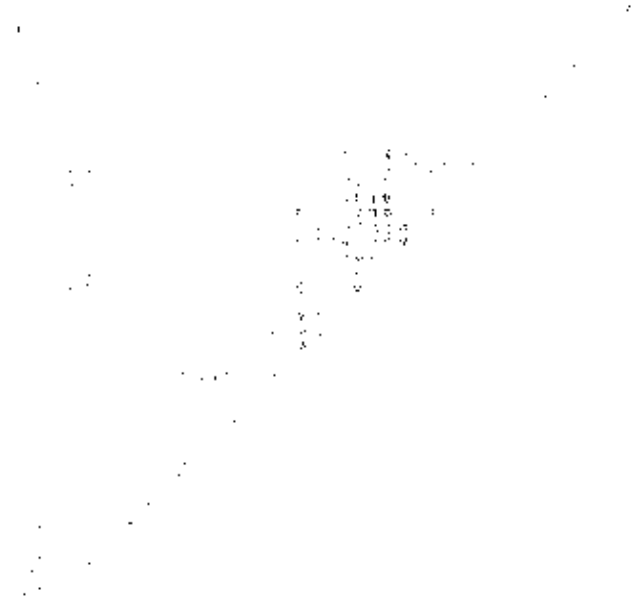


Figure 2. Matrice de corrélation des variables du jeu de données

On remarque aussi que certaines variables sont faiblement liées aux autres (case peu colorées sur toute une ligne) comme diffSeTime2, diffSeTime6, diffSeTime8, diffSeTime12, diffSeTime14. Il est possible que la suite de l'étude élimine ces variables si sélection il y a.





F16 Aircraft flying control

This data set addresses a control problem, namely flying a F16 aircraft. The attributes describe the status of the aeroplane, while the goal is to predict the control action on the ailerons of the aircraft. In fact, the controller is made by an algorithm which takes variables from the airplane status and measures. In this algorithm, there is a step where a prediction model has to estimate a certain goal variable which will be used in the controller model. So our target in this project is to create this model and determine the most important variables.

The Goal

The goal is a continuous variable which corresponds to a technical variable used in the controller. It's also our target variable in the model. Here we can see some information on its distribution :

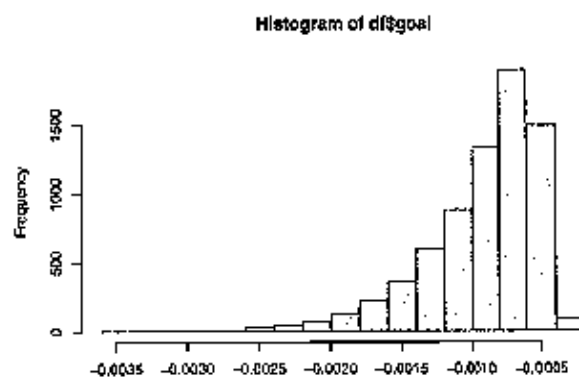


fig 1.1 : Histogram of the goal variable which shows the repartition of the distribution it.

We can see that the distribution follows a sort of exponential law. There are no absurd variables so we can keep all rows of our data set.

The Variables

All our variables are continuous variables. They come from different measures on the F-16. Some of these variables are quite explicit for a non-expert (like the Climb Rate) others are not easily understandable (like p , q or Sgz). The best we can say on these non-explicit variables is that they are standard technical measures on an airplane fly. There are another type of variable named `SeTime` and `diffSeTime` which correspond to each one to 14 columns; those variables correspond to a sort of delay and differential on measurement response.

Now we will look at the correlation matrix between the variables. As we have too many parameters, we will separate the matrices in three parts.

```
## Warning in cor(df_dse): 1'écart type est nul
```

SeTime / Goal Cor PLOT

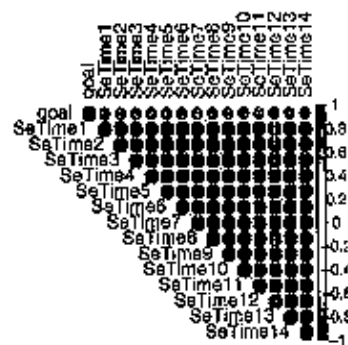


fig 2.1 : Correlations between goal and all SeTime variables

diffSeTime / Goal Cor PLOT

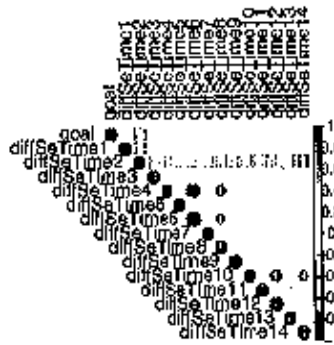


fig 2.2 : Correlations between goal and all diffSeTime variables

We can see on fig 2.1 a huge correlation between SeTimes variable and negative correlation between SeTimes and Goal. We could already think that those variables will be important in the model. Then on fig 2.1, we see that there are not so much correlation between diffSeTime and Goal variables. We will see later if this variable is significant for our model but yet, it seems not. We remark too that on this plot, the column diffSeTime2 seems to contain many none value, but after looking at it, we found no None values and all values are numeric (type : "double"). So we can keep this column in the data set as it's still a numeric values but if it causes problem later we will remove it. In fact, as all values of the column is 0, it will not impact the model.

