Section 5 :
Inférence sur β et σ^2

STAT-F-406
Master en sciences mathématiques, Master en statistique

ACTU-F4001
Master en sciences actuarielles

Davy Paindaveine

Université libre de Bruxelles

2024–2025

Contenu du chapitre

Inférence sur β_j

Inférence sur σ^2

Tests d'hypothèses linéaires sur β

Inférence sur β_j: problèmes considérés

Pour $j \in \{1, \ldots, k\}$ fixé, on veut construire ici

(i) des intervalles de confiance (IC) pour β_j, et

(ii) des tests (au niveau $\alpha \in]0, 1[$) pour

$$\begin{cases} \mathcal{H}_0 : \beta_j = c \\ \mathcal{H}_1 : \beta_j \neq c, \end{cases}$$

où c est un réel fixé.

On considérera

- des procédures exactes (valables pour tout $n > k$ fixé, même petit),
- des procédures asymptotiques (valables seulement pour n "grand").

Les premières requièrent le modèle fort avec normalité, tandis que les secondes ne requièrent que le modèle semi-fort homoscédastique.

Procédures exactes

Dans le modèle fort avec normalité, le lemme de Fisher

(i) $\hat{\beta}|\mathcal{X} \sim \mathcal{N}(\beta, \frac{\sigma^2}{n} Q^{-1})$

(ii) $\frac{n\hat{\sigma}^2}{\sigma^2}|\mathcal{X} \sim \chi^2_{n-k}$

(iii) $\hat{\beta} \perp\!\!\!\perp \hat{\sigma}^2$ conditionnellement à $\mathcal{X}$

implique que, conditionnellement à $\mathcal{X}$,

$$\frac{\sqrt{n-k}(\hat{\beta}_j - \beta_j)}{\hat{\sigma}\sqrt{(Q^{-1})_{jj}}} = \frac{\frac{\sqrt{n}(\hat{\beta}_j - \beta_j)}{\sigma\sqrt{(Q^{-1})_{jj}}}}{\sqrt{(\frac{n\hat{\sigma}^2}{\sigma^2})/(n-k)}} \sim t_{n-k},$$

donc aussi inconditionnellement.

IC exacts

Pour tout $\alpha \in \,]0,1[$, on a donc

$$P\left[-\,t_{n-k,1-\alpha/2} \leq \frac{\sqrt{n-k}(\hat{\beta}_j - \beta_j)}{\hat{\sigma}\sqrt{(Q^{-1})_{jj}}} \leq t_{n-k,1-\alpha/2}\right] = 1-\alpha,$$

ce qui livre

$$P\left[\hat{\beta}_j - \frac{t_{n-k,1-\alpha/2}}{\sqrt{n-k}}\hat{\sigma}\sqrt{(Q^{-1})_{jj}} \leq \beta_j \leq \hat{\beta}_j + \frac{t_{n-k,1-\alpha/2}}{\sqrt{n-k}}\hat{\sigma}\sqrt{(Q^{-1})_{jj}}\right] = 1-\alpha.$$

Ceci est un IC pour β_j au niveau de confiance $1-\alpha$.

Cet IC est exact: il est valide pour tout $n(>k)$ fixé.

IC exacts

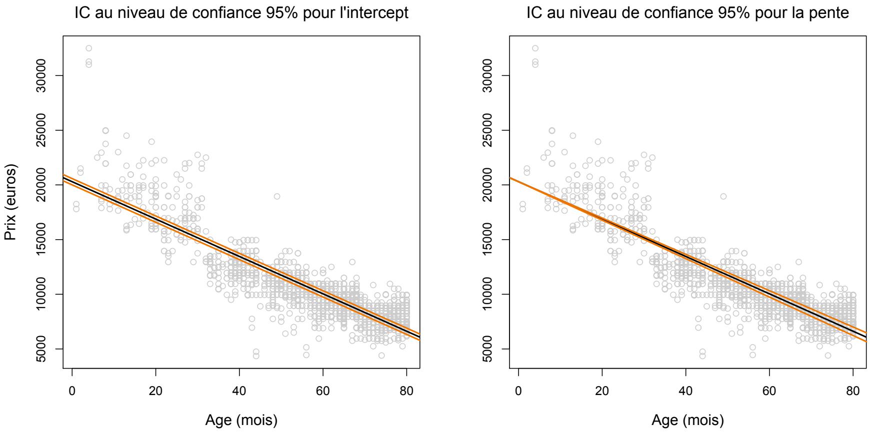
```
Call:
lm(formula = Price ~ Age)

Residuals:
    Min      1Q  Median      3Q     Max 
-8423.0  -997.4   -24.6   878.5 12889.7 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 20294.059    146.097  138.91   <2e-16 ***
Age          -170.934      2.478  -68.98   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1746 on 1434 degrees of freedom
Multiple R-squared:  0.7684,	Adjusted R-squared:  0.7682 
F-statistic:  4758 on 1 and 1434 DF,  p-value: < 2.2e-16
```

