christine.keribin@universite-paris-saclay.fr

Modèle linéaire

Éléments de corrigé

Des procédés de hautes températures sont souvent utilisés pour traiter des pièces métalliques telles que les engrenages, en modifiant leurs caractéristiques de surface. La profondeur de trempe est une propriété importante de la pièce, et contribue à sa fiabilité. Des essais destructifs sont effectués en découpant la pièce pour mesurer la profondeur de trempe au niveau des dents de l'engrenage. Une ingénieure méthodes et procédés cherche à expliquer la profondeur de trempe `y` en fonction de quatre variables `soaktime`, `soakpct`, `difftime`, `diffpct`, à partir d'un jeu de données de 32 observations qui forment un modèle régulier [1].

1. Analyser les graphes de la dernière page.

   Correction. `y` *semble corrélée positivement avec* `soaktime` *et éventuellement* `difftime`. `y` *pourrait avoir une liaison négative avec* `diffpct`. *Les variables explicatives semblent peu corrélées entre elles. Un calcul de la matrice de corrélation permet d'obtenir des données chiffrées sur ces constatations visuelles. Deux points ayant des* `soaktime` *élevés semblent avoir un comportement un peu différent qu'il conviendrait d'analyser un peu plus finement.*

2. Modéliser par une régression linéaire multiple. Préciser la matrice du plan d'expérience, la dimension du modèle. En supposant les données gaussiennes, quelle est la loi de l'estimateur des moindres carrés?

   Correction. *Modèle de régression linéaire*

   $$y_i = \mu + \beta_{ST}\ \ soaktime_i + \beta_{SP}\ soakpct_i + \beta_{DT}\ difftime_i + \beta_{DP}\ diffpct_i + \varepsilon_i$$

   où le bruit ε_i *est iid gaussien* $\mathcal{N}(0, \sigma^2)$. μ *est le paramètre d'intercept et le paramètre à estimer est* $\theta = (\mu, \beta_{ST}, \beta_{SP}, \beta_{DT}, \beta_{DP})$ *de dimension* $p = 5$ *qui est également celle du modèle (car le modèle est régulier, ceci est indiqué dans l'énoncé. Dans les faits, c'est vérifié lors du calcul de* $(X'X)^{-1}$ *sur les valeurs du jeu de données).*

   On met le modèle sous forme matricielle: $Y = X\theta + \varepsilon$ *où la matrice du plan d'expérience est la concaténation d'une colonne de 1 (intercept) et quatre colonnes représentant chacune les valeurs des quatre variables pour les différentes pièces du lot; le vecteur* ε *est iid centré, gaussien, de variance* $\sigma^2 Id_n$.

   La variance σ^2 *joue le rôle d'un paramètre de nuisance qu'il faudra aussi estimer.*

   D'après le cours: $\widehat{\theta} = (X'X)^{-1}X'Y \sim \mathcal{N}(\theta, \sigma^2(X'X)^{-1}$ *et* $\widehat{\sigma}^2 = ||Y - X\widehat{\theta}||^2/(n-p)$.

3. Les résultats de l'estimation par moindres carrés sont consignés dans le tableau suivant.

[1] Generalized Linear Models. R. Myers, D. Montgomery, G. Vining. Wiley 2002

Parameter	Estimate	Std_Error	t_value	Pr(>|t|)
Intercept	0.0205609	0.0089236	2.304	0.0291
soaktime	0.0023847	0.0001533	15.560	5.28e-15
soakpct	-0.0038039	0.0057357	-0.663	0.5128
difftime	0.0083840	0.0012055	6.955	1.79e-07
diffpct	-0.0057911	0.0069046	-0.839	0.4090

Déterminer l'équation de la fonction de régression.

Correction. *La fonction de régression est*

$$\begin{aligned}\widehat{y} &= \widehat{\mu} + \widehat{\beta}_{ST}\, soaktime + \widehat{\beta}_{SP}\, soakpct + \widehat{\beta}_{DT}\, difftime + \widehat{\beta}_{DP}\, diffpct \\ &= (1\ soaktime\ soakpct\ difftime\ diffpct)\ \widehat{\theta}\end{aligned}$$

avec les valeurs des coefficients indiqués dans la colonne `Estimate`.

(a) Rappeler ce que représente chacune des colonnes.