Now we plot the correlation matrix of the others datas, goal, SeTime1 and diffSeTime1.

Goal / Other Variables Cor PLOT

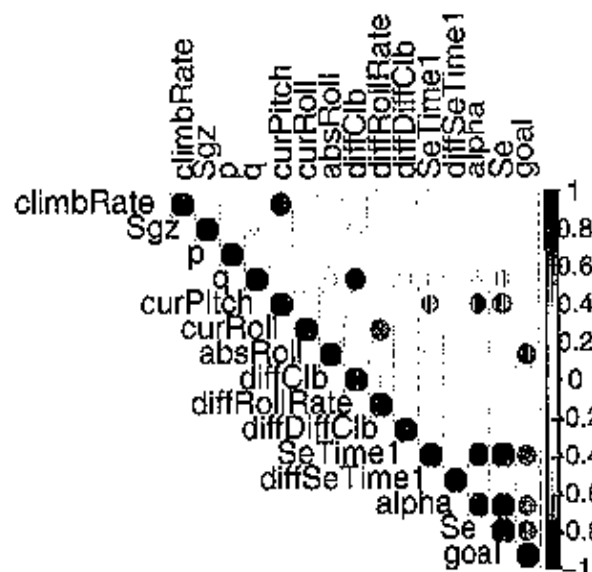


fig 2.3 : Correlations between goal, SeTime1, diffSeTime1 and others variables

On this plot, we see that some variables seems to be more significant than others for our model (like absRoll, alpha, Se). While some variable seems to be insignificant (like diffRollRate or diffDiffCib) but we still have to study a model to see if it have or not a real impact.

Analysis and description of data – Ailerons

1. The general nature of the problem

- Title of the database: Ailerons
- Sources: Experiments of Rui Camacho (rcamacho@garfield.fe.up.pt)

This data set addresses a control problem, namely flying a F16 aircraft. The attributes describe the status of the aeroplane, while the goal is to predict the control action on the ailerons of the aircraft.

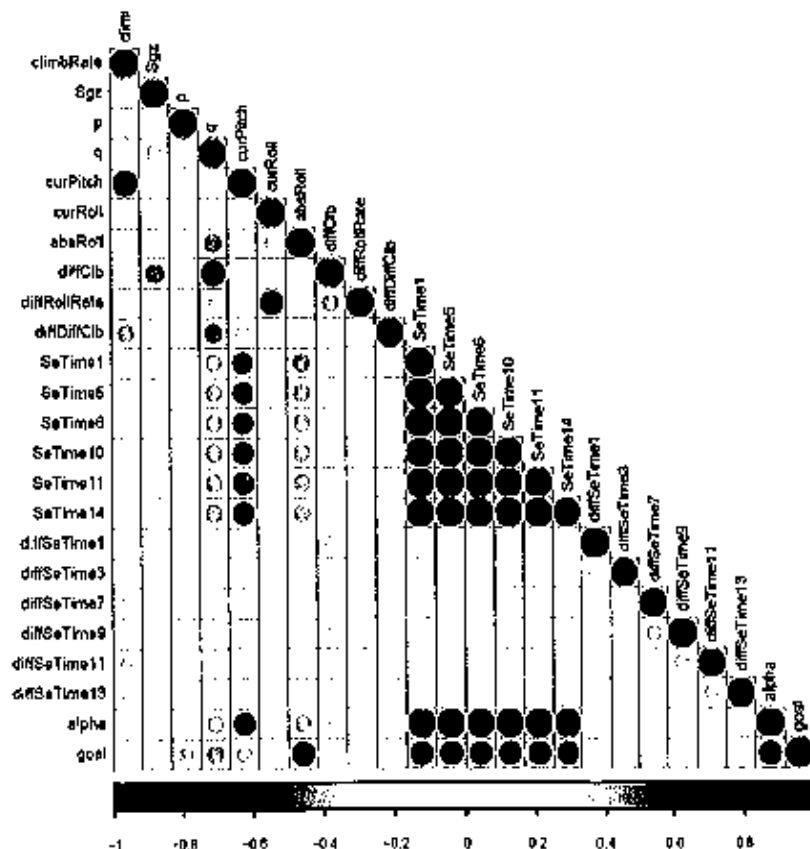
2. The variables available

The data set is composed of 7154 instances of 41 variables. All the variables of the set are continuous. The first 40 variables are the co-variables which are available to predict the target variable called "goal".

