Statistiques appliquées aux sciences de l'ingénieur

ÉCOLE CENTRALE LYON

BE – 3

Plans d'expérience et régression logistique

Julie HAZARD
Alexei REIMAN

29/11/2017

Sommaire

Exercice 1 – Plan d'expériences

Lors de ce BE, nous allons travailler sur les défauts de fabrications sur la planéité de plaques de silicium. En effet, à cause de ces défauts récurrents, le rendement de l'entreprise n'est pas optimal et une étude de l'apparition de ces mêmes défauts pourrait améliorer le rendement en comprenant la cause des problèmes. On considère 6 facteurs pouvant intervenir sur ces rendements :

- Ltime le temps de laminage
- Ltemp la température de laminage
- Lpress la pression de laminage
- Ctemp la température de cuisson
- Ctime le temps de cuisson
- Catmos l'atmosphère de cuisson

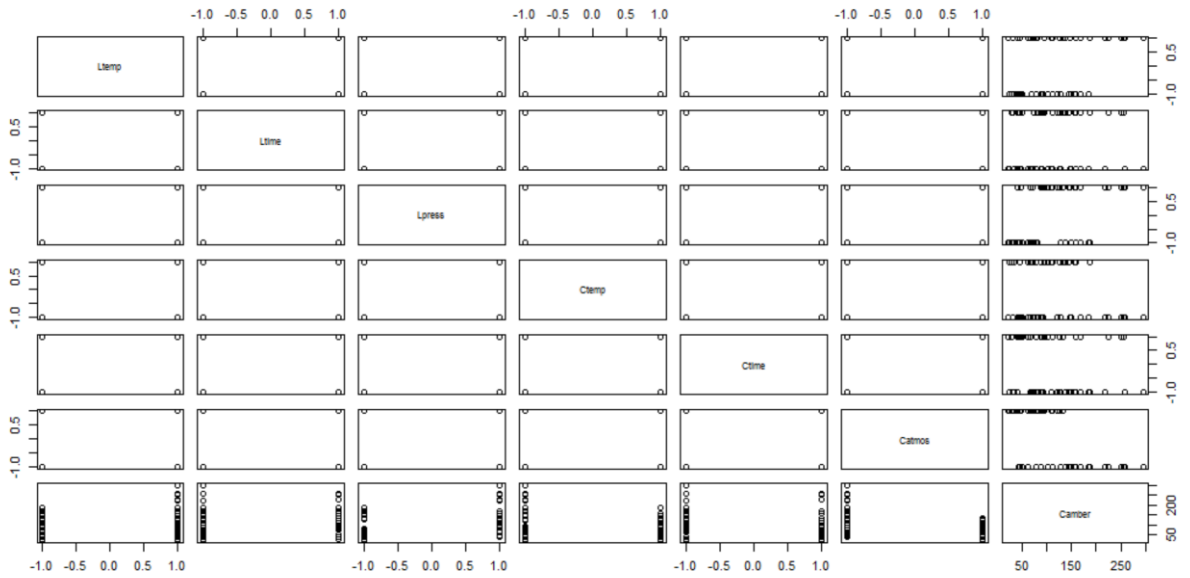


Figure 1 : Plot Silicium

On voit dans ce plan que les 6 variables décrites précédemment prennent chacune soit la valeur 1, soit -1 et que Camber décrit ensuite la courbure observée de la plaque. Toutes les expériences réalisées sont de niveau 1 ou -1, on a par ailleurs 64 expériences, soit 2^6. On a une matrice orthogonale, On pourrait donc penser que nous avons un plan orthogonal complet. Toutefois, on a répété chaque expérience 4 fois car il semble y avoir beaucoup de variabilité sur la sortie, on a donc en réalité seulement 16 expériences, soit 2^4. Nous sommes donc avec un plan fractionnaire 2^{6-2}.

On considère X, la matrice dans laquelle on a mis les 64 expériences réalisées sans le résultat de cambrure. En calculant la transposée de X, on trouve le résultat de la figure 2 : une matrice diagonale avec que des 64 sur la diagonale.

	Ltemp	Ltime	Lpress	Ctemp	Ctime	Catmos
Ltemp	64	0	0	0	0	0
Ltime	0	64	0	0	0	0
Lpress	0	0	64	0	0	0
Ctemp	0	0	0	64	0	0
Ctime	0	0	0	0	64	0
Catmos	0	0	0	0	0	64

Figure 2 : Transposée de X

Pour ne pas prendre en compte les répétitions d'expérience, on ne prend en compte que les 16 premières parmi les 64 du fichier original

Nous allons à présent chercher les interactions triples de ce plan : on s'aperçoit que Ctime est l'interaction triple de Ltemp, Ltime et Lpress. De même Catmos est l'interaction triple de Ltemp, Lpress et Ctemp. En revanche, on a Ctime*Catmos=Ltime*Ctemp.