Correction. *Pour chaque ligne j*

- `Estimate` *: valeur observée $\widehat{\theta}_j$ de l'EMC de θ_j*
- `Std_Error` *: estimation de l'écart type de $\widehat{\theta}_j$, soit $s_j = \widehat{\sigma}\sqrt{[(X'X)^{-1}]_{jj}}$, avec $\widehat{\sigma}^2 = \sum_i (y_j - \widehat{y}_i)^2/(32-5)$*
- `t_value`*: la valeur observée du test de Student de $(H_0) : \theta_j = 0$ contre $(H_1) : \theta_j \neq 0$, soit $t_{obs_j} = (\widehat{\theta}_j - 0)/s_j$ de région de rejet $\{|T| > q^{\mathcal{T}(n-p)}_{1-\alpha/2}\}$*
- `Pr(>|t|)`*: la p-value du test précédent, soit $pval_j = \mathbb{P}(|T| > |t_{obs_j}|)$ quand $T \sim \mathcal{T}(32-5)$*

La variable `diffpct` est-elle significative?

Correction. *Test de Student de $(H_0) : \theta_{DP} = 0$ contre $(H_1) : \theta_{DP} \neq 0$, $|t_{obs}| = 0.84 < q^{\mathcal{T}(32-5)}_{1-\alpha/2} = 2.02$. Les données ne sont pas significatives pour rejeter (H_0). On conserve donc le modèle restreint $(\mu, \beta_{ST}, \beta_{SP}, 0, \beta_{DP})$ par défaut, la variable* `difftime` *n'est pas utile pour expliquer le modèle. Risque de la décision de seconde espèce, inconnu.*

On peut aussi prendre la décision avec la pvalue du test $.41 > 5\% = \alpha$. On ne rejette pas (H_0) avec un risque (de seconde espèce) inconnu.

Le coefficient θ_{DT} est-il positif?

Correction. *Test de Student unilatéral: $(H_0) : \theta_{DT} \leq 0$ contre $(H_1) : \theta_{DT} > 0$. La région de rejet est unilatéral à droite $\mathcal{R} = \{T_{DT} > q^{\mathcal{T}(32-5)}_{1-\alpha} = 1.7\}$. La valeur observée appartient à la région de rejet car $6.93 > 1.7$. On rejette la négativité, et on affirme au risque $\alpha = 5\%$ que le coefficient est positif.*

Remarque: le test de $(H_0) : \theta_{DT} > 0$ contre $(H_1) : \theta_{DT} < 0$ de région de rejet $\mathcal{R} = \{T_{DT} < q^{\mathcal{T}(32-5)}_{1-\alpha} = -1.7\}$ amène à ne pas rejeter (H_0) mais avec un risque de seconde espèce inconnu. Il ne faut donc pas utiliser cette façon de définir la région de rejet qui ne permet pas de calibrer le risque de la décision qui nous intéresse.

(b) Déterminer un intervalle de confiance de β_{DT}, puis un intervalle de confiance de β_{DP}.

Correction. *Comme* $\widehat{\theta} \sim \mathcal{N}(\theta, \sigma^2(X'X)^{-1})$*, on déduit pour chaque composante* θ_j
$\mathbb{P}_{\theta_j}(|\widehat{\theta}_j - \theta_j|/s_j > q^{\mathcal{T}_{n-p}}_{1-\alpha/2}) = 1-\alpha$ *équivalent à* $\mathbb{P}_{\theta_j}(|\widehat{\theta}_j - \theta_j|/s_j \leq q^{\mathcal{T}_{n-p}}_{1-\alpha/2}) = 1-\alpha$ *d'où* $\mathbb{P}(\theta_j \in IC(\theta_j)) = 1-\alpha$ *avec*

$$IC(\theta_j) = \left[\widehat{\theta}_j - q^{\mathcal{T}_{n-p}}_{1-\alpha/2} s_j;\ \widehat{\theta}_j + q^{\mathcal{T}_{n-p}}_{1-\alpha/2} s_j\right] \text{ avec } s_j = \widehat{\sigma}\sqrt{[(X'X)^{-1}]_{jj}}$$

Cet intervalle $IC(\theta_j)$ *dont les bornes sont aléatoires et tel que* $\mathbb{P}(\theta_j \in IC(\theta_j)) = 1-\alpha$ *est un intervalle de confiance de* θ_j *de niveau* $1-\alpha$*.*

