Régression Linéaire Multiple

N. Jégou

Université Rennes 2

M1 2SEP

Les données ozone

Individu	maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v
20010601	87	15.6	18.5	18.4	4	4	8	0.7	-1.7	-0.7	84
20010602	82	17.0	18.4	17.7	5	5	7	-4.3	-4.0	-3.0	87
20010603	92	15.3	17.6	19.5	2	5	4	3.0	1.9	0.5	82
20010604	114	16.2	19.7	22.5	1	1	0	1.0	0.3	-0.2	92
20010605	94	17.4	20.5	20.4	8	8	7	-0.5	-3.0	-4.3	114
20010606	80	17.7	19.8	18.3	6	6	7	-5.6	-5.0	-6.0	94
20010607	79	16.8	15.6	14.9	7	8	8	-4.3	-1.9	-3.88	80
20010610	79	14.9	17.5	18.9	5	5	4	0.0	-1.0	-1.4	99
20010611	101	16.1	19.6	21.4	2	4	4	-0.8	-1.0	-2.3	79
20010612	106	18.3	21.9	22.9	5	6	8	1.3	-2.3	-3.9	101

Au total $n = 112$ mesures.

Objectifs

- Des variables quantitatives :
  - Y : `max03` variable à expliquer
  - X_1 =`T9`, X_2 =`T12`,...,X_p =`max03v` :
    $p = 10$ variables explicatives
- Expliquer le pic d'ozone Y par les variables $X_1, \ldots, X_p$
- On dispose de données :

$$(Y_i, x_{i1}, \ldots, x_{ip})_{i=1,\ldots,n} \in \mathbb{R}^{p+1}$$

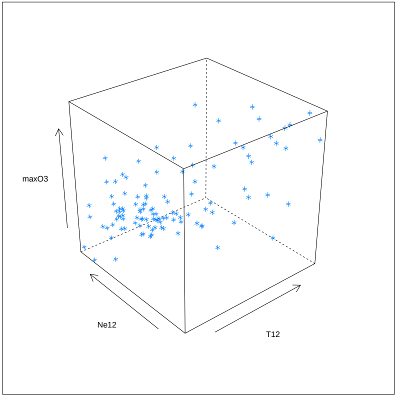
- $\rightarrow$ Prévoir pics de pollution
- $\rightarrow$ Déclencher des mesures de santé publique

Graphes

Les données $\in \mathbb{R}^{p+1}$: impossible à représenter (pour $p > 2$)

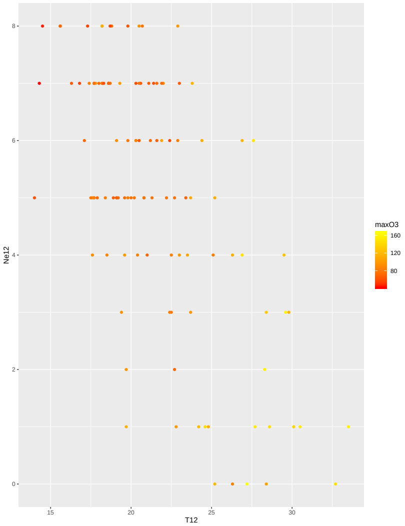
Graphes

Pour $p = 2$, représentation possible en 3 dimensions :



Graphes

Pour $p = 2$, représentation possible en 3 dimensions :



Le modèle linéaire

- On généralise les idées vues avec une variable
- On suppose l'existence d'une fonction $f : \mathbb{R}^p \to \mathbb{R}$ telle que

$$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon$$

- Modèle linéaire : f est supposée linéaire

$$\exists\ (\beta_0, \beta_1, \ldots, \beta_p) \in \mathbb{R}^{p+1} : Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$$

- f est caractérisée par le paramètre $(\beta_0, \beta_1, \ldots, \beta_p)$: modèle paramétrique

Le modèle linéaire

- On dispose de données dont on suppose qu'elles sont des réplications de ce modèle
- Y et ε sont supposées aleatoires
- Les $X_1, \ldots, X_p$ sont non aléatoires