IC exacts


IC au niveau de confiance 95% pour l'intercept
IC au niveau de confiance 95% pour la pente
Prix (euros)
5000
10000
15000
20000
25000
30000
0
20
40
60
80
Age (mois)

Tests exacts

Pour le problème de test

$$\begin{cases} \mathcal{H}_0 : \beta_j = c \\ \mathcal{H}_1 : \beta_j \neq c, \end{cases}$$

le test naturel au niveau α rejette $\mathcal{H}_0$ si et seulement si

$$|T| > t_{n-k,1-\alpha/2}, \qquad \text{où } T \stackrel{\text{def}}{=} \frac{\sqrt{n-k}(\hat{\beta}_j - c)}{\hat{\sigma}\sqrt{(Q^{-1})_{jj}}}.$$

Les logiciels statistiques retournent le plus souvent la p-valeur

$$P_{\mathcal{H}_0}\left[|T| > |T_{\text{observé}}|\right],$$

et on rejette au niveau α si et seulement si cette p-valeur est $< \alpha$.

Tests exacts

```
Call:
lm(formula = Price ~ Age + KM + HP + CC + Weight + Doors + Automatic +
    FuelType + MetColor)

Residuals:
     Min       1Q   Median       3Q      Max
-10642.3   -737.7      3.1    731.3   6451.5

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -3.801e+03  1.304e+03  -2.915 0.003613 **
Age            -1.220e+02  2.602e+00 -46.889  < 2e-16 ***
KM             -1.621e-02  1.313e-03 -12.347  < 2e-16 ***
HP              6.081e+01  5.756e+00  10.565  < 2e-16 ***
CC             -4.174e+00  5.453e-01  -7.656 3.53e-14 ***
Weight          2.001e+01  1.203e+00  16.629  < 2e-16 ***
Doors          -7.776e+00  4.006e+01  -0.194 0.846129
Automatic       3.303e+02  1.571e+02   2.102 0.035708 *
FuelTypeDiesel  3.390e+03  5.188e+02   6.535 8.86e-11 ***
FuelTypePetrol  1.121e+03  3.324e+02   3.372 0.000767 ***
MetColor        5.716e+01  7.494e+01   0.763 0.445738
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1316 on 1425 degrees of freedom
Multiple R-squared:  0.8693,  Adjusted R-squared:  0.8684
F-statistic:   948 on 10 and 1425 DF,  p-value: < 2.2e-16
```

Tests exacts

Bien entendu, on peut définir de la même manière des tests exacts pour les problèmes unilatéraux

$$\begin{cases} \mathcal{H}_0 : \beta_j \leq c \\ \mathcal{H}_1 : \beta_j > c \end{cases} \quad \text{et} \quad \begin{cases} \mathcal{H}_0 : \beta_j \geq c \\ \mathcal{H}_1 : \beta_j < c, \end{cases}$$

pour lesquels on rejette $\mathcal{H}_0$ seulement pour les grandes valeurs de T et seulement pour les petites valeurs de T, respectivement.

Procédures asymptotiques

Dans le modèle semi-fort homoscédastique, les résultats

(i) $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{D}} \mathcal{N}_k\big(0, \sigma^2(\mathrm{E}[XX'])^{-1}\big)$

(ii) $\hat{\sigma}^2 \xrightarrow{\mathrm{P}} \sigma^2$

impliquent que

$$\begin{aligned} \frac{\sqrt{n}(\hat{\beta}_j - \beta_j)}{\hat{\sigma}\sqrt{(Q^{-1})_{jj}}} &= \frac{\sigma\sqrt{((\mathrm{E}[XX'])^{-1})_{jj}}}{\hat{\sigma}\sqrt{(Q^{-1})_{jj}}} \times \frac{\sqrt{n}(\hat{\beta}_j - \beta_j)}{\sigma\sqrt{((\mathrm{E}[XX'])^{-1})_{jj}}} \\ &\xrightarrow{\mathcal{D}} 1 \times \mathcal{N}(0,1) = \mathcal{N}(0,1), \end{aligned}$$

par le lemme de Slutsky.

