MOOC Statistique pour ingénieur
Thème 4 : Régression linéaire
Vidéo 3 : Tests, prévision

Anca Badea

Institut Mines-Télécom
Mines Saint-Étienne

Sommaire

Sommaire

Tests sur β_i

et si le paramètre β_i était nul ?

- $H_0 : \beta_i = 0$
  $H_1 : \beta_i \neq 0$
- statistique de test : $T_{\hat{\beta}_i} = \dfrac{\hat{\beta}_i}{s_{\hat{\beta}_i}} \sim \mathcal{T}(n-2)$
- fixer le risque α
- décision : rejet ou pas de H_0
  en fonction de l'appartenance de la valeur de la statistique de test à la région critique
  $]-\infty, -t_{\alpha/2}[\cup]t_{\alpha/2}, \infty[$ ou pas

Tests sur β_i

et si le paramètre β_i était nul ?

- $H_0 : \beta_i = 0$
  $H_1 : \beta_i \neq 0$
- fixer le risque α
- statistique de test : $T_{\hat{\beta}_i} = \dfrac{\hat{\beta}_i}{s_{\hat{\beta}_i}} \sim \mathcal{T}(n-2)$
- décision : rejet ou pas de H_0 en fonction de la $p-$valeur
  si $p_{val} \leq \alpha \rightarrow$ rejet
  avec $p_{val} = \mathbb{P}_{H_0}(|T_{\hat{\beta}_i}| \geq |t_{\hat{\beta}_i}|)$

Test de Fisher

significativité de la régression en utilisant l'analyse de la variance

- $H_0 : \beta_1 = 0$
  $H_1 : \beta_1 \neq 0$
- statistique de test : $F = \dfrac{\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2/1}{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2/(n-2)} \sim \mathcal{F}_{1,n-2}$
- fixer le risque α
- décision : rejet ou pas de H_0
  en fonction de l'appartenance de la valeur de la statistique de test à la région critique $]f_\alpha, \infty[$ ou pas
  (avec f_α la valeur telle que $\mathbb{P}(F \leq f_\alpha) = 1 - \alpha$)

  ou en fonction de la $p-$valeur : si $p_{val} \leq \alpha \longrightarrow$ rejet ($p_{val} = \mathbb{P}_{H_0}(F \geq f)$)

Exemple alligators

$$t_{\widehat{\beta}_1} = \frac{\widehat{\beta}_1}{s_{\widehat{\beta}_1}} = 25,8, \quad t_{\widehat{\beta}_0} = \frac{\widehat{\beta}_0}{s_{\widehat{\beta}_0}} = -16,93$$

décisions tests : rejet de H_0 pour les tests sur β_0 et β_1

Variable	Coefficient	Ecart-type	t	p_{val}
Intercept	$-8,48$	$0,5$	$-16,93$	3 e-10
lnLength	$3,43$	$0,13$	$25,8$	1e-12

$\widehat{\sigma}^*$	ddl	R^2	R^2_{adj}
$0,12$	13	$0,9808$	$0,9794$

F-statistique : 666	sur 1 et 13 ddl	p-val : 1e-12

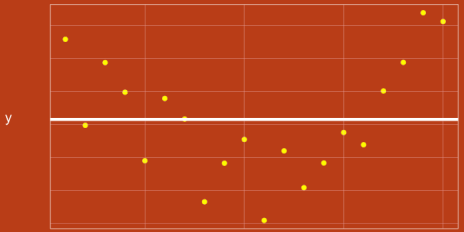
Importance du test

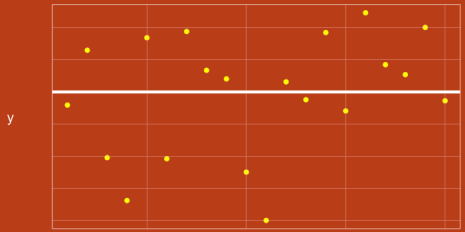
- $H_0 : \beta_1 = 0$
  $H_1 : \beta_1 \neq 0$

si on ne peut pas rejeter H_0

il n'y a pas de relation linéaire entre *x* et *y*

x n'explique pas la variabilité de *y*





la meilleure prédiction de *y* est $\bar{y}$

Commentaires

Si on rejette H_0 :

x a une certaine importance pour expliquer la variabilité de *y*

- soit le modèle de régression simple est adéquat,
- soit, bien que l'effet linéaire est présent, des meilleurs résultats pourraient être obtenus en rajoutant des termes d'ordre plus élevés

Sommaire

Prévision de la variable expliquée

Estimer $\mathbb{E}(Y|x_0) = \beta_0 + \beta_1 x_0$ pour une valeur particulière x_0
(x_0 comprise dans la plage de variation de x)