- D'où l'écriture (en ligne) du modèle

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, \cdots, n$$

Le modèle linéaire

Ecriture matricielle :

$$Y = X\beta + \varepsilon,$$

avec

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & X_{11} & \dots & X_{1p} \\ 1 & X_{21} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \dots & X_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \text{ et } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Hypothèses

- Hypothèses sur le design :
  - $\mathcal{H}_1$: X de rang $p+1$ (de plein rang)
    i.e. Deux colonnes de X ne sont pas (parfaitement) corrélées

    $\rightarrow$ unicité de $\hat{\beta}$

- Hypothèses sur le résidu ε :
  - $\mathcal{H}_2$: $\mathbb{E}(\varepsilon) = 0 \qquad \mathrm{var}(\varepsilon) = \sigma^2 I_n$
    centré / non corrélation des ε_i
  - $\mathcal{H}_3$: $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$
    $\rightarrow$ IC, tests

Estimateur des Moindres Carrés Ordinaires

- Le principe généralise celui de la régression simple
- On minimise le coût quadratique

$$S(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$$

- On cherche $(\beta_0, \beta_1, \ldots, \beta_p)$ tel que $S(\beta_0, \beta_1, \ldots, \beta_p)$ minimum

Estimateur des Moindres Carrés Ordinaires

- L'estimateur des MCO (définition)

$$(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p) = \text{argmin}_{(\beta_0,\beta_1,\ldots,\beta_p)\in\mathbb{R}^{p+1}} S(\beta_0, \beta_1, \ldots, \beta_p)$$

- L'expression (matricielle) est :

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Estimateur des Moindres Carrés Ordinaires

- Valeurs ajustées (prévues aux points du design) :

$$\hat{Y} = X\hat{\beta}$$

Soit, pour $i = 1, \ldots, n$

$$\hat{Y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}$$

- Estimation des résidus :

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$$

- Estimateur de σ^2 : moyenne des résidus estimés

$$\hat{\sigma}^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

Propriétés de l'estimateur

- Estimateur $\hat{\beta}$ sans biais :

$$\forall\ j = 0, \ldots, p \qquad \mathbb{E}[\hat{\beta}_j] = \beta_j$$

- Estimateur de faible variance :

$$\mathrm{var}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

- Estimateur $\hat{\sigma}^2$ sans biais

$$\mathbb{E}[\hat{\sigma}^2] = \sigma^2$$

Intervalles de confiance

- Sous l'hypothèse $\mathcal{H}_3$, on a la loi des $\hat{\beta}_j$:

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 (X'X)^{-1}_{jj})$$

- On en déduit des IC sur les paramètres inconnus β_j :

$$\hat{\beta}_j - t_{1-\alpha/2}(n-p)\sqrt{\hat{\mathrm{var}}(\hat{\beta}_j)} \leq \beta_j \leq \hat{\beta}_j + t_{1-\alpha/2}(n-p)\sqrt{\hat{\mathrm{var}}(\hat{\beta}_j)}$$

avec la variance estimée

$$\hat{\mathrm{var}}(\hat{\beta}_j) = \hat{\sigma}^2 (X'X)^{-1}_{jj}$$

Test de nullité d'un paramètre β_j

- On fixe l'erreur de première espèce
- Deux hypothèses pour formaliser l'influence (ou non) de X_j :

$$\mathcal{H}_0 : \beta_j = 0 \qquad \text{vs} \qquad \mathcal{H}_1 : \beta_j \neq 0$$

- La statistique de test :

$$T = \frac{\hat{\beta}_j}{\hat{\text{var}}(\hat{\beta}_j)} \overset{\mathcal{H}_0}{\sim} \mathcal{T}(n-p)$$

- Observation de la stat. de test : T_{obs}
  $\rightarrow$ p-value
- Décision / Conclusion