Application numérique au niveau de confiance $1-\alpha$*:*

- $IC(\theta_{DT}) = [0.006; 0.011]$*; 0 n'appartient pas à l'IC, la variable* `difftime` *est significative (risque 5%).*
- $IC(\theta_{DP}) = [-0.02; 0.008]$*; 0 appartient à l'IC ,la variable* `diffpct` *n'est pas significative (risque de seconde espèce inconnu).*

(c) Estimer par intervalle de confiance de niveau 95% l'espérance moyenne de la profondeur de trempe quand `soaktime=0.6`, `soakpct=1`, `difftime=0.6`, `diffpct=0.9`, puis prédire par intervalle de confiance de niveau 95% une valeur individuelle de profondeur de trempe sous les mêmes conditions. Pouvez-vous faire l'application numérique?

Correction. *Soit* $x = [1\ a\ b\ c\ d]$ *une nouvelle condition d'expérience. Comme* $x\widehat{\theta} \sim \mathcal{N}(x\theta, \sigma^2 x(X'X)^{-1}x')$*, on déduit un IC de* $\mathbb{E}(Y|x) = x\theta$ *de niveau* $1-\alpha$

$$IC(x\theta) = \left[x\widehat{\theta} - q^{\mathcal{T}_{n-p}}_{1-\alpha/2}\widehat{\sigma}_x;\ x\widehat{\theta} + q^{\mathcal{T}_{n-p}}_{1-\alpha/2}\widehat{\sigma}_x\right] \text{ avec } \widehat{\sigma}_x = \widehat{\sigma}\sqrt{x(X'X)^{-1}x'}$$

Pour une valeur individuelle, il ne faut pas oublier d'ajouter la variabilité naturelle de Y*:*

$$IC(\mathbb{E}(Y^{new})) = \left[x\widehat{\theta} - q^{\mathcal{T}_{n-p}}_{1-\alpha/2}\tilde{\sigma}_x;\ x\widehat{\theta} + q^{\mathcal{T}_{n-p}}_{1-\alpha/2}\tilde{\sigma}_x\right] \text{ avec } \tilde{\sigma}_x = \widehat{\sigma}\sqrt{x(X'X)^{-1}x' + 1}$$

Il manque la valeur observée de la matrice de variance de $\widehat{\theta}$ *pour faire l'application numérique. Nous n'avons que la racine carrée de sa diagonale (les* s_j*).*

Un stagiaire calcule les intervalles observés suivants: $[0.016; 0.02]$ pour l'IC de la moyenne et $[0.013; 0.023]$ pour l'intervalle de prédiction. Cela vous semble-t-il cohérent?

Correction. *Non, l'intervalle de prédiction doit être plus large que l'intervalle de la valeur moyenne.*

(d) Rappeler la construction du test de significativité globale du modèle iid (constant) contre le modèle d'étude (modèle complet).

Correction. *On teste (H_0) (modèle iid, tous les coefficients sont nuls sauf μ) contre (H_1) (au moins un des coefficients hors μ n'est pas nul). La statistique de test de Fisher suit sous (H_0) une loi de Fisher:*

$$F = \frac{(SCR(\widehat{\theta}_{H_0}) - SCR(\widehat{\theta}_{H_1}))/(p-1)}{SCR(\widehat{\theta}_{H_1})/(n-p)} \sim \mathcal{F}(p-1=4, n-p=32-5=27)$$

La région de rejet de niveau α est $\{F > q_{1-\alpha}^{\mathcal{F}(p-1,n-p)}\}$

Vérifier le calcul de la statistique de test à partir du tableau d'analyse de variance.

	Res.Df	SCR	Df	SCM	F
Modèle constant	31	0.00433			
Modèle complet	27	0.00014	4	0.00419	202

Correction. *On lit $SCR(H_1) = 0.0014$; $SCR(H_0) = 0.00433$; $SCM = SCR(H_0) - SCR(H_1) = 0.00419$, $F = (SCM/(5-1))/(SCR/(32-5)) = 202 > q_{1-\alpha}^{\mathcal{F}(4,27)} = 2.73$. On rejette la non significativité. Le modèle avec les 4 variables (complet) explique mieux les données que le modèle constant. On choisit donc le modèle complet, avec un risque de première espèce de 5%.*

Déterminer une estimation ponctuelle de la variance du bruit dans le modèle complet.