Nous sommes donc dans un plan de résolution $(6-2)=4$.

En faisant une régression linéaire puis une Anova, on obtient la figure 3.

```
Call:
lm(formula = Camber ~ Ltemp + Ltime + Lpress + Ctemp + Ctime +
    Catmos, data = silicium)

Residuals:
    Min      1Q  Median      3Q     Max
-67.672 -22.703  -3.875  28.797  81.328

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  107.016      4.793  22.328  < 2e-16 ***
Ltemp         19.453      4.793   4.059 0.000152 ***
Ltime          2.891      4.793   0.603 0.548823
Lpress        28.016      4.793   5.845 2.57e-07 ***
Ctemp         -7.109      4.793  -1.483 0.143492
Ctime        -17.234      4.793  -3.596 0.000676 ***
Catmos       -38.734      4.793  -8.082 5.03e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38.34 on 57 degrees of freedom
Multiple R-squared:  0.6975,    Adjusted R-squared:  0.6657
F-statistic: 21.91 on 6 and 57 DF,  p-value: 3.566e-13
```

```
Analysis of Variance Table

Response: Camber
           Df Sum Sq Mean Sq F value    Pr(>F)
Ltemp       1  24219   24219 16.4740 0.0001520 ***
Ltime       1    535     535  0.3638 0.5488232
Lpress      1  50232   50232 34.1681 2.574e-07 ***
Ctemp       1   3235    3235  2.2003 0.1434922
Ctime       1  19010   19010 12.9304 0.0006762 ***
Catmos      1  96023   96023 65.3150 5.028e-11 ***
Residuals 57  83798    1470
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3 : Régression Linéaire et Anova

On peut retrouver notre Standard Error par calcul avec $\sqrt{\hat{\sigma}^2({}^tXX^{-1})}$. On sait que ${}^tX$est une matrice diagonale avec que des 64, on cherche donc $\frac{\hat{\sigma}}{\sqrt{64}}$. Quel que soit le coefficient, on aura toujours la même valeur, ce qui nous donne bien 38,34.

On déduit facilement de la figure 3 que Ctemp et Ltime ne sont pas significatives et que l'on peut les supprimer. On peut à présent faire un modèle en ne prenant en compte que les facteurs influents. On obtient avec la régression linéaire la figure 4. On s'aperçoit que les coefficients des paramètres significatifs sont exactement les mêmes et que par ailleurs toutes les variables passent le test de Student, on peut donc rester dans cette configuration-là.

```
Call:
lm(formula = Camber ~ Ltemp + Lpress + Ctime + Catmos, data = silicium)

Residuals:
    Min      1Q  Median      3Q     Max
-74.953 -26.547  -4.016  25.844  85.547

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  107.016      4.816  22.222  < 2e-16 ***
Ltemp         19.453      4.816   4.040 0.000157 ***
Lpress        28.016      4.816   5.818 2.59e-07 ***
Ctime        -17.234      4.816  -3.579 0.000698 ***
Catmos       -38.734      4.816  -8.043 4.62e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38.53 on 59 degrees of freedom
Multiple R-squared:  0.6839,    Adjusted R-squared:  0.6625
F-statistic: 31.92 on 4 and 59 DF,  p-value: 3.711e-14
```

Figure 4 : Summary de la régression linéaire en modèle réduit

On peut maintenant chercher à minimiser à la courbure de la plaque. Pour cela, on peut se concentrer sur les coefficients de la colonne Estimate du summary de notre régression linéaire. S'ils sont positifs on peut essayer de minimiser la valeur associée et s'ils sont négatifs au contraire on va chercher à maximiser la valeur associée.

On a Ltemp et Lpress qui ont un Estimate positif et Ctime, Catmos négatif. On peut donc associer -1 aux premiers et 1 aux deuxièmes dans R pour faire l'optimisation cherchée. On obtient :