Test de significativité globale

- C'est un test entre modèles emboîtés
- Les deux modèles envisagés sont
  Modèle $\mathbf{H}_1$ (complet) : $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p Xp + \varepsilon$
  Modèle $\mathbf{H}_0$: $Y = \beta_0 + \varepsilon$
- La statistique de test est une statistique de Fisher :

$$F = \frac{\|\hat{Y}_{\mathbf{H}_1} - \hat{Y}_{\mathbf{H}_0}\|^2/p}{\hat{\sigma}^2} \overset{\mathbf{H}_0}{\sim} \mathcal{F}(p, n-(p+1))$$

- Décision entre :
  - Tous les $\beta_j, j \geq 1$ sont nuls : $\mathbf{H}_0$
  - Au moins un des $\beta_j, j \geq 1$ non nul : $\mathbf{H}_1$

Le coin de R

Estimation dans le modèle complet :

```
> reg.complet <- lm(max03 ~ ., data = don)
> summary(reg.complet)

Call:
lm(formula = max03 ~ ., data = don)

Residuals:
    Min      1Q  Median      3Q     Max
-53.566  -8.727  -0.403   7.599  39.458

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.24442   13.47190   0.909   0.3656
T9          -0.01901    1.12515  -0.017   0.9866
T12          2.22115    1.43294   1.550   0.1243
T15          0.55853    1.14464   0.488   0.6266
Ne9         -2.18909    0.93824  -2.333   0.0216 *
Ne12        -0.42102    1.36766  -0.308   0.7588
Ne15         0.18373    1.00279   0.183   0.8550
Vx9          0.94791    0.91228   1.039   0.3013
Vx12         0.03120    1.05523   0.030   0.9765
Vx15         0.41859    0.91568   0.457   0.6486
max03v       0.35198    0.06289   5.597 1.88e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.36 on 101 degrees of freedom
Multiple R-squared:  0.7638,Adjusted R-squared:  0.7405
F-statistic: 32.67 on 10 and 101 DF,  p-value: < 2.2e-16
```

Le coin de R

Intervalles de confiance sur les β_j :

```
> confint(reg.complet)
                  2.5 %     97.5 %
(Intercept) -14.4802083 38.9690481
T9           -2.2510136  2.2129851
T12          -0.6214249  5.0637287
T15          -1.7121182  2.8291799
Ne9          -4.0502993 -0.3278850
Ne12         -3.1340934  2.2920630
Ne15         -1.8055357  2.1729977
Vx9          -0.8618028  2.7576211
Vx12         -2.0620871  2.1244836
Vx15         -1.3978619  2.2350470
maxO3v        0.2272237  0.4767292
```

Le coin de R

Prédictions

```
> new.x <- data.frame(maxO3=88, T9=15, T12=18, T15=20,
  +  Ne9=3, Ne12=4, Ne15=4, Vx9=0.7, Vx12=1.3, Vx15=1.4, maxO3v=85)

> predict.lm(reg.complet, newdata = new.x)
       1
86.80227
```

Le coin de R

Test global du modèle

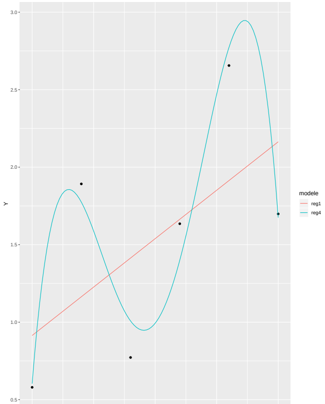
```
> reg0 <- lm(max03 ~ 1, data = don)
> anova(reg0, reg.complet)
Analysis of Variance Table

Model 1: max03 ~ 1
Model 2: max03 ~ T9 + T12 + T15 + Ne9 + Ne12 + Ne15 + Vx9 + Vx12 + Vx15 +
    max03v
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    111 88192
2    101 20827 10     67364 32.668 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Modèle complet = Meilleur modèle ?

- Illustration avec une variable X
- Deux modèles

