Université Paris Nanterre
Master 1 Iséfar
TD de Modèles de régression

# Feuille de TD n° 1 : Régression linéaire simple

**N.B.** Cette feuille d'exercices est inspirée de

`https://online.stat.psu.edu/stat501/lesson/1/1.1`

## 1 Relation déterministe ou aléatoire

1. Il est connu que, si $F$ désigne une température en dégrés Farenheit et $C$ désigne la même température en degrés Celcius, alors

$$F = \frac{9}{5} \times C + 32. \tag{1}$$

Reproduire grâce à **R** la Figure 1, qui représente graphiquement la relation (1).

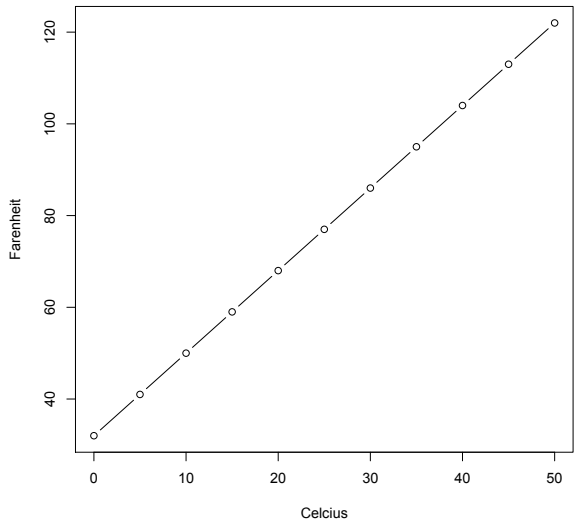


Figure 1 – Degrés Farenheit en fonction de degrés Celcius.

2. Le jeu de données **skincancer**, disponible sur la plateforme `Cours en Ligne`, fournit pour 48 états des Etats-Unis le nombre, pour 10 millions d'habitants, de décès dus au cancer de la peau, ainsi que le degré de latitude nord du centre de l'état. Les données ont été recueillies en 1950.

   (a) Importer le jeu de données dans **R** grâce à la commande

   ```
   skin <- read.table("skincancer.txt",header=T)
   ```

   Dans le fichier disponible, les mesures de mortalité sont stockées dans le vecteur `Mort` tandis que les latitudes sont stockées dans le vecteur Lat.

   (b) Vérifier que les données ont été correctement importée avec les commandes

   ```
   summary(skin)
   head(skin)
   ```

   (c) Reproduire dans `R` la FIGURE 2, qui représente graphiquement le jeu de données ainsi que la droite des moindres carrés.

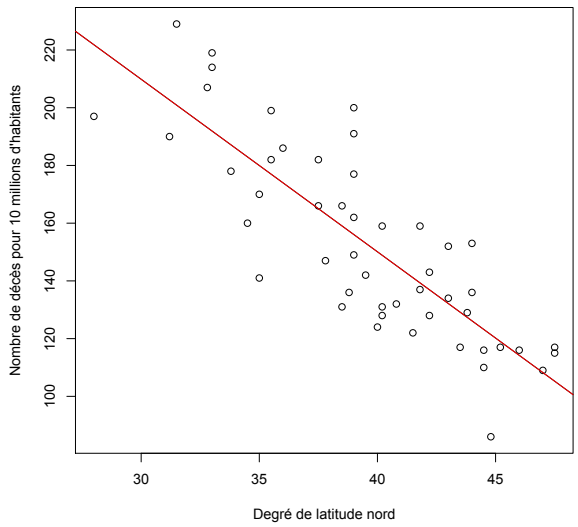


FIGURE 2 – Mortalité en fonction de Latitude

Pour cela, on attachera les données à l'espace de travail, on construira le nuage de points, puis on ajoutera la droite des moindres carrés grâce à la commande

```
abline(lm(Mort~Lat),col=2)
```

   (d) Exécuter la commande

   ```
   summary(lm(Mort~Lat))
   ```

   Observer que la sortie **R** indique notamment

| Coefficients: | |
|---|---|
| Estimate | |
| (Intercept) | 389.1894 |
| Lat | -5.9776 |

En déduire que l'équation de la droite rouge en FIGURE 2 est

$$y = \hat{\mu} + \hat{a}x. \tag{2}$$

où $\hat{\mu} \approx 389.19$ et $\hat{a} \approx -5.98$.

