Modèle Linéaire - Régression

Estimation

Thierry Dhorne

Institut Universitaire de Technologie
Université de Bretagne Sud

Année Universitaire 2014-2015

Estimation des paramètres

- le modèle est toujours

$$Y_r = \beta_0 + \beta_1 x_r + E_r,\ \mathbb{V}(E_r) = \sigma_E^2$$

- il faut en estimer
  - les 2 paramètres d'espérance β_0 et β_1
  - le paramètre de variance σ_E^2
- les estimateurs de β_0 et de β_1 sont notés B_0 et B_1
- les estimations sont notées b_0 et b_1
- l'estimateur de σ_E^2 est notés S_E^2
- l'estimation est notée s_E^2

Estimateurs et critère

- on estime le plus souvent les paramètres d'espérance par la méthode des moindres carrés

★ il existe un lien avec l'approche probabiliste gaussienne

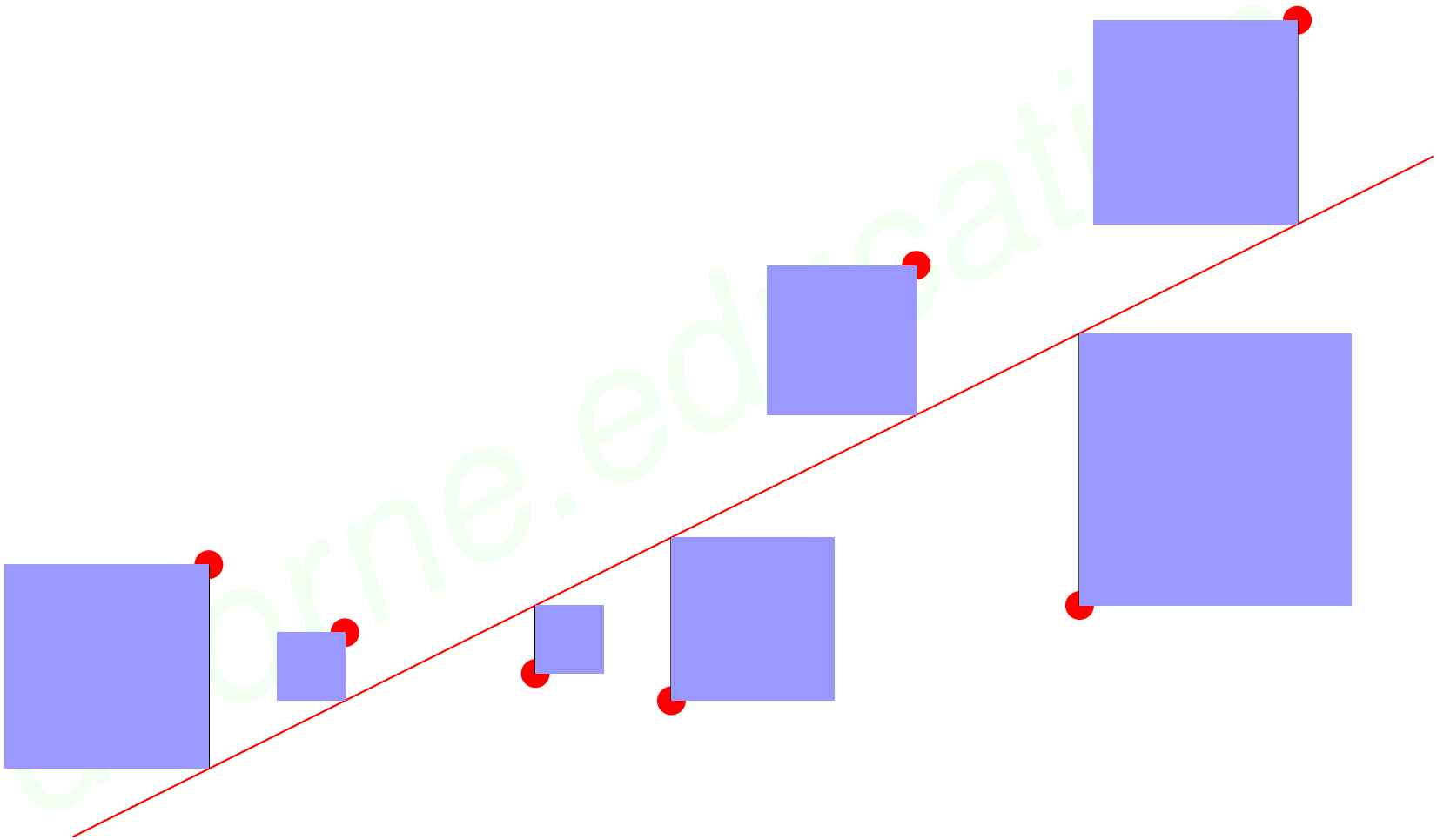
★ d'autres critères peuvent être utilisés