Correction. *Dans le modèle complet, l'estimation de σ^2 est $\widehat{\sigma}^2 = SCR(H_1)/(32-5) = 5.19e-06$.*

(e) Le coefficient de détermination vaut $R^2 = 0.96767$. Les graphes des résidus sont représentés en annexe. Commenter la validation de ce modèle.

Correction. *Le $R^2 = SCM/SCR(H_0)$ avec $SCM = SCR(H_0) - SCR(H_1)$, où (H_0) est le modèle constant, et (H_1) le modèle complet. Le R^2 est proche de 1, ce qui reflète un très bon ajustement. Ceci est confirmé dans l'étude des résidus.*

- *les résidus bruts sont très faibles parce que les valeurs observées sont elles-mêmes proches de 0 et estimées avec une grande précision.*
- *les résidus studentisés ne montrent pas de dépendance ou de tendance, ils sont raisonnablement compris entre -2 et 2, sauf deux points*
- *le qqplot indique un bon ajustement de la loi des résidus à une loi gaussienne.*

Retrouver la statistique de Fisher de significativité globale en fonction du R^2.

Correction. *On a par Pythagore: $SCR(H_1) = SCR(H_0) - SCM$. D'où*

$$F = \frac{SCM/(p-1)}{SCR(H_1)/(n-p)} = \frac{SCM/(p-1)}{(SCR(H_0) - SCM/(n-p)} = \frac{R^2}{1-R^2}\frac{n-p}{p-1}$$

La statistique de test fait un compromis entre une très bonne qualité d'ajustement (R^2 proche de 1) qui fait croître la statistique de test et un nombre important de variables explicatives p qui la fait diminuer.

(f) On considère le modèle (M2) qui ne contient que les deux variables `soaktime` et `difftime`. Compléter le tableau d'analyse de la variance. Quel modèle retenez-vous?

	Res.Df	SCR	Df	SCM	F
M2	??a	0.000147			
complet	??b	??c	??d	??e	??f

Correction. *Il s'agit d'un test de modèle emboîté: test de Fisher de sous modèle entre* (H_0)*: M2:* $\theta_{SP} = 0$ *et* $\theta_{DP} = 0$*contre* (H_1) *: M:* $\theta_{SP} \neq 0$ *ou* $\theta_{DP} \neq 0$.

Le nombre de degrés de liberté des résidus de $(M2)$ *est* $a = n - 3 = 29$*. On a vu que* $b = 32 - 5 = 27$ *et* $c = 0.00014$.

La somme des carrés modèle est $SCM = SCR(M2) - SCR(complet) = 0.000147 - c = e = 7.e - 06$ *et son nombre de degrés de liberté est* $d = a - b = 2$*. Sous* (H_0)*, la stat de Fisher suit une loi de Fisher de paramètres* (d, b) *et vaut* $(e/d)/(c/b) = 0.675 < q_{1-\alpha^{\mathcal{F}(2,27)}} = 3.35$*. La statistique observée n'appartient pas à la région de rejet, on ne peut rejeter le modèle* $(M2)$ *qu'on conserve par défaut avec un risque de seconde espèce inconnu. Les variables* `soakpct` *et* `diffpct` *ne sont pas significatives.*

La pvalue observée vaut $\mathbb{P}(F > 0.675) = 0.52 > 5\%$*, donc on conserve M2 avec un risque de seconde espèce inconnu:*

Calculer le coefficient de détermination de $(M2)$.

Correction. $R^2 = SCM(constant, M2)/SCR(constant)$*. Application numérique:* $(0.00433 - 0.0001467)/0.00433 = 0.966$ *très légèrement inférieur à celui du modèle complet qui vaut* 0.9677.

On note que le fait de supprimer deux variables n'a quasiment pas d'impact sur la qualité d'ajustement. Ce qui va bien dans le sens qu'elles ne sont pas explicatives.

(g) On considère maintenant les variables `x1=soaktime*soakpct` et `x2=difftime*diffpct`
Un modèle utilisant ces variables est-il toujours linéaire?