(e) Que penser de l'hypothèse qu'un individu résidant à une plus grande latitude nord aux Etats-Unis (et donc peut-être moins exposé aux rayons du soleil), a un risque plus faible de décéder d'un cancer de la peau qu'un individu résidant à une plus faible latitude.

(f) Peut-on considérer qu'il existe une relation linéaire déterministe entre le nombre de décès dus au cancer de la peau et la latitude, ou qu'il s'agit plutôt d'une relation linéaire aléatoire, c'est-à-dire présentant une tendance linéaire et une dispersion ?

## 2 La droite des moindres carrés

Nous travaillons encore dans cette section sur le jeu de données **skincancer**, disponible sur la plateforme `Cours en Ligne`. Nous notons $x_i$ la latitude nord du centre du $i$-ème état et $y_i$ la mesure de mortalité correspondante, pour tout $i$. Nous notons de plus $\widehat{y}_i = \hat{\mu} + \hat{a}x_i$ pour tout $i$, où $\hat{\mu}$ et $\hat{a}$ sont tirés de l'équation de la droite des moindres carrés (2).

1. Sachant que la mesure de latitude de l'Alabama est de 33, calculer la prédiction $\widehat{y}_i$ correspondante grâce à la commande

   ```
   predict(lm(Mort~Lat) , data.frame(Lat=33))
   ```

   Calculer alors l'erreur de prédiction $y_i - \widehat{y}_i$ en Alabama.

2. Calculer la somme de carrés résiduels

   $$\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2,$$

   et la comparer à $\sum_{i=1}^{n} (y_i - (\mu + ax_i))^2$ pour différentes valeurs de $(a, \mu)$. Quel résultat général peut-on énoncer ?

3. Montrer que $\hat{\mu} = \bar{y} - \hat{a}\bar{x}$ et

   $$\hat{a} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

   où $n$ désigne le nombre d'états étudiés, $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$ et $\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$.

4. Montrer que la droite des moindres carrés passe par le point de coordonnées $(\bar{x}, \bar{y})$.
5. Vérifier que si $x_i - \bar{x}$ était du même signe que $y_i - \bar{y}$ pour tout $i$, alors la pente de la droite des moindres carrés serait positive. Interpréter.
6. Retrouver la somme de carrés résiduels avec la commande

   `anova(lm(Mort~Lat))`

# 3 Le modèle de régression linéaire simple

Nous travaillons encore dans cette section sur le jeu de données **skincancer**, et utilisons les mêmes notations que dans la Section 2. Nous supposons que $y_1, \ldots, y_n$ sont les réalisations de variables aléatoires indépendantes $Y_1, \ldots, Y_n$ telles qu'il existe des réels $a_0$ et $\mu_0$ vérifiant $E(Y_i) = \mu_0 + a_0 x_i$ pour tout $i$. Nous supposons

$$Y_i = \mu_0 + a_0 x_i + \varepsilon_i, \ i = 1, \ldots, n \tag{3}$$

où $\varepsilon_1, \ldots, \varepsilon_n$ sont des variables aléatoires non observées, centrées, indépendantes entre elles, et de même loi gaussienne centrée. Nous notons $\sigma^2$ leur variance commune (inconnue). Nous utilisons la même notation $\hat{\mu}, \hat{a}$ pour les estimations et pour les estimateurs correspondants.

1. Vérifier qu'il s'agit d'un modèle linéaire, et écrire ce modèle sous forme matricielle. Quelle est la dimension de l'espace vectoriel associé ?
2. Rappeler la définition de l'estimateur du maximum de vraisemblance de $\sigma$, puis celle de l'estimateur corrigé. Vérifier que l'on obtient sa valeur avec la commande

   `sqrt(sum((predict(lm(Mort~Lat))-Mort)^2/(length(Mort)-2)))`
3. Soit $\hat{\varepsilon}$ le vecteur de composantes $Y_i - (\hat{\mu} + \hat{a} x_i)$, pour tout $i \in \{1, \ldots, n\}$. Donner la loi de $\hat{\varepsilon}$ puis celle de $\|\hat{\varepsilon}\|^2$. Expliquer pourquoi une plus grande variance $\sigma^2$ conduit à une plus grande erreur de prédiction.
4. Soient

$$\begin{aligned} SSR &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \quad \text{(Regression sum of squares)} \\ SSE &= \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \quad \text{(Error sum of squares)} \\ SST &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \text{(Total sum of squares)} \\ R^2 &= SSR/SST. \end{aligned}$$

   (a) En utilisant le Théorème de Pythagore, justifier que $SST = SSR + SSE$.