- le critère minimisé est

$$\sum_{r=1}^{n}(Y_r - \beta_0 - \beta_1 x_r)^2$$

- on estime la variance à partir des estimateurs de l'espérance

- on peut représenter graphiquement ce critère



Recherche des estimateurs B_0^{MC} et B_1^{MC}

- B_0^{MC} et B_1^{MC} sont solutions de la minimisation du critère

$$(B_0^{MC}, B_1^{MC}) = \arg_{B_0,B_1} \min \sum_{r=1}^{n} (Y_r - B_0 - B_1 x_r)^2$$

- c'est-à-dire des deux équations

$$\begin{aligned} \frac{\partial}{\partial B_0^{MC}} \sum_{r=1}^{n} (Y_r - B_0^{MC} - B_1^{MC} x_r)^2 &= 0 \\ \frac{\partial}{\partial B_1^{MC}} \sum_{r=1}^{n} (Y_r - B_0^{MC} - B_1^{MC} x_r)^2 &= 0 \end{aligned}$$

★ dans la suite pour simplifier l'écriture nous utilisons $B_0 = B_0^{MC}$ et $B_1 = B_1^{MC}$

Dérivation et annulation de la dérivée par rapport à B_0

$$\frac{\partial}{\partial B_0}\sum_{r=1}^{n}(Y_r - B_0 - B_1 x_r)^2 = 0$$
$$\Updownarrow$$
$$-2\sum_{r=1}^{n}(Y_r - B_0 - B_1 x_r) = 0$$
$$\Updownarrow$$
$$\sum_{r=1}^{n} Y_r - B_0 \sum_{r=1}^{n} 1 - B_1 \sum_{r=1}^{n} x_r = 0$$
$$\Updownarrow$$
$$n\overline{Y} - nB_0 - nB_1\overline{x} = 0$$
$$\Updownarrow$$
$$\overline{Y} - B_0 - B_1\overline{x} = 0$$

ce qui s'écrit

La droite des moindres carrés passe par le point moyen

$$\overline{Y} = B_0 + B_1\overline{x}$$

Dérivation et annulation de la dérivée par rapport à B_1

$$\frac{\partial}{\partial B_1}\sum_{r=1}^{n}(Y_r - B_0 - B_1 x_r)^2 = 0$$
$$\Updownarrow$$
$$-2\sum_{r=1}^{n}(Y_r - B_0 - B_1 x_r)x_r = 0$$
$$\Updownarrow$$
$$\sum_{r=1}^{n} x_r Y_r - B_0\sum_{r=1}^{n} x_r - B_1\sum_{r=1}^{n} x_r^2 = 0$$
$$\Updownarrow$$
$$\sum_{r=1}^{n} x_r Y_r - nB_0\bar{x} - B_1\sum_{r=1}^{n} x_r^2 = 0$$
$$\Updownarrow$$
$$\sum_{r=1}^{n} x_r Y_r - n\bar{x}\overline{Y} + nB_1\bar{x}^2 - B_1\sum_{r=1}^{n} x_r^2 = 0$$
$$\Updownarrow$$
$$\sum_{r=1}^{n}(x_r - \bar{x})(Y_r - \overline{Y}) - B_1\sum_{r=1}^{n}(x_r - \bar{x})^2 = 0$$

Expression des estimateurs

- l'estimateur B_1 est

Estimateur de la pente de la régression

$$B_1 = \frac{\sum_{r=1}^{n}(x_r - \overline{x})(Y_r - \overline{Y})}{\sum_{r=1}^{n}(x_r - \overline{x})^2}$$

- l'estimateur B_0 est

Estimateur de l'ordonnée à l'origine

$$B_0 = \overline{Y} - B_1\overline{x}$$

Lien avec la corrélation linéaire

Rappels

- on utilise les quantités
  - $s_x^2 = \sum_1^n (x_r - \bar{x})^2$
  - $S_{xY} = \sum_1^n (x_r - \bar{x})(Y_r - \overline{Y})$
  - $S_Y^2 = \sum_1^n (Y_r - \overline{Y})^2$
- le coeffcient de corrélation linéaire est

$$R_{xY} = \frac{\sum_1^n (x_r - \bar{x})(Y_r - \overline{Y})}{\sqrt{\sum_1^n (x_r - \bar{x})^2 \times \sum_1^n (Y_r - \overline{Y})^2}} = \frac{S_{xY}}{s_x \times S_Y}$$

