Modèles Linéaires Appliqués

Arthur Charpentier

Automne 2020

OLS #3 (régression sur une variable continue - 2)

Droite de régression (x continue)

Si ε suit une loi normale, et en supposant l'indépendance des résidus, on obtient la normalité de $\widehat{\beta}_0$ et $\widehat{\beta}_1$.
Aussi, une statistique naturelle de test de $H_0 : \beta_1 = 0$ contre $H_1 : \beta_0 \neq 0$ est basé sur

$$T = \frac{\widehat{\beta}_1 - 0}{\sqrt{\widehat{\text{Var}}[\widehat{\beta}_1]}}$$

qui suit, sous H_0, une loi de Student $\mathcal{S}td(n-2)$.

Droite de régression (x continue)

Comme c'est un test bi-latéral, on utilise alors une des deux méthodes

- région critique: si $t^{-1}_{n-2}(1-\alpha)$ désigne le quantile de niveau $1-\alpha$ de la loi $\mathcal{S}td(n-2)$, on rejette H_0 si $|T| > t^{-1}_{1,n-2}(1-\alpha/2)$,
- p-value: on rejette H_0 si $p = \mathbb{P}[|Y| > |T||Y \sim \mathcal{S}td(n-2)] < \alpha$.

Considerons le test $H_0 : \beta_1 = 0$ contre $H_1 : \beta_0 \neq 0$. La statistique de Fisher (analyse de la variance) vise à comparer les résidus de deux modèles : (0) $y_i = \beta_0 + \varepsilon_i$ et (0) $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, i.e.

$$F = \frac{(TSS - RSS)/1}{RSS/(n-2)} = \frac{ESS}{RSS/(n-2)}$$

qui suit, sous H_0 une loi de Fisher $\mathcal{F}_{1,n-2}$.

Droite de régression (x continue)

On utilise alors une des deux méthodes

- région critique: si $F^{-1}_{1,n-2}(1-\alpha)$ désigne le quantile de niveau $1-\alpha$ de la loi $\mathcal{F}_{\infty,\backslash-\in}$, on rejette H_0 si $F > F^{-1}_{1,n-2}(1-\alpha)$,
- p-value: on rejette H_0 si $p = \mathbb{P}[Y > F | Y \sim \mathcal{F}_{\infty,\backslash-\in}] < \alpha$.

On peut noter que

$$F = (n-2)\frac{R^2}{1-R^2}$$

Dans le cas de la significativité de la pente, i.e. $H_0 : \beta_1 = 0$, notons que

$$T^2 = \frac{\widehat{\beta}_1^2}{\widehat{\sigma}^2/s_x^2} = W = \frac{\widehat{\beta}_1^2 s_x^2}{\widehat{\sigma}^2} = \frac{ESS}{RSS/(n-2)} = F.$$

Droite de régression (x continue)

La prévision associée à x_i est

$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

et si on considere une nouvelle observation x, on aurait

$$\widehat{y}_x = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

et on aura comme observation

$$y_x = \beta_0 + \beta_1 x + \varepsilon$$

où ε est un bruit imprévisible.
Notons que

$$y_x - \widehat{y}_x = \big(\beta_0 + \beta_1 x + \varepsilon\big) - \big(\widehat{\beta}_0 + \widehat{\beta}_1 x\big)$$

soit

$$y_x - \widehat{y}_x = \varepsilon + \big(\beta_0 - \widehat{\beta}_0\big) + \big(\beta_1 - \widehat{\beta}_1\big) x$$

de telle sorte que

$$\mathbb{E}[y_x - \widehat{y}_x] = 0$$

Droite de régression (x continue)

Aussi, $\widehat{y}_x$ est un estimateur dans biais de y_x.
Si on continue,

$$\text{Var}[y_x - \widehat{y}_x] = \text{Var}[\varepsilon] + \text{Var}\big(\widehat{\beta}_0 + \widehat{\beta}_1 x\big)$$

qui se réécrit

$$\text{Var}[y_x - \widehat{y}_x] = \sigma^2 + \left(\frac{\sigma^2}{n} + \frac{\sigma^2(x-\overline{x})^2}{s_x^2}\right)$$

Frees (2014) appelle "standard error of prediction" la racine carrée de l'estimateur de cette grandeur

$$\sqrt{\widehat{\text{Var}}[y_x - \widehat{y}_x]} = \widehat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{s_x^2}}$$

Droite de régression (x continue)

En notant que si les residus sont supposes Gaussien, on peut alors construire un intervalle de confiance de prédiction pour y_x :

$$\underbrace{\widehat{\beta}_0 + \widehat{\beta}_1 x}_{\widehat{y}_x} \pm t_{n-2,1-\alpha/2} \cdot \widehat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{s_x^2}}$$

On retrouve dans l'expression précédante deux sources d'erreur - l'erreur de modèle, venant du fait que la réalisation y est $\beta_0 + \beta_1 x$ auquel s'ajoute un bruit ε - l'erreur d'estimation, venant du fait que $\beta_0 + \beta_1 x$ est incertain

Droite de régression (x continue)

Pour l'erreur d'estimation, notons que

$$\text{Var}[\widehat{y}_x] = \left(\frac{\sigma^2}{n} + \frac{\sigma^2(x-\overline{x})^2}{s_x^2}\right)$$

d'où un intervalle de confiance pour $\widehat{y}_x$ de la forme

$$\underbrace{\widehat{\beta}_0 + \widehat{\beta}_1 x}_{\widehat{y}_x} \pm t_{n-2,1-\alpha/2} \cdot \widehat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{s_x^2}}$$