- estimateur : $\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$
- estimation ponctuelle : $\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$
- estimation par intervalle de confiance
  - loi de $\widehat{Y}_0$: loi normale (car CL des Y_i)
  - espérance :
    $\mathbb{E}(\widehat{\beta}_0 + \widehat{\beta}_1 x_0) = \beta_0 + \beta_1 x_0$
  - variance :
    $\mathbb{V}(\widehat{\beta}_0 + \widehat{\beta}_1 x_0) = \mathbb{V}(\overline{Y} + \widehat{\beta}_1(x_0 - \overline{x})) = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{n\, s_x^2}\right]$

    car $\mathbb{C}ov(\overline{Y}, \widehat{\beta}_1) = 0$

Prévision de la variable expliquée

$$\frac{\widehat{Y}_0 - (\beta_0 + \beta_1 x_0)}{\widehat{\sigma}^* \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{n\, s_x^2}}} \sim \mathcal{T}(n-2)$$

loi de Student à $n-2$ degrés de libertés

$$Ic_{1-\alpha}(\beta_0 + \beta_1 x_0) =$$

$$\left[\widehat{\beta}_0 + \widehat{\beta}_1 x_0 - t_{\alpha/2} \widehat{\sigma}^* \sqrt{\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{n\, s_x^2} \right]},\ \widehat{\beta}_0 + \widehat{\beta}_1 x_0 + t_{\alpha/2} \widehat{\sigma}^* \sqrt{\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{n\, s_x^2} \right]} \right]$$

Prévision de la variable expliquée

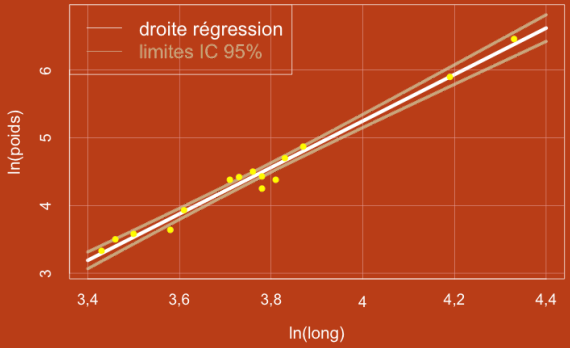
- la largeur dépend de x_0 :
  - minimum pour $x_0 = \overline{x}$
  - augmente pour des valeurs croissantes de $|x_0 - \overline{x}|$.
- meilleures prévisions seront pour des valeurs de x proches de $\overline{x}$
- la précision se détériore vers des extrémités de la plage de x

Exemple alligators

$n = 15,\ \bar{x} \approx 3,76,\ s_x^2 \approx 0,06,$
$\widehat{\beta}_1 \approx 3,43,\ \widehat{\beta}_0 \approx -8,48$
$\widehat{\sigma}^{*2} \approx 0,02$
$t_{0,025} = 2,16$

$x_0 = 4$

$\widehat{y}_0 \approx -8,48 + 3,43 \times 4 = 5,25$
$Ic_{0,95}(\beta_0 + \beta_1 x_0) \approx [5,15\ ;\ 5,35]$



Prévision de la variable expliquée

Estimer $Y_0 = \beta_0 + \beta_1 x_0 + \varepsilon$ pour une valeur particulière x_0
(x_0 comprise dans la plage de variation de x)

- estimateur : $\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$
- calculons la variance de $Y_0 - \widehat{Y}_0$

$$\mathbb{V}(Y_0 - \widehat{Y}_0) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{n\, s_x^2} \right]$$

car Y_0 et $\widehat{Y}_0$ sont indépendantes

Intervalle de prévision pour des nouvelles observations

$$Ic_{1-\alpha}(y_0) = \left[\widehat{y}_0 - t_{\alpha/2}\widehat{\sigma}^* \sqrt{\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{n\, s_x^2}\right]},\ \widehat{y}_0 + t_{\alpha/2}\widehat{\sigma}^* \sqrt{\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{n\, s_x^2}\right]}\right]$$

l'intervalle de prédiction est toujours plus large que l'intervalle de confiance

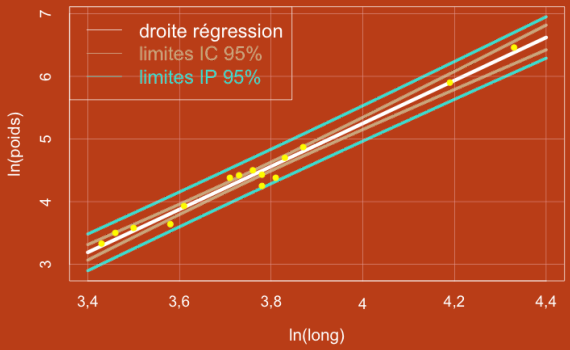
Exemple alligators

$n = 15,\ \bar{x} \approx 3,76,\ s_x^2 \approx 0,06,$
$\widehat{\beta}_1 \approx 3,43,\ \widehat{\beta}_0 \approx -8,48$
$\widehat{\sigma}^{*2} \approx 0,02$
$t_{0,025} = 2,16$

$x_0 = 4$

$\widehat{y}_0 \approx -8,48 + 3,43 \times 4 = 5,25$
$Ic_{0,95}(\beta_0 + \beta_1 x_0) \approx [5,15\ ;\ 5,35]$

$Ic_{0,95}(y_0) \approx [4,97\ ;\ 5,53]$




INSTITUT
Mines-Télécom