Lien avec la corrélation linéaire

Écriture

- on peut écrire l'estimateur de la pente

$$B_1 = \frac{S_{xY}}{S_x^2}$$

★ c'est une fonction dissymétrique en x et Y

→ on en déduit le

Lien entre estimateur de la pente et corrélation linéaire

$$B_1 = R_{xY}\frac{S_Y}{s_x}$$

- on remarque que si $R_{xY} = 0$ alors $B_1 = 0$

Nature des estimateurs

- B_0 et B_1 sont des fonctions de Y_r et x_r
- en tant que fonction de Y ce sont des variables aléatoires

→ on peut calculer

  - leur espérance
  - leur variance

Espérances des estimateurs : B_1

$$
\begin{aligned}
B_1 &= \frac{\sum_{r=1}^{n}(x_r-\overline{x})(Y_r-\overline{Y})}{\sum_{r=1}^{n}(x_r-\overline{x})^2} \\
\mathbb{E}(B_1) &= \frac{\sum_{r=1}^{n}(x_r-\overline{x})(\mathbb{E}(Y_r)-\mathbb{E}(\overline{Y}))}{\sum_{r=1}^{n}(x_r-\overline{x})^2} \\
&= \frac{\sum_{r=1}^{n}(x_r-\overline{x})(\beta_0+\beta_1 x_r-\beta_0-\beta_1\overline{x_r})}{\sum_{r=1}^{n}(x_r-\overline{x})^2} \\
&= \frac{\sum_{r=1}^{n}(x_r-\overline{x})\beta_1(x_r-\overline{x_r})}{\sum_{r=1}^{n}(x_r-\overline{x})^2} \\
&= \beta_1\frac{\sum_{r=1}^{n}(x_r-\overline{x})^2}{\sum_{r=1}^{n}(x_r-\overline{x})^2} \\
&= \beta_1
\end{aligned}
$$

Espérances des estimateurs : B_0

$$\begin{aligned} B_0 &= \overline{Y} - B_1\overline{x} \\ \mathbb{E}(B_0) &= \mathbb{E}(\overline{Y}) - \mathbb{E}(B_1)\overline{x} \\ &= \beta_0 + \beta_1\overline{x} - \beta_1\overline{x} \\ &= \beta_0 \end{aligned}$$

Variances des estimateurs

- $Var(Y_r) = Var(E_r) = \sigma_E^2$
- $Var(B_0) = \sigma_E^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_x^2}\right]$
- $Var(B_1) = \frac{\sigma_E^2}{S_x^2}$
- Ce sont des estimateurs de **variance minimum**
- Variances faibles si S_X^2 fort
- c-à-d x_r éloignés de $\bar{x}$.

Propriétés des estimateurs
Résumé

- les calculs ne sont pas fait ici mais

→ $\mathbb{E}(B_0) = \beta_0$ et $\mathbb{E}(B_1) = \beta_1$

→ ce sont des estimateurs **sans biais**

Autour des estimations

- Les paramètres étant estimés, tout se déduit,

→ On prédit les variables aléatoires

→ Le prédicteur de Y_r

$$B_0 + B_1 x_r$$

→ La prédiction de Y_r

$$b_0 + b_1 x_r$$

→ L'erreur estimée ou **résidu**

$$y_r - \hat{Y}_r$$

Estimation de σ_E^2

- on estime σ_E^2 à partir de la somme des carrés des écarts des aléatoires observables Y_r à leurs prédictions Y_r

Somme des Carrés des Écarts Résiduels

$$SCER = \sum_{r=1}^{n} (Y_r - B_1 x_r - B_0)^2$$

→ en la divisant par $n-2$ on obtient un estimateur non biaisé

Estimateur de la variance résiduelle

$$S_E^2 = \frac{\sum_{r=1}^{n} (Y_r - B_1 x_r - B_0)^2}{n-2}$$

★ il s'agit aussi de l'estimateur du maximum de vraisemblance gaussien corrigé pour le non biais