The covariates are: the climb rate (climbRate), the vertical acceleration (Sgz), the roll rate (p), the pitch rate (q), the current pitch and roll angles (curPitch and curRoll), absRoll, the derivatives of the climb rate and the roll rate (diffCib and diffRollRate), the double derivative of the climb rate (diffDiffCib), the angle of attack (alpha) and Se. The 28 remaining co-variables are called SeTime1 and diffSeTime1 with i in [1,14].

Among these variables, the co-variate diffSeTime2 only contains values equal to 0, so we decided to drop it.

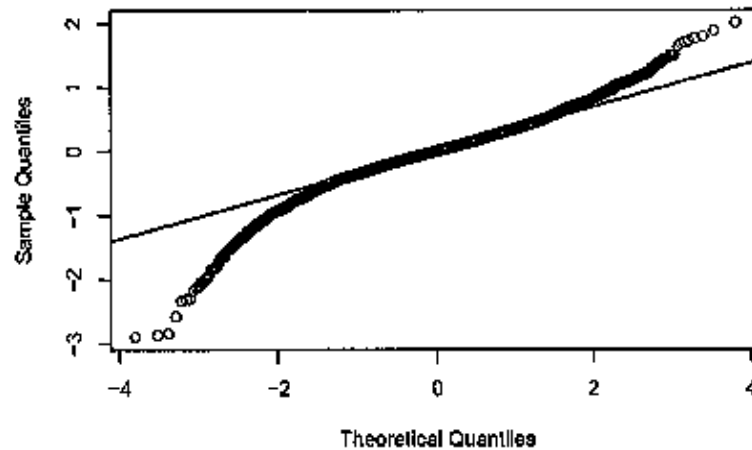
Graph 1: correlation matrix between a selection of co-variables and the target variable



On this graph, we can see that the most correlated co-variables to the goal variable are alpha, absRoll, p, q, curPitch and the SetTime variables. Furthermore, the SeTime variables are strongly correlated to each other which induce X to be degenerated. Indeed, the rank of X is 32 while it has 39 columns.

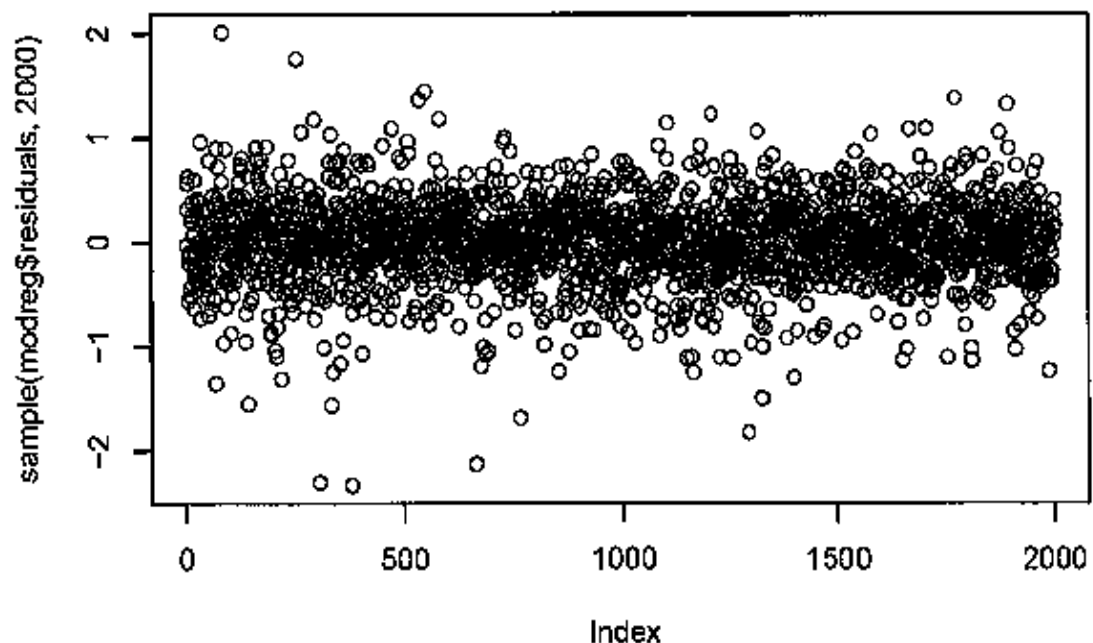
3. The target variable

Graph 2: Q-Q plot



Because of our target variable being continuous, it cannot be a logistic regression problem. So the link between target variable and the explanatory variables is linear.

Graph 3: Student residual graph (2000 points)



We can find that all the points are roughly symmetrically distributed around the x-axis, so it will be reasonable to use linear regression.