Big Data

Régression linéaire

benoit.lardeux@isen-ouest.yncrea.fr

Inspiré des notes de cours de M. Saumard, ISEN Brest

Exemple de cas d’application

- Vente d’un produit (en milliers d’unités)
- En fonction du budget publicitaire (en milliers d’euros) pour la TV, la radio, les journaux (papiers)
- Objectif: Optimiser le budget publicitaire pour en vendre le plus

- Questions:
    - Existe-t ’il une relation entre les ventes et le budget publicitaire?
    - Quel média contribue aux ventes?
    - Peut-on prédire les futures ventes?

Modélisation du problème marketing

- Première approche:

$$Y \approx \beta_0 + \beta_1 X$$

  - Y représente les ventes
  - X représente le budget pub pour la tv
  - β_0 et β_1 sont des paramètres à déterminer du modèle

ISEN ALL IS DIGITAL! OUEST yncréa

Les étapes d'une régression linéaire

- Formulation et hypothèses du modèle
- Estimation des paramètres
- Qualité d'ajustement
- Tests d'hypothèses

ISEN ALL IS DIGITAL! OUEST | yncréa

Hypothèses du problème

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ pour } i = 1, \dots, n$$

- X fixe
- $E[\varepsilon_i] = 0$ erreur centrée et $var(\varepsilon_i) = \sigma^2$ (homoscédasticité)
- β_0 et β_1 sont constants (pas d'évolutions, pas de rupture de modèle)
- Pour l'inférence, on supposera de plus $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$

Remarque: On parle d'homoscédasticité lorsque la variance des erreurs stochastiques de la regression est la même pour chaque observation i (de 1 à n observations)

ISEN ALL IS DIGITAL! OUEST | yncréa

Estimation des paramètres

- Moindres carrés

$$min_{\beta_0,\beta_1} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

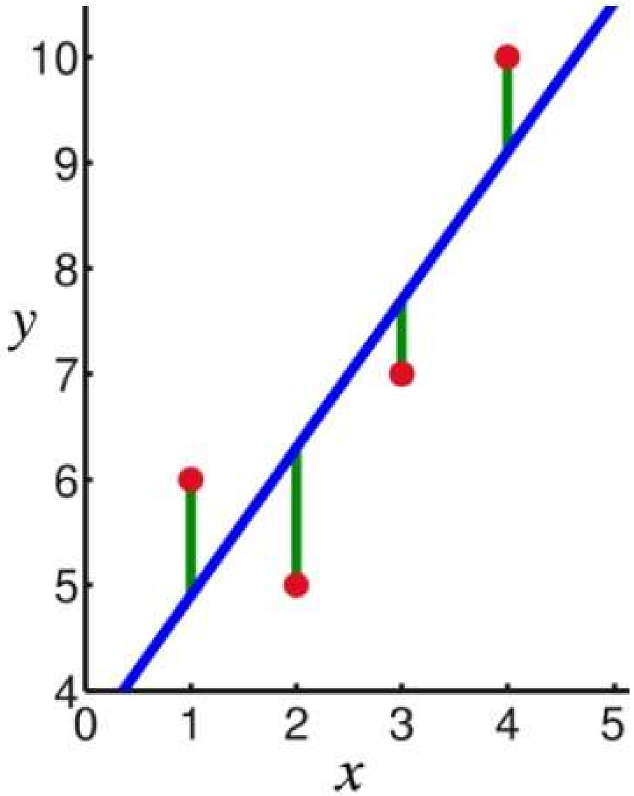
- Résolution
  - Posons $\bar{x} = \frac{1}{n}\sum x_i$ et $\bar{y} = \frac{1}{n}\sum y_i$,

$$s_x^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2 \;, \quad s_y^2 = \frac{1}{n-1}\sum(y_i - \bar{y})^2 \quad \text{et} \quad s_{xy} = \frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y})$$

  - Alors

$$\widehat{\beta_1} = \frac{s_{xy}}{s_x^2} \qquad \text{et} \qquad \widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\,\bar{x}$$

Exemple



ISEN ALL IS DIGITAL! OUEST | yncréa

Prédiction, résidus et variance estimée

- Prédiction

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$$

- Résidus estimés

$$e_i = y_i - \widehat{y_i}$$

- Variance du modèle σ^2 estimée par

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2$$

Précision des estimateurs

- Question: ces estimations sont-elles précises?