   (b) Calculer $SSR$, $SSE$ et $SST$ en utilisant la fonction `predict`. Retrouver les résultats avec la commande

`anova(lm(Mort~Lat))`

(c) Calculer $R^2$.

(d) Calculer le coefficient de corrélation linéaire entre les $x_i$ et les $y_i$ :

$$\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}.$$

# 4 Simulations

1. Simuler des observations $y_i = 1 + 2x_i + \varepsilon_i$, $i = 1, \ldots, 20$, où on a tiré les $x_i$ uniformément dans $[-2, 2]$ et les $\varepsilon_i$ selon la loi gaussienne centrée de variance $1/4$, toutes ces variables étant indépendantes entre elles (fixer la graine du générateur aléatoire à 123).
2. Représenter graphiquement les données.
3. Réaliser une régression linéaire simple et enregistrer les résultats dans `out`.
4. Valider le modèle.
5. L'effet `x` est-il significatif ? Comment tester que la pente de la droite de régression est positive ? Qu'elle est égale à 2 ?
6. Que produisent les commandes suivantes ? On aura préalablement ordonné les données selon `x`.

```
p.conf=predict(out,interval="confidence")
p.pred=predict(out,interval="prediction")
colnames(p.conf)=list("fit","lconf","uconf")
colnames(p.pred)=list("fit","lpred","upred")
intervalles=cbind(p.conf,p.pred[,-1])
```

7. Ajouter un titre approprié au graphe produit par les commandes

```
matplot(x,intervalles,type="l",col=c(1,2,2,3,3))
legend("topleft",lty="solid",col=2:3,c("confiance","prediction"))
```

Université Paris Nanterre
Master 1 Iséfar
TD de Modèles de régression

# Feuille de TD n° 2 : Le coefficient de détermination

Soient $y_1, \ldots, y_n$ des réels, $y$ le vecteur correspondant, $V$ un sous-espace vectoriel de $\mathbb{R}^n$, et $W$ l'espace vectoriel engendré par $\mathbb{1}_n$, le vecteur de longueur $n$ dont toutes les composantes sont égales à 1. On rappelle que le coefficient de détermination associé au modèle $V$ est alors défini par

$$R^2 = \frac{\|P_V y - P_W y\|^2}{\|y - P_W y\|^2},$$

où $P_V$ et $P_W$ sont les matrices de projection orthogonale sur $V$ et $W$, respectivement.

1. Pourquoi le coefficient de détermination n'a-t-il d'interprêtation que si $W$ est un sous-espace de $V$ ?
2. Vérifier que si $V_0$ est un sous-espace vectoriel de $V$, et si $W$est un sous-espace vectoriel de $V_0$ (et donc aussi de $V$), alors le $R^2$ associé à $V_0$ est nécessairement inférieur au $R^2$ associé à $V$.
3. En notant $\hat{y}_i$ la $i$-ème composante du vecteur $P_V y$, et $\bar{y} = n^{-1} \sum_{i=1}^n y_i$, vérifier que

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

4. Quelle est la valeur de $R^2$ dans le cas où $y \in V$ ?
5. Nous considérons dans cette question le cas où $V$ est l'espace vectoriel engendré par $\mathbb{1}_n$ et un vecteur $x$, dont on notera les composantes $x_1, \ldots, x_n$. On note $\rho$ le coefficient de corrélation linéaire entre les $x_i$ et les $y_i$, défini par

$$\rho = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}},$$

où $\bar{x} = n^{-1} \sum_{i=1}^n x_i$ et $\bar{y} = n^{-1} \sum_{i=1}^n y_i$. On note de plus $\hat{\mu} = \bar{y} - \hat{a}\bar{x}$ et

$$\hat{a} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

de sorte que l'équation de la droite des moindres carrés est

$$y = \hat{\mu} + \hat{a}x.$$

(a) Vérifier que

$$R^2 = \frac{\hat{a}^2 \sum_{i=1}^n (\hat{x}_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

(b) En déduire que $R^2 = \rho^2$.

6. Générer dans **R** les vecteurs `x` et `y` de la façon suivante

```
x <- seq(-2,2,by=.2)
y <- (9/5)*x+32
```

Représenter graphiquement le nuage de points de coordonnées $(x_i, y_i)$, $i \in \{1, \ldots, n\}$ ; calculer $\rho^2$ dans **R** en utilisant la fonction `cor`. La commande

```
summary(lm(y~x))
```

permet-elle de retrouver la valeur de $\rho^2$ ? Quel modèle statistique avons-nous considéré dans cette commande ?