Application sur les données moules

Listage


```
> summary(lm(pds.chair~pds.animal,moules))

Residuals :
     Min       1Q   Median       3Q      Max
-1.14823 -0.16649  0.00057  0.22968  0.97175

Coefficients :
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.08101    0.16325  -0.496    0.622
pds.animal   0.35273    0.02451  14.391   <2e-16 ***

Residual standard error : 0.4071 on 57 degrees of freedom
Multiple R-Squared : 0.7842,   Adjusted R-squared : 0.7804
F-statistic : 207.1 on 1 and 57 DF,  p-value : < 2.2e-16
```

Application sur les données moules

Interprétation du listage

- estimations des paramètres de l'espérance

$$b_0 = -0.08101, b_1 = 0.35273$$

- estimation de l'écart-type résiduel

$$s = 0.4071$$

→ incertitude sur les estimations

estimation de $var(B_0) = 0.16325^2$

estimation de $var(B_1) = 0.02451^2$

- prédicteur linéaire

$$\hat{Y} = B_0 + B_1 x$$

→ prédiction

$$y = -0.08101 + 0.3573x$$

Application sur les données moules

Listing du modèle sans β_0


```
 summary(lm(pds.chair~pds.animal-1,moules))

Residuals :
     Min       1Q   Median       3Q      Max
-1.12293 -0.17753 -0.02853  0.22111  0.96345

Coefficients :
           Estimate Std. Error t value Pr(>|t|)
pds.animal 0.341227   0.007905   43.16   <2e-16 ***

Residual standard error : 0.4044 on 58 degrees of freedom
Multiple R-Squared : 0.9698,   Adjusted R-squared : 0.9693
F-statistic :  1863 on 1 and 58 DF,  p-value : < 2.2e-16
```

Relation entre estimateurs des paramètres d'espérance

- les estimateurs B_0 et B_1 sont fortement corrélés

→ intuitivement une observation qui «agit» sur la pente «agit»aussi sur l'ordonnée à l'origine

→ tendance à «pivoter» autour du point moyen

→ à cause de la contrainte : la droite passe par $\overline{Y}, \overline{x}$

- on postule que les variables aléatoires résiduelles sont gaussiennes

$$E_r \rightsquigarrow \mathcal{N}(0, \sigma_E^2)$$

→ ceci entraîne que les Y_r sont aussi gaussiens

$$Y_r \rightsquigarrow \mathcal{N}(\mu_{Y/x}, \sigma_{Y/x}^2)$$

→ avec

$$\begin{aligned} \sigma_{Y/x}^2 &= \sigma_E^2 \\ \mu_{Y/x} &= \beta_0 + \beta_1 x \end{aligned}$$

- alternative à tester
- statistique de test (règle de décision)
- loi de la statistique
- régions d'acceptation et de rejet
- normalisation de la région (en fonction du risque α)
- décision

Test sur la pente
Hypothèses

- en langage courant

  H_0 : le modèle n'a pas d'intérêt

  H_1 : le modèle a un intérêt

- en langage statistique

  H_0 : la variable x n'explique pas la variable Y

  H_1 : la variable x explique la variable Y

- formellement

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Test sur la pente

Statistique de test

- l'estimateur de β_1 est B_1,
- sa variance est $Var(B_1) = \frac{\sigma^2_{Y/x}}{S^2_x}$

→ on construit une statistique centrée

$$B_1 - 0$$

→ que l'on normalise par son écart type

$$\frac{B_1}{\sqrt{\mathbb{V}(B_1)}} = \frac{B_1 S^2_x}{\sqrt{\sigma^2_{Y/x}}}$$

→ on estime $\sigma^2_{Y/x}$ par

$$\frac{SCER}{n-2}$$

- la statistique

$$T_{\beta_1} = B_1 \times \frac{S_x^2 \times \sqrt{n-2}}{\sqrt{SCER}}$$

→ correspond à une loi de Gauss divisée par un χ^2

→ il s'agit d'une loi t de Student

★ à $n-2$ degrés de liberté

Test sur la pente
Principe

- si t_{β_1} **est voisin de 0** on décide H_0

→ entre 2 bornes (quantiles) qui dépendent du risque α

→ plus le risque est faible plus les bornes s'éloignent de 0,

→ plus on a tendance à décider H_0

- les logiciels fournissent la *probabilité critique*

→ si cette probabilité est inférieure à α on décide H_1.

→ le test est dit *significatif*

- Test **d'absence de relation** entre Y et x
- **Risque** souvent 5%

→ Grossièrement si n est assez grand, les bornes $-2, +2$

- Absence de relation :

→ la probabilité critique> 0.05.