Correction. *La linéarité étant en le paramètre, le modèle est toujours linéaire même en considérant ces variables.*

$$Y \sim \mathcal{N}(\mu + \beta_1\, x_1 + \beta_2\, x_2; \sigma^2 Id_n)$$

L'ajustement du modèle (M3) n'utilisant que ces deux variables (et l'intercept) donne les résultats suivants:

```
Coefficients:
Parameter          Estimate Std_Error  t_value
Intercept           0.01089    0.00112    9.72
soaktime * soakpct  0.00269    0.00015   17.93
difftime * diffpct  0.00932    0.00146    6.38

Residual standard error: 0.002581 on ?? degrees of freedom
R2:  0.9553
```

Rappeler la définition de l'erreur standard résiduelle.

Correction. $\sqrt{\widehat{\sigma}^2} = \sqrt{SCR(M3)/(n - dim(M3))} = 0.002581$. *Le nombre de degrés de liberté est* $n - dim(M3) = 32 - 3 = 29$

Est-il possible de tester ce modèle contre les autres?

Correction. *On ne peut pas tester avec le test de Fisher car les modèles ne sont pas emboîtés. Mais on peut comparer les R2 entre deux modèles de même dimension (ie M2 et M3), et on choisit M2: son R2 est (légèrement) meilleur (mais attention à la variabilité due à l'échantillonnage), et ses variables explicatives sont plus simples.*

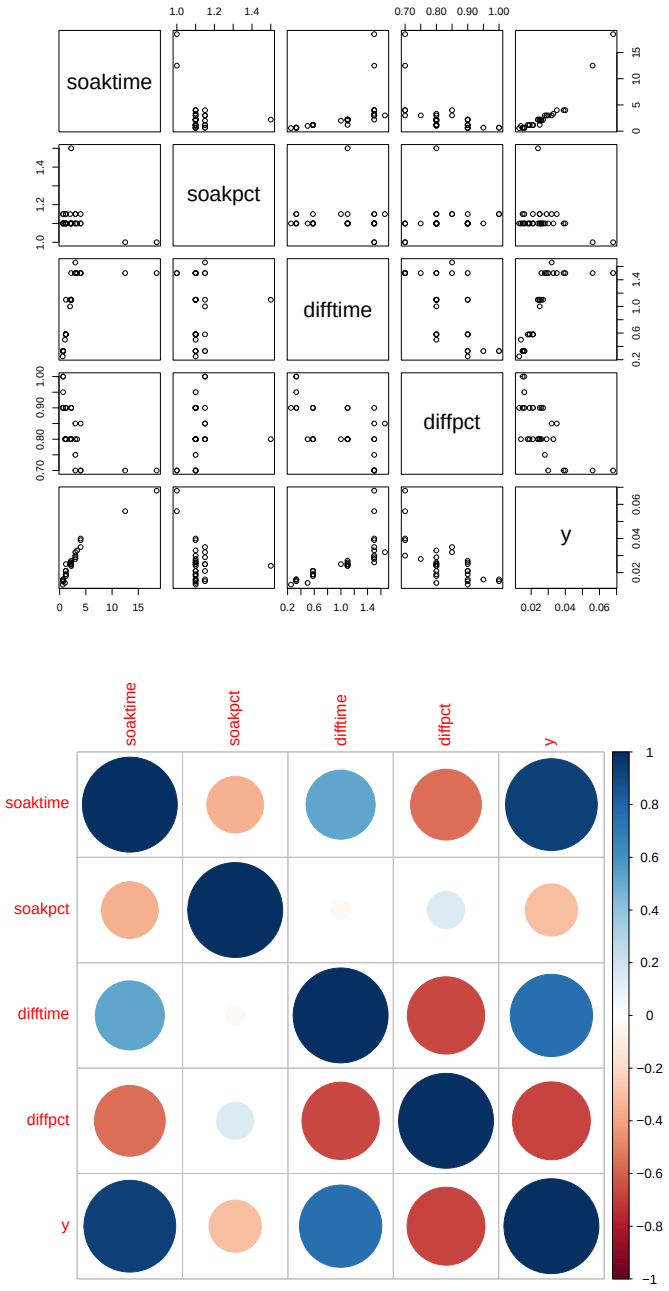

Figure 1: En haut: scatter plot des données - En bas: représentation des corrélations