7. Reprendre la question précédente en remplaçant `y` par

```
y <- (9/5)*x+32+rnorm(length(x))
```

puis par

```
y <- (9/5)*x+32+rnorm(length(x), sd=2)
```

Commenter les résultats obtenus.

8. (a) Reprendre encore la même question en remplaçant `y` par

```
y <- (9/5)*x^2+32+rnorm(length(x))
```

(b) Décrire avec soin le modèle statistique implémenté par les commandes

```
t <- x^2
summary(lm(y~x+t))
```

puis représenter sur le même graphique le nuage de points, la fonction de régression et la fonction de régression ajustée.

(c) Quel est ici la valeur du coefficient de corrélation linéaire ? Commenter.

9. Nous travaillons dans cette question sur le jeu de données **carstopping**, disponible sur la plateforme `Cours en Ligne`. Le jeu de données est tiré de la page

`https://online.stat.psu.edu/stat501/lesson/1/1.8/1.8.2`

sur laquelle il est décrit ainsi : *The American Automobile Association has published data (Defensive Driving : Managing Time and Space, 1991) that looks at the relationship between the average stopping distance ( y = distance, in feet) and the speed of a car (x = speed, in miles per hour). The data set contains 63 such data points.*

(a) Représenter graphiquement les données, ainsi que la droite des moindres carrés.

(b) Quel coefficient de détermination obtient-on dans le modèle de régression linéaire simple ? Cette valeur permet-elle d'affirmer que ce modèle est très bon ?

(c) Proposer un autre modèle linéaire pour ces données, et représenter graphiquement l'ajustement obtenu.

10. Nous travaillons dans cette question sur le jeu de données **mccoo**, disponible sur la plateforme `Cours en Ligne`. Le jeu de données est tiré de la page

    `https://online.stat.psu.edu/stat501/lesson/1/1.8/1.8.2`

    sur laquelle il est décrit ainsi : *The dataset contains data on the running back Eric McCoo's rushing yards (mccoo) for each game of the 1998 Penn State football season. It also contains Penn State's final score (score).*

    (a) Représenter graphiquement les données, ainsi que la droite des moindres carrés.

    (b) Quel coefficient de détermination obtient-on dans le modèle de régression linéaire simple ?

    (c) Quel coefficient de détermination obtient-on dans le modèle de régression linéaire simple si l'on supprime la donnée pour laquelle la valeur de `McCoo` est égale à 206 ? Commenter.

Université Paris Nanterre
Master 1 ISÉFAR
TD de Modèles de régression

# Feuille de TD n° 3 : Régression linéaire multiple

## 1 Données bodyfat

Nous travaillons dans cette partie sur le jeu de données **bodyfat**, disponible sur la plateforme `Cours en Ligne`. Le jeu de données est tiré de la page

`https://online.stat.psu.edu/stat501/lesson/5/5.5`

sur laquelle il est décrit ainsi : *For a sample of $n = 20$ individuals, we have measurements of $y$ = body fat, $x_1$ = triceps skinfold thickness, $x_2$= thigh circumference, and $x_3$ = midarm circumference.*

NB : Le fichier accessible sur la page internet référencée ci-dessus est encodé au format `UTF-16`, ce qui peut poser des problèmes de lecture. Vous pouvez télécharger le fichier, l'ouvrir et le sauvegarder au format `UTF-8` pour supprimer ce problème. Ou bien utilisez la version disponible sur `Cours en Ligne`.

Soient les commandes et les sorties **R** suivantes :

```
> dat <- read.table("bodyfat.txt", header=T)
> pairs(dat)
> fit1 <- lm(Bodyfat~Triceps+Thigh+Midarm, data=dat)
> fit0 <- lm(Bodyfat~1, data=dat)
> fit2 <- lm(Bodyfat~., data=dat)
 > summary(fit1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173    0.258
Triceps        4.334      3.016   1.437     ??
Thigh         -2.857      2.582    ??      0.285
Midarm        -2.186      1.595  -1.370    0.190

Residual standard error: 2.48 on ?? degrees of freedom
Multiple R-squared:  0.8014,Adjusted R-squared:  0.7641
F-statistic: 21.52 on ?? and ?? DF,  p-value: 7.343e-06
```

