

Statistiques, séance n°5

La régression linéaire : une courte introduction au modèle linéaire

EPITA

Mai 2020

Problématique

- Nous entrons désormais dans le domaine des **modèles prédictifs**,
- A partir de données, en général portant sur des variables numériques
- Comprendre les relations entre les variables
- Typiquement, peut-on « expliquer » l'une des variables en fonction des autres ?
- Exemples multiples (météo, économie, finance, marketing, ...)

Régression simple

- Nous disposons de deux variables aléatoires X et Y .
- Nous disposons d'un échantillon de réalisations de ces variables (x_i, y_i)
- Ces deux variables ne sont pas indépendantes à priori
- Que peut-on dire de Y connaissant X ?
- On aimerait bien « expliquer » le comportement de Y en fonction de X , trouver une fonction f telle que $f(X)$ « ressemble » à Y
- Au mieux $\mathbb{E}(Y|X)$

Modèle linéaire

- Le modèle :

- $Y = \alpha + \beta X + \varepsilon$

$$\mathbb{E}(\varepsilon) = 0 \quad \text{Var}(\varepsilon) = \sigma^2$$

- $\mathbb{E}(Y) = \alpha + \beta \mathbb{E}(X)$

- Attention, c'est un modèle !!!

- Quelles sont les bonnes valeurs de α, β ?

- ... tout en gardant un regard critique sur le modèle !!!

- Notamment le fait qu'il n'y a pas de raison à priori, pour que σ^2 soit petit

Modèle (2)

- $Y - \mathbb{E}(Y) = \beta(X - \mathbb{E}(X)) + \varepsilon$
- $\mathbb{E}\left((Y - \mathbb{E}(Y))(X - \mathbb{E}(X))\right) = \beta\sigma^2(X) + \mathbb{E}(\varepsilon)\mathbb{E}(X - \mathbb{E}(X))$
 $= \beta\sigma^2(X)$
- $\text{cov}(X, Y) = \beta V(X)$
- $\beta = \frac{\text{cov}(X, Y)}{V(X)} = \rho \frac{\sigma_y}{\sigma_x} \qquad \mathbb{E}(Y) - \beta\mathbb{E}(X) = \alpha$
- Les coefficients se « calculent bien » dans le modèle linéaire
- $\hat{\beta} = r \frac{s_y}{s_x} \qquad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$

- Echantillon de n observations indépendantes
- ε tel que : $\mathbb{E}(\varepsilon) = 0$ $Var(\varepsilon) = \sigma^2$
- α, β, σ^2 estimés par la méthode des moindres carrés
- c.a.d. minimiser $G(\alpha, \beta) = \sum (y_i - \beta x_i - \alpha)^2$
- Annulation de dérivées partielles $\frac{\partial G}{\partial \alpha}$ $\frac{\partial G}{\partial \beta}$

Régression linéaire avec R

- Il faudrait un cours spécifique, le modèle linéaire permet aussi :
- $Y = \alpha_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$
- Les coefficients se calculent bien à l'aide de calcul matriciel
- ... mais il faut garder le regard critique sur un modèle
- En R commande `lm` (cf. démos)