IC asymptotiques

En procédant comme ci-dessus, on obtient alors que, quand $n \to \infty$,

$$P\left[\hat{\beta}_j - \frac{z_{\alpha/2}}{\sqrt{n}}\hat{\sigma}\sqrt{(Q^{-1})_{jj}} \leq \beta_j \leq \hat{\beta}_j + \frac{z_{\alpha/2}}{\sqrt{n}}\hat{\sigma}\sqrt{(Q^{-1})_{jj}}\right] \to 1 - \alpha.$$

Ceci livre donc un IC asymptotique pour β_j au niveau de confiance $1 - \alpha$.

On vérifiera que, quand $n \to \infty$, les IC exacts et asymptotiques coïncident (de sorte qu'ils fournissent la même solution dans le modèle fort avec normalité pour n grand).

C'est la raison pour laquelle les logiciels statistiques ne retournent que (les quantités qui permettent de calculer) les IC exacts.

Tests asymptotiques

Pour le problème de test

$$\begin{cases} \mathcal{H}_0 : \beta_j = c \\ \mathcal{H}_1 : \beta_j \neq c, \end{cases}$$

le test asymptotique au niveau α rejette $\mathcal{H}_0$ si et seulement si

$$\left| \frac{\sqrt{n}(\hat{\beta}_j - c)}{\hat{\sigma}\sqrt{(Q^{-1})_{jj}}} \right| > z_{\alpha/2}.$$

De nouveau, on vérifiera que, quand $n \to \infty$, les tests exacts et asymptotiques coïncident.

Inférence sur σ^2: problèmes considérés

Ici, on veut construire

(i) des intervalles de confiance (IC) pour σ^2, et

(ii) des tests (au niveau $\alpha \in]0,1[$) pour

$$\begin{cases} \mathcal{H}_0 : \sigma^2 = c \\ \mathcal{H}_1 : \sigma^2 \neq c, \end{cases}$$

où c est un réel positif fixé.

Dans le modèle fort avec normalité, cela peut être fait de façon exacte en utilisant le fait que, conditionnellement à $\mathcal{X}$,

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-k}$$

donc aussi inconditionnellement.

IC et tests exacts

Pour tout $\alpha \in]0,1[$, on a donc

$$P\left[\chi^2_{n-k,\alpha/2} \leq \frac{n\hat{\sigma}^2}{\sigma^2} \leq \chi^2_{n-k,1-\alpha/2}\right] = 1-\alpha,$$

ce qui livre

$$P\left[\frac{n\hat{\sigma}^2}{\chi^2_{n-k,1-\alpha/2}} \leq \sigma^2 \leq \frac{n\hat{\sigma}^2}{\chi^2_{n-k,\alpha/2}}\right] = 1-\alpha.$$

Ceci est un IC pour σ^2 au niveau de confiance $1-\alpha$.

Le test naturel pour le problème bilatéral ci-dessus rejette $\mathcal{H}_0$ au niveau α si et seulement si

$$\frac{n\hat{\sigma}^2}{c} \notin [\chi^2_{n-k,\alpha/2}, \chi^2_{n-k,1-\alpha/2}].$$

On peut évidemment construire des tests unilatéraux de façon similaire.

IC et tests asymptotiques

Dans le modèle semi-fort homoscédastique, on pourrait fonder des procédures asymptotiques sur le fait que

$$\sqrt{n}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathrm{Var}[\varepsilon^2]),$$

pour autant que $\mathrm{E}[\varepsilon^4] < \infty$ (exercice!)

Tests d'hypothèses linéaires sur β

Ici, on veut construire des tests (au niveau $\alpha \in]0,1[$) pour

$$\begin{cases} \mathcal{H}_0 : R\beta = r \\ \mathcal{H}_1 : R\beta \neq r, \end{cases}$$

où R est une matrice $p \times k$ ($p \leq k$) fixée et où $r \in \mathbb{R}^p$ est fixé.

On supposera que R est de rang maximal p (pourquoi?)

Le cas particulier le plus important: $\mathcal{H}_0 : \beta_2 = \ldots = \beta_k = 0$ (β_1 correspondant à l'intercept), ce qui est obtenu pour $R = (0\ I_{k-1})$ et $r = 0$.

Test exact

Rappelons que, dans le modèle fort avec normalité,

(i) $\hat{\beta}|\mathcal{X} \sim \mathcal{N}_k(\beta, \frac{\sigma^2}{n}Q^{-1})$

(ii) $\frac{n\hat{\sigma}^2}{\sigma^2}|\mathcal{X} \sim \chi^2_{n-k}$

(iii) $\hat{\beta} \perp\!\!\!\perp \hat{\sigma}^2$ conditionnellement à $\mathcal{X}$.