1. A quoi sert l'option `header=T` dans l'appel **R** à `read.table` ?
2. Commenter la représentation graphique fournie par la commande `pairs(dat)`.
3. A quels modèles statistiques les objets `fit1`, `fit0` et `fit2` correspondent-ils ? Il est demandé de décrire ces modèles de façon rigoureuse, et d'en fournir la formulation matricielle.
4. Dans le modèle correspondant à l'objet `fit1`, donner l'équation de prédiction obtenue, c'est-à-dire l'équation permettant de prédire, pour un individu dont les valeurs des variables `Triceps`, `Thigh` et `Midarm` sont connues, la valeur correspondante de `Bodyfat`. Implémenter dans **R** une fonction réalisant ces prédictions : la fonction prendra en entrée des valeurs pour `Triceps`, `Thigh` et `Midarm`, et donnera en sortie la valeur prédite correspondante de `Bodyfat`.
5. Recalculer dans **R** (ou de tête lorsque c'est possible) les valeurs masquées par des `??`. Pour cela, il est demandé d'extraire les éléments non masqués utiles de l'objet `summary(fit1)`, et de faire le calcul avec des fonctions usuelles de **R**. Le calcul devra être justifié.
6. En utilisant la fonction `vcov` de **R**, calculer la matrice $({}^tXX)^{-1}$. Vous pourrez vérifier votre réponse sur la page internet citée en début de partie.
7. En utilisant la fonction `vcov`, recalculer les valeurs fournies dans la colonne `Std. Error` de l'objet `fit1`.
8. La variable `Thigh` semble-t-elle jouer un rôle significatif ? Vous argumenterez votre réponse et interprêterez le résultat (vous expliquerez pourquoi, selon vous, cette variable est, ou n'est pas, significative).
9. Les variables `Triceps` et `Midarm` semblent-elles jouer un rôle significatif ? Vous argumenterez votre réponse et utiliserez pour cela la fonction step de **R**.
10. Quel modèle retenez-vous pour ce jeu de données ? Vous validerez le modèle proposé.

# 2 Données coolhearts

Nous travaillons dans cette partie sur le jeu de données **coolhearts**, disponible sur la plateforme `Cours en Ligne`. Le jeu de données est tiré de la page

`https://online.stat.psu.edu/stat501/lesson/6/6.1`

sur laquelle il est décrit ainsi : *When heart muscle is deprived of oxygen, the tissue dies and leads to a heart attack ("myocardial infarction"). Apparently, cooling the heart reduces the size of the heart attack. It is not known, however, whether cooling is only effective if it takes place before the blood flow to the heart becomes restricted. Some researchers (Hale, et al, 1997) hypothesized that cooling the heart would be effective in reducing the size of the heart attack even if it takes place after the blood flow becomes restricted.*

*To investigate their hypothesis, the researchers conducted an experiment on 32 anesthetized rabbits that were subjected to a heart attack. The researchers established three experimental groups :*

— *Rabbits whose hearts were cooled to 6° C within 5 minutes of the blocked artery ("early cooling", Group 1)*
— *Rabbits whose hearts were cooled to 6° C within 25 minutes of the blocked artery ("late cooling", Group 2)*
— *Rabbits whose hearts were not cooled at all ("no cooling", Group 3)*

*At the end of the experiment, the researchers measured the size of the infarcted (i.e., damaged) area (in grams) in each of the 32 rabbits. But, as you can imagine, there is great variability in the size of hearts. The size of a rabbit's infarcted area may be large only because it has a larger heart. Therefore, in order to adjust for differences in heart sizes, the researchers also measured the size of the region at risk for infarction (in grams) in each of the 32 rabbits..*

Pour chacun des $n = 32$ lapins, nous disposons de la taille de la zone infarcie (`Inf.`), la taille de la région à risque d'infarctus (`Area`) et le groupe expérimental (`Group`).

1. Stocker les données dans un data.frame `dat`, puis les représenter graphiquement.
2. Représenter la droite des moindres carrés de chaque groupe expérimental.
3. La variable `Group` est-elle qualitative ou quantitative ?
4. Soient les commandes **R** suivantes :

```
attach(dat)
gp <- as.factor(Group)
fit.quant <- lm(Inf.~Area+Group)
fit.qual <- lm(Inf.~Area+gp)
fit.int <- lm(Inf.~gp+Area:gp)
fit.int2 <- lm(Inf.~Area*gp)
```