```
Y ~ 1 + X (modele 1)
Y ~ 1 + X + X^2 + X^3 + X^4 (modele 4)
```

- Ajustements :



Intérêt d'un modèle parcimonieux

- Beaucoup de variables X dans le modèle :
  - Risque de surajustement : mauvais en prévision
  - Souvent des $\hat{\beta}_j$ non significativement $\neq 0$
  - Variables explicatives possiblement très corrélées
- Intérêt d'un modèle parcimonieux :
  - Retenir les variables significatives
  - Coût éventuellement moins élevé
  - Recherche de meilleurs prévisions

Le R^2 pour sélectionner un modèle ?

- Le R^2 mesure la qualité d'ajustement :

$$R^2 = \frac{\text{var}(\hat{Y})}{\text{var}(Y)}$$

- Problème : quand on ajoute une variable, le R^2 augmente
- Le critère du R^2 conduira à choisir le modèle complet

- $\Rightarrow$ ce n'est pas un bon critère

Autres idées ?

- Enlever les variables non significatives du modèle complet ?

  Problème : le fait d'enlever une seule variable modifie p-values et estimation des paramètres $\Rightarrow$ sélection pas à pas ?

- Tests emboités sur tous les modèles possibles ?

  Problème : beaucoup trop de modèles possibles

Sélection de modèle par critères pénalisés

- Dilemme :
  - Ajouter des variables : $\uparrow$ la qualité d'ajustement
  - $\Rightarrow$ risque de surajustement

- Les critères pénalisés résultent d'un compromis entre
  - Un terme favorisant la qualité d'ajustement
  - Un terme pénalisant le nombre de variables

Sélection de modèle par critères pénalisés

- Terme favorisant la qualité d'ajustement :

$$\phi_1\left(\sum(y_i - \hat{y}_i)^2\right) \qquad \phi_1 \searrow$$

- Terme pénalisant le nombre de variables :

$$\phi_2\,(\text{nb. de variables}) \qquad \phi_2 \nearrow$$

- Compromis

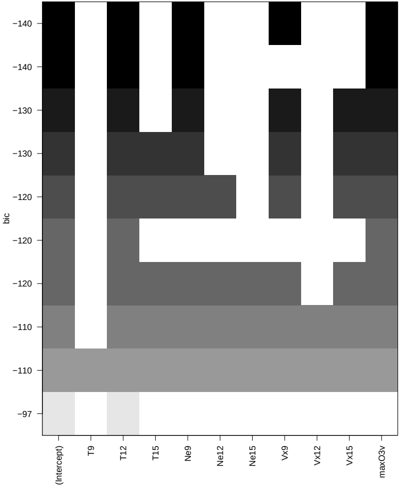
$$\text{argmin}\left\{\phi_1\left(\sum(y_i - \hat{y}_i)^2\right) + \phi_2\,(\text{nb. de variables})\right\}$$

- Selon la fonction ϕ_2, on a les critères : AIC, BIC, Cp,...

Mise en œuvre avec R

```
> library(leaps)
> critere.penalise <- regsubsets(max03~., int=T, nbest=1,
+       nvmax=10, method = "exhaustive",
+       really.big = T, data = don)
> plot(critere.penalise, scale = "bic")
```

Mise en œuvre avec R



Comparaison de modèles : Apprentissage - Validation

- Idée : se mettre en situation de prévision et comparer les performances des modèles
- Méthodologie
  - On sépare les données en deux échantillons : Apprentissage / Validation
  - On estime les $\hat{\beta}_j$ du modèle sur l'échantillon d'apprentissage
  - On déduit les prévisions $\hat{Y}^*$ sur l'échantillon de validation
  - On calcule une erreur de prévision en comparant aux Y^* réellement observés sur l'échantillon de validation :

$$||Y^* - \hat{Y}^*||^2 = \frac{1}{n_{\text{validation}}} \sum_{i=1}^{n_{\text{validation}}} (Y_i^* - \hat{Y}_i^*)^2$$

- Nécessite suffisament de données

Comparaison de modèles : Apprentissage - Validation