```
                          > CourbureSilicium
                          [1] 43 27 28 28
                          > moyenneCamberSilicium
                          [1] 31.5
> predict(siliciumLmsansautresfacteurs, frame, interval="confidence", level=0.95)
       fit      lwr      upr
1 3.578125 -17.9689 25.12515
> frame
  Ltemp Lpress Ctime Catmos
1    -1     -1     1      1
```

On a donc un intervalle compris entre la lower value et la upper value (-17,9689 et 25,12515). On a donc en arrondi un intervalle de confiance de [-18 ; 25,2]. En valeur absolue, on peut arrondir à [0 ; 25]. Comme on a choisi un intervalle de confiance à 95%, on a :

$$\left[\bar{x} - 2\frac{\sigma(X)}{\sqrt{n}} ; \bar{x} + 2\frac{\sigma(X)}{\sqrt{n}}\right]$$

Nos réalisations seront donc inférieures à $25 + 2 * \text{STd Error} = 101{,}68$.

On cherche maintenant à voir l'impact sur la courbure d'une augmentation de 5 degrés de la température de laminage. On est dans la situation $Y = C_0 + \beta X + \varepsilon$. Passer de 0 à 1 dans Ltemp correspond en réalité – d'après le summary précédent de la régression linéaire – à une augmentation de la courbure de 19,453. Or cela correspond par ailleurs à un passage de 65 à 75° pour Ltemp. On a donc que pour une augmentation de 5°, la courbure va augmenter de $\frac{19{,}453}{2} = 9{,}7265$.

Les hypothèses sont du modèle ne sont pas vraiment vérifiées. En effet, on trouve un R-squared de 0,68 ce qui est assez faible. On peut penser que c'est dû au fait que le modèle soit seulement additif et qu'on ne prend pas en compte les interactions. On voit la représentation de la régression de ce modèle en figure 5.

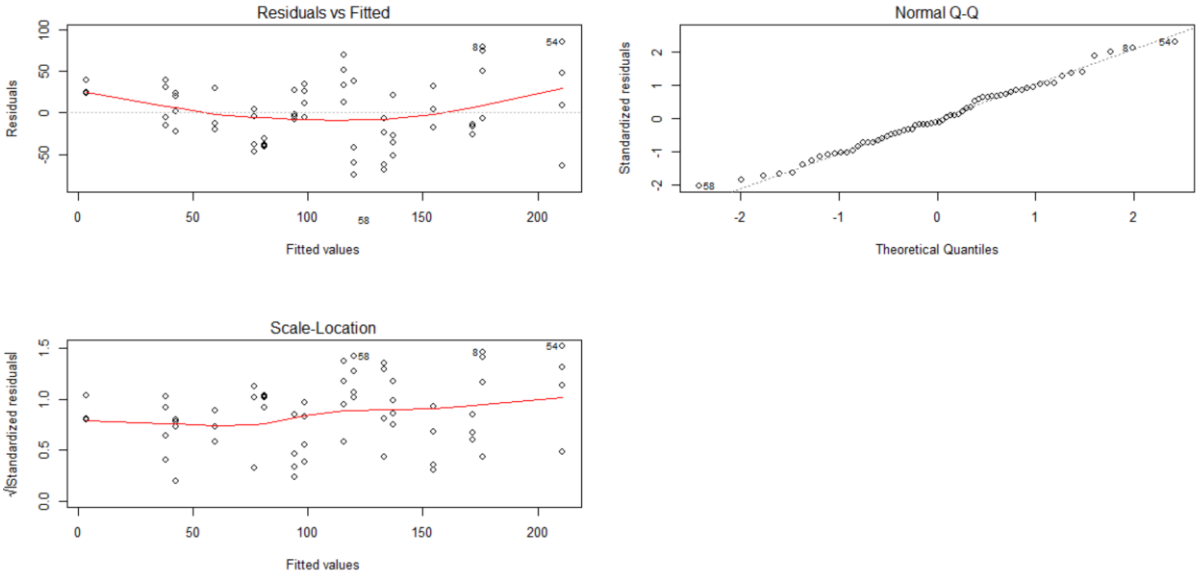


Figure 5 : Représentation – Régression Linéaire du modèle « simple »

Pour proposer un meilleur modèle, on pourrait envisager de rajouter les interactions dans le modèle. Dans ce nouveau modèle, on trouve alors un R-squared de 0,74 ce qui est mieux que le précédent sans être pour autant vraiment bien. En revanche, l'intervalle de confiance est plus grand que le précédent. Il serait donc moins intéressant de travailler avec ce modèle (représenté ci-dessous).