   A quels modèles statistiques les objets `fit.quant`, `fit.qual`, `fit.int` et `fit.int2` correspondent-ils ? Il est demandé de décrire ces modèles de façon rigoureuse, et d'en fournir la formulation matricielle.
5. Lequel des modèles proposés vous semble-t-il préférable ?
6. Dans chaque modèle, indiquer s'il faut ou non poser une contrainte d'identifiabilité, et préciser quelle est la contrainte d'identifiabilité posée par défaut par `R`. Afficher dans **R** la matrice de design associée à chacun des modèles.
7. Quelle est la dimension de chacun de ces modèles ?
8. Donner les équations des droites des moindres carrés représentées en question 2.
9. A quelle mesure de `Inf.` peut-on s'attendre pour un individu dont le coeur n'a pas été refroidi, et dont la valeur de `Area` est de 1 ? On donnera ici une prédiction ponctuelle, un intervalle de prédiction, et un intervalle de confiance pour l'espérance.
10. Au vu de l'expérience réalisée, peut-on valider l'hypothèse de Hale et al. ?
11. Peut-on considérer que le refroidissement tardif a le même effet que le refroidissement précoce ?

# 3 Données Credit

Nous travaillons dans cette partie sur le jeu de données **Credit**, disponible sur la plateforme `Cours en Ligne`. Le jeu de données est tiré de [1], où il est décrit ainsi : *The* `Credit` *data records* `balance` *(average credit card debt for a number of individuals) as well as several quantitative predictors :* `age`*,* `cards` *(number of credit cards),* `education` *(years of education),* `income` *(in thousands of dollars),* `limit` *(credit limit), and* `rating` *(credit rating). In addition to these quantitative variables, we also have four qualitative variables :* `gender`*,* `student` *(student status),* `status` *(marital status), and* `ethnicity` *(Caucasian, African American or Asian).*

1. Stocker les données dans un data.frame `mydata`, puis les représenter graphiquement.
2. Pour combien d'individus la variable `Balance` pend-elle la valeur 0 ? Cela est-il cohérent avec un modèle qui suppose que les valeurs dans `Balance` sont les réalisations de variables aléatoires Gaussiennes indépendantes ?
3. Stocker dans un data.frame `mydata0` les observations pour lesquelles le débit moyen des cartes de crédit est strictement positif.
4. Sur les observations pour lesquelles `Balance` prend une valeur strictement positive, faire la régression linéaire simple de `Balance` sur `Limit`. Comparer ce modèle avec le modèle de régression polynomiale d'ordre deux. Il est demandé de décrire précisément les modèles statistiques, de représenter graphiquement les ajustements obtenus, et de justifier le test d'hypothèses réalisé.
5. Sur les données de `mydata0`, ajuster un modèle linéaire sur `Balance` incluant un effet linéaire de `Limit` et un effet logarithmique de `Limit`. Peut-on comparer ce modèle avec celui de la question précédente grâce à un test de Fisher ? Commenter le résultat de la fonction `anova` de **R** appliquée aux deux objets de type `lm` associés à ce deux modèles.
6. Sur les données de `mydata0`, représenter graphiquement `Balance` en fonction de `Limit` et de `Student`. L'effet de `Limit` semble-t-il différer selon les modalités de `Student` ?
7. Sur les données de `mydata0`, ajuster un modèle de régression linéaire pour la variable réponse `Balance`, incluant une interaction entre les variables explicatives `Student` et `Limit` et un effet d'ordre deux de `Limit`. Stocker les résultats dans un objet `fit7`.
8. Comparer ce modèle à celui de la question 4.
9. L'effet d'ordre deux de `Limit` est-il significatif dans le modèle associé à `fit7` ?
10. Calculer les coefficients de corrélation linéaire entre les variables quantitatives. Quelle est la variable la plus corrélée linéairement avec `Balance` ?
11. Rechercher un modèle parcimonieux en utilisant la fonction `step` ou la fonction `leaps:regsubsets` de **R**. Comparer ce modèle à celui de la question 7. Le modèle retenu vous semble-t-il convenable ? Donner l'équation de régression correspondante.

# Références

[1] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

Université Paris Nanterre
Master 1 Iséfar
TD de Modèles de régression

# Feuille de TD n° 4 : Modèle linéaire généralisé

## 1 EMV

Soient $Y_1, \ldots, Y_n$ des observations indépendantes à valeurs dans $\{0, 1\}$. On suppose que ces observations sont régies par un modèle de régression logistique dans lequel chaque $Y_i$ est associée à un vecteur ligne de variables explicatives $x_i$. Dans la suite, on note $\beta$ le vecteur colonne des paramètres inconnus et $p$ sa dimension.