Donc $R\hat{\beta}|\mathcal{X} \sim \mathcal{N}_p(R\beta, \frac{\sigma^2}{n}RQ^{-1}R')$, ce qui livre, sous $\mathcal{H}_0 : R\beta = r$,

$$R\hat{\beta} - r|\mathcal{X} \sim \mathcal{N}_p(0, \frac{\sigma^2}{n}RQ^{-1}R').$$

Par conséquent, $(R\hat{\beta} - r)'(\frac{\sigma^2}{n}RQ^{-1}R')^{-1}(R\hat{\beta} - r)|\mathcal{X} \sim \chi^2_p$, de sorte que

$$W \stackrel{\text{def}}{=} \frac{(R\hat{\beta} - r)'(\frac{\sigma^2}{n}RQ^{-1}R')^{-1}(R\hat{\beta} - r)/p}{(\frac{n\hat{\sigma}^2}{\sigma^2})/(n-k)}|\mathcal{X} \sim F_{p,n-k},$$

Test exact

Puisque ceci implique que

$$W = \frac{(n-k)(R\hat{\beta}-r)'(RQ^{-1}R')^{-1}(R\hat{\beta}-r)}{p\hat{\sigma}^2} \sim F_{p,n-k},$$

le test qui en résulte rejette $\mathcal{H}_0 : R\beta = r$ au niveau α si et seulement si

$$W > F_{p,n-k,1-\alpha}.$$

Les logiciels statistiques retournent le plus souvent la p-valeur

$$P_{\mathcal{H}_0}\left[W > W_{\text{observé}}\right],$$

et on rejette au niveau α si et seulement si cette p-valeur est $< \alpha$.

$$\mathcal{H}_0 : \beta_2 = \ldots = \beta_k = 0 \; (k = 11, p = k - 1, n = 1436)$$

```
Call:
lm(formula = Price ~ Age + KM + HP + CC + Weight + Doors + Automatic +
    FuelType + MetColor)

Residuals:
     Min       1Q   Median       3Q      Max
-10642.3   -737.7      3.1    731.3   6451.5

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -3.801e+03  1.304e+03  -2.915 0.003613 **
Age            -1.220e+02  2.602e+00 -46.889  < 2e-16 ***
KM             -1.621e-02  1.313e-03 -12.347  < 2e-16 ***
HP              6.081e+01  5.756e+00  10.565  < 2e-16 ***
CC             -4.174e+00  5.453e-01  -7.656 3.53e-14 ***
Weight          2.001e+01  1.203e+00  16.629  < 2e-16 ***
Doors          -7.776e+00  4.006e+01  -0.194 0.846129
Automatic       3.303e+02  1.571e+02   2.102 0.035708 *
FuelTypeDiesel  3.390e+03  5.188e+02   6.535 8.86e-11 ***
FuelTypePetrol  1.121e+03  3.324e+02   3.372 0.000767 ***
MetColor        5.716e+01  7.494e+01   0.763 0.445738
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1316 on 1425 degrees of freedom
Multiple R-squared:  0.8693,  Adjusted R-squared:  0.8684
F-statistic:   948 on 10 and 1425 DF,  p-value: < 2.2e-16
```

Test asymptotique

Dans le modèle semi-fort homoscédastique, les résultats

(i) $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{D}} \mathcal{N}_k\big(0, \sigma^2(\mathrm{E}[XX'])^{-1}\big)$

(ii) $\hat{\sigma}^2 \xrightarrow{\mathrm{P}} \sigma^2$

impliquent que, sous $\mathcal{H}_0 : R\beta = r$,

$$\sqrt{n}(R\hat{\beta} - r) \xrightarrow{\mathcal{D}} \mathcal{N}_k\big(0, \sigma^2 R(\mathrm{E}[XX'])^{-1}R'\big),$$

ce qui livre

$$\frac{n}{\sigma^2}(R\hat{\beta} - r)'(R(\mathrm{E}[XX'])^{-1}R')^{-1}(R\hat{\beta} - r) \xrightarrow{\mathcal{D}} \chi^2_p.$$

Par le lemme de Slutsky, on a donc que, sous $\mathcal{H}_0 : R\beta = r$,

$$W_{\mathrm{asympt}} \stackrel{\mathrm{def}}{=} \frac{n(R\hat{\beta} - r)'(RQ^{-1}R')^{-1}(R\hat{\beta} - r)}{\hat{\sigma}^2} \xrightarrow{\mathcal{D}} \chi^2_p.$$

Test asymptotique

Le test asymptotique associé rejette $\mathcal{H}_0 : R\beta = r$ au niveau α si et seulement si

$$W_{\text{asympt}} > \chi^2_{p,1-\alpha}.$$

Quand $n \to \infty$, les tests exact et asymptotique coïncident (pourquoi?), ce qui est la raison pour laquelle les logiciels statistiques ne fournissent que les quantités exactes.