```
> predict(Lmnouveaumodele, frame2, interval="confidence", level=0.95)
        fit       lwr     upr
1 -19.14062 -42.52162 4.24037
```

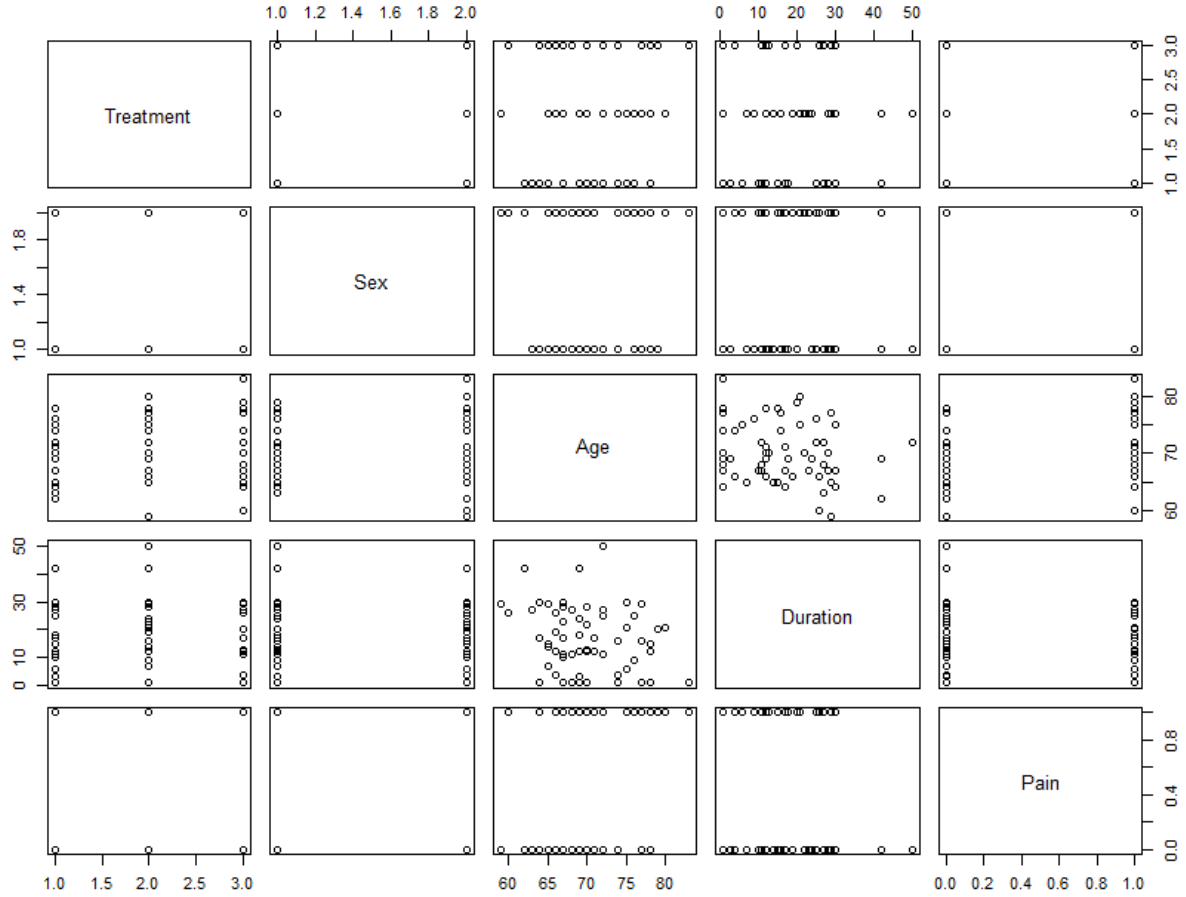
Exercice 2 – Régression logistique

Question 1

Nous travaillons dans cet exercice sur le fichier neuralgia.tkt qui contient 60 observations. La variable à expliquer est la variable Pain, c'est une variable à deux modalités (Oui il y a Pain ou non il n'y a pas Pain). Le fichier contient 4 variables explicatives :

- Treatment : variable à trois modalités
- Sex : variable à deux modalités
- Age : variable à modalités multiples
- Duration : variable à modalités multiples

Voici ce que nous donne R comme représentation des données de notre fichier :



Question 2

80% des observations de notre fichiers correspond à 48 observations.

Question 3

L'évènement modélisé ici est l'évènement « il y a Pain ». Soit π(x) la probabilité de l'évènement que l'on cherche à modéliser. Alors le modèle logit consiste à écrire que :

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p}}$$

Où $\beta_0,\ldots,\beta_p$ sont les paramètres à estimer. La fonction Logit est la fonction définie par :

$$p \to \log\left(\frac{p}{1-p}\right)$$

Question 4

Voici le ce que nous donne R pour la commande anova (test = « chist ») :

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	
NULL			47	61.105		
Treatment	2	13.6064	45	47.499	0.00111	**
Sex	1	2.1103	44	45.389	0.14631	
Age	1	3.0826	43	42.306	0.07914	.
Duration	1	1.1367	42	41.169	0.28635	

Nous observons que la variable la plus influente est la variable *Treatment*. La variable âge peut aussi être considérée comme une variable influente. Les variables les moins influentes sont les variables *Sex* et *Duration*.

Voici le ce que nous donne R pour la commande Anova de type III utilisant le rapport de vraisemblance:

Response: Pain

	LR Chisq	Df	Pr(>Chisq)	
Treatment	14.8888	2	0.0005847	***
Sex	1.7685	1	0.1835673	
Age	1.8332	1	0.1757503	
Duration	1.1367	1	0.2863536	

Nous pouvons observer que la variable la plus influente est la variable Treatment. Les trois autres variables sont beaucoup moins influentes.

Voici le ce que nous donne R pour la commande Anova de type III utilisant le test de Wald:

Response: Pain

	Df	Chisq	Pr(>Chisq)	
(Intercept)	1	2.0547	0.151741	
Treatment	2	10.5355	0.005155	**
Sex	1	1.6642	0.197037	
Age	1	1.6038	0.205362	
Duration	1	1.0709	0.300747	

Ici encore le test nous donne une variable Treatment très influente par rapport aux autres. Les trois autres variables sont beaucoup moins influentes.

Question 5

A chaque pas, une variable est ajoutée au modèle. C'est celle dont la valeur de p (test modèles emboités Fisher) est minimum. La procédure s'arrête lorsque toutes les variables sont introduites ou lorsque p reste plus grande qu'une valeur seuil fixée par défaut à 0.5. Voici ce que nous donne R :

Response: Pain

	LR Chisq	Df	Pr(>Chisq)	
Treatment	20.1322	2	4.25e-05	***
Age	11.1182	1	0.0008548	***
Sex	6.2766	1	0.0122340	*

Nous voyons que le traitement est encore le facteur le plus influent. La méthode forward a éliminé le facteur *duration*. Dans les tests précédents on avait trouvé que ce facteur était le moins influent.

Question 6

En éliminant le facteur duration de notre test, voici les prédictions sur l'échantillon test que l'on obtient avec R :