1. Exprimer la loi de $Y_i$ en fonction de $\beta$.
2. Exprimer la vraisemblance du modèle, puis sa log-vraisemblance.
3. Calculer la dérivée de la log-vraisemblance, puis exprimer la matrice des dérivées partielles secondes en fonction de la densité de la loi logistique.
4. Montrer que la matrice des dérivées partielles secondes (matrice Hessienne) est semi-définie négative, et qu'elle est même définie négative si la matrice $X$ constituée des lignes $x_i$ est de rang $p$.
5. En déduire que si $X$ est de rang $p$ et $\hat{\beta}$ est un vecteur qui annule la dérivée de la log-vraisemblance, alors $\hat{\beta}$ est l'unique EMV.

## 2 Données leukemia_remission : régression logistique

Nous travaillons dans cette partie sur le jeu de données **leukemia_remission**, disponible sur la plateforme `Cours en Ligne`. Le jeu de données est tiré de la page

`https://online.stat.psu.edu/stat501/lesson/15/15.1`

sur laquelle il est décrit ainsi : *The Leukemia Remission data set has a response variable of whether leukemia remission occurred (REMISS), which is given by a 1.*

*The predictor variables are cellularity of the marrow clot section (CELL), smear differential percentage of blasts (SMEAR), percentage of absolute marrow leukemia cell infiltrate (INFIL), percentage labeling index of the bone marrow leukemia cells (LI), absolute number of blasts in the peripheral blood (BLAST), and the highest temperature prior to start of treatment (TEMP).*

1. Dans **R**, créer un data frame `mydata` contenant les données.
2. Ajuster une régression logistique sur la variable `REMISS`. On se contentera ici d'inclure toutes les autres variables du jeu de données, de façon additive et sans interaction.
3. Le modèle précédent vous semble-t-il parcimonieux ?
4. Proposer un modèle parcimonieux. Décrire soigneusement le modèle en question.
5. Estimer la fonction qui à un vecteur de variables explicatives $x$ associe la probabilité pour qu'un individu associé à $x$ ait une rémission.
6. Quelle estimation obtient-on pour la probabilité qu'un individu ait une rémission, lorsque l'individu est caractérisé par les valeurs suivantes ?

| CELL | MEAR | INFIL | LI | BLAST | TEMP |
|---|---|---|---|---|---|
| 0.8 | 0.83 | 0.66 | 0.9 | 1.1 | 1 |

7. Considérons la régression logistique de `REMISS` sur la seule variable explicative `LI`.
   (a) Représenter graphiquement la probabilité de succès ajustée dans ce modèle.
   (b) Ce modèle est-il préférable au précédent ?
8. Peut-on bâtir un test d'adéquation fondé sur la déviance ?

# 3 Modèle de Poisson

On observe des variables aléatoires indépendantes $Y_1, \ldots, Y_n$ de loi de Poisson. On suppose que pour tout $i \in \{1, \ldots, n\}$,

$$E(Y_i) = \exp\left(\beta_1 + \sum_{j=2}^{p} \beta_j x_{ij}\right), \tag{4}$$

où $p$ est un entier tel que $1 < p < n$, $\beta_1, \ldots, \beta_p$ sont des paramètres inconnus et pour tout $i \in \{1, \ldots, n\}$ et $j \in \{2, \ldots, p\}$, $x_{ij} \in \mathbb{R}$ est une variable explicative observée. Dans la suite, on notera $\beta$ le vecteur colonne de composantes $\beta_1, \ldots, \beta_p$ et, pour tout $i \in \{1, \ldots, n\}$, on notera $x_i$ le vecteur ligne de composantes $1, x_{i2}, \ldots, x_{ip}$.

1. Vérifier que
$$E(Y_i) = \exp\left(x_i \beta\right).$$
2. Montrer qu'il s'agit d'un modèle linéaire généralisé. Quelle est la fonction de lien ?

On suppose dans la suite que la matrice constituée des lignes $x_1, \ldots, x_n$ est de rang $p$.

3. Pourquoi considérer une matrice de rang $p$ ?
4. Montrer que le vecteur des scores admet comme première composante
$$\sum_{i=1}^{n} \left(Y_i - \exp(x_i \beta)\right).$$

On suppose dans la suite que l'estimateur du maximum de vraisemblance de $\beta$ est bien défini, et on le note $\hat{\beta}$. On ne cherchera pas à donner une écriture explicite de cet estimateur, puisque l'estimateur ne possède pas d'écriture explicite (en pratique, on le calcule en utilisant une méthode itérative implémentée sur les logiciels tels que **R**).

5. Déduire de la question précédente que

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \exp(x_i \hat{\beta}).$$

6. Calculer la déviance du modèle, puis utiliser la question précédente pour montrer qu'elle s'écrit

$$2 \sum_{i:Y_i \neq 0} Y_i \left( \log Y_i - x_i \hat{\beta} \right),$$

où la somme porte sur les indices $i$ tels que $Y_i \neq 0$.

7. On souhaite tester $H_0$ : "$\beta_2 = \beta_3$" contre $H_1$ : "$\beta_2 \neq \beta_3$". On note $\hat{\beta}_0$ l'estimateur du maximum de vraisemblance de $\beta$ calculé sous l'hypothèse $H_0$ (on suppose qu'il est bien défini).

   (a) Montrer que la statistique de test du rapport des vraisemblances est

$$2 \sum_{i=1}^{n} Y_i \left( x_i \hat{\beta} - x_i \hat{\beta}_0 \right).$$

   (b) Exprimer la $p$-valeur du test du rapport des vraisemblances. S'agit-il d'un test exact ou d'un test asymptotique ?

8. Dans cette question, nous allons simuler des données sous **R** puis illustrer les résultats théoriques obtenus dans les questions précédentes. Les commentaires attendus ci-dessous doivent donc faire le lien entre les résultats obtenus dans **R** et ces résultats théoriques.

   (a) Dans **R**, exécuter les commandes suivantes, et expliquer ce qu'elles produisent.

```
set.seed(123)
n <- 30
x1 <- rnorm(n)
x2 <- rnorm(n)
X <- cbind(1,x1,x2)
beta <- c(1,2,2.1)
m <- exp(X%*%beta)
Y <- rpois(n,m)
table(Y)
```

   (b) Dans **R**, éxécuter les commandes suivantes et décrire précisément le modèle correspondant à l'objet `fit`.

```
fit <- glm(Y ~ x1+x2, family = poisson(link = "log"))
summary(fit)
```

(c) Dans **R**, éxécuter les commandes suivantes et commenter le résultat.

```
p <- predict (fit,type="response")
sum(p)
sum(Y)
```

(d) Dans **R**, éxécuter les commandes suivantes et commenter le résultat.

```
l <- predict(fit,type="link")
dev <- 2*sum((Y[Y>0])*(log(Y[Y>0])-l[Y>0]))
```

(e) Dans **R**, éxécuter les commandes suivantes et commenter le résultat.

```
x=x1+x2
fit0 <- glm(Y ~ x, family = poisson(link = "log"))
summary(fit0)
anova(fit0,fit,test="Chisq")
stat <- 2*sum(Y*(predict(fit,type="link") - predict(fit0,type="link")))
stat
1-pchisq(stat,1)
```

# 4 Données GermanCredit : régression logistique

Nous travaillons dans cette partie sur le jeu de données **GermanCredit**, disponible sur la plateforme `Cours en Ligne`. Le jeu de données est tiré de la page

`https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)`

sur laquelle il est décrit ainsi : *This dataset classifies people described by a set of attributes as good or bad credit risks. Comes in two formats (one all numeric). Also comes with a cost matrix.* Nous n'exploiterons pas la matrice de coûts dans cet exercice.

1. Dans **R**, créer un data frame `credit` contenant les données grâce aux commandes

```
noms_variables <- c("status_account","duration","history","purpose",
       "amount","savings","employment","installement_rate",
       "status_personnal","other_debtors","residence_since",
       "property","age","other_installment","housing",
       "nb_credits","job","liable","telephone","foreign","class")
credit0 <- read.table("GermanCredit.txt", col.names = noms_variables)
credit0$class <- as.factor(credit0$class-1)
credit <- subset(credit0, select=-c(foreign))
```

Noter que nous avons exclu la variable foreign de l'étude, et que nous avons attribué un nom à chacun des attributs. Comment s'interprête l'objet `class` du data frame `credit` ?

2. Décrire les données (on pourra s'appuyer sur la fonction `summary` de **R** ou sur des représentations graphiques).
3. Ajuster une régression logistique sur la variable `class`.
4. Le modèle précédent vous semble-t-il parcimonieux ?
5. Proposer un modèle parcimonieux. Décrire soigneusement le modèle en question.
6. Estimer la fonction qui à chaque vecteur d'attributs $x$ associe la probabilité pour qu'un individu possédant les attributs $x$ soit un bon client.