```
> prediction
         4          5         20         22         28         30         36         37         43
0.75285295 0.04033596 0.09818647 0.04661176 0.32125372 0.89297059 0.51406526 0.95577578 0.59455707
        44         51         59
0.32125372 0.80852388 0.51708974
```

On voit que l'échantillon 4,30,36,37,43,51 et 59 ont donné une probabilité pour l'évènement « il y a Pain » supérieur à 0,5. On conclut donc que ces patients vont probablement ressentir de la douleur.

Nous allons maintenant calculer la matrice de confusion du test. Voici ce que nous donne R :

	0	1
FALSE	3	2
TRUE	0	7

Seulement 2 instances ont été mal classée sur 12, c'est-à-dire que 2 patients qui avaient une douleur ont été classés comme des patients qui ne ressentent pas de douleur. Ceci qui représente 17% d'instances mal classées par la méthode.

Avec la commande anova (test= « chisq ») on avait vu que les facteurs les plus influents étaient le *Traitment* et *Age*. Nous allons donc dans un deuxième temps éliminer ces deux autres facteurs. Voici les résultats obtenus par R :

```
> predictionBis
         4          5         20         22         28         30         36         37         43
0.59634177 0.06910830 0.05981348 0.08703931 0.20335828 0.91661148 0.37806149 0.90402714 0.45288364
        44         51         59
0.20335828 0.66796679 0.35582292
```

On observe que les instances 4,30,37 et 51 59 ont donné une probabilité pour l'évènement « il y a Pain » supérieur à 0,5. On conclut donc que ces patients vont probablement ressentir de la douleur.

Voici la matrice de confusion donnée par R :

	0	1
FALSE	3	5
TRUE	0	4

5 instances ont été mal classée sur 12, c'est-à-dire que 5 patients qui avaient une douleur ont été classés comme des patients qui ne ressentent pas de douleur. Ceci représente 42% d'instances mal classées par la méthode. C'est beaucoup plus élevé que dans le modèle précédent. On ne peut donc pas négliger l'effet du facteur *Sex*. Ainsi le meilleur modèle est celui où l'on néglige seulement le facteur *Duration*.

Question 7

On trouve en moyenne 2.48 instances mal classé avec cette méthode, c'est-à-dire en moyenne 21 % des instances sont mal classés.