- Calcul de l'erreur standard de $\widehat{\beta_0}$ et $\widehat{\beta_1}$:

$$SE(\widehat{\beta_0})^2 = \sigma^2\left[\frac{1}{n}\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right],$$

$$SE(\widehat{\beta_1})^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2},$$

où $\sigma^2 = Var(\varepsilon)$

Précision des estimateurs

- Intervalles de confiance à 95% pour $\widehat{\beta_0}$ et $\widehat{\beta_1}$:

$$\widehat{\beta_0} \pm 1{,}96\, SE(\widehat{\beta_0})$$

$$\widehat{\beta_1} \pm 1{,}96\, SE(\widehat{\beta_1})$$

Etude de cas

- Imaginons que, pour les données publicitaires, on ait les intervalles de confiance suivants:

$$IC(95\%)(\beta_0) = [6{,}130 ; 7{,}935]$$
$$IC(95\%)(\beta_1) = [0{,}042 ; 0{,}053]$$

Alors, on peut dire qu'en absence de publicité, les ventes en moyenne tomberont entre 6,130 et 7,935 unités. De plus, pour chaque 1000 € investis en pub, il y aura en moyenne une augmentation des ventes de 42 à 53 unités

Tests d’hypothèses sur les coefficients

- Testons
  - H_0 : Il n’y a pas de relation entre X et Y

  versus
  - H_1 : Il y a une relation entre X et Y

  - Cela correspond mathématiquement à tester
    - $H_0 : \beta_1 = 0$

    versus
    - $H_1 : \beta_1 \neq 0$

ISEN ALL IS DIGITAL! OUEST | yncréa

Tests d'hypothèses sur les coefficients

- On utilise la statistique t:

$$t = \frac{\widehat{\beta_1}}{SE(\widehat{\beta_1})}$$

- Sous H_0, t suit une loi de Student à $n - 2$ degrés de liberté

- On calcule $a = P_{H_0}(|T| > |t|)$ et on compare à 5%
  - Si a est supérieur à 0,05, on ne refuse pas H_0
  - Si a est plus petit que 0,05, on rejette l'hypothèse H_0

Etude de cas

	Coefficient	Erreur standard	Statistique t	P valeur
Intercept	7,0325	0,4578	15,36	<0,0001
TV	0,0475	0,0027	17,59	<0,0001

Précision du modèle

- 2 types d'indicateurs concernant la précision du modèle
  - RSE (erreur standard des résidus)
  - R^2 (coefficient de détermination)

Précision du modèle

- RSE

$$RSE = \sqrt{\frac{1}{n-2} RSS} \ ,$$

où $RSS = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$

où RSS est la somme des résidus au carré

ISEN ALL IS DIGITAL! OUEST yncréa

Précision du modèle

- R^2

$$R^2 = 1 - \frac{RSS}{TSS}$$

où TSS est la somme totale des carrés ($\sum_1^n (y_i - \bar{y})^2$)

- $0 < R^2 < 1$

- Plus R^2 est proche de 1, mieux le modèle explique les données

Régression linéaire multiple

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \,, \quad \text{pour } i = 1, \dots, n$$

- Où
  - Les x_{ij} sont des nombres connus, non aléatoires
  - Les paramètres β_j du modèle sont inconnus, mais non aléatoires
  - Les ε_i sont des variables aléatoires inconnues

Remarque: Pour avoir une constante, on peut prendre $x_{i1} = 1$

Forme matricielle

- Le modèle s'écrit sous forme matricielle:

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

- Où
  - Y est un vecteur aléatoire de dimension *n*
  - X est une matrice de taille *n x p* connue, appelée matrice du plan d'expérience
  - β est le vecteur de dimension *p* des paramètres inconnus du modèle
  - ε est le vecteur de dimension *n* des erreurs

- Hypothèses du modèle
  - $Rg(X) = p$
  - Les erreurs sont centrées, de même variance σ^2 et non corrélées entre elles

Moindre carrés ordinaires (MCO)

- L'estimateur des MCO du vecteur inconnu β est

$$\hat{\beta} = ({}^{t}XX)^{-1}\ {}^{t}XY$$

- Il vérifie $\hat{\beta} = Argmin_{\beta \in \mathbb{R}} ||Y - \beta X||^2$

Des propriétés de l’estimateur

- Estimateur sans biais:

$$\mathbb{E}[\hat{\beta}] = \beta$$

- Parmi les estimateurs sans biais, il est de variance minimum et:

$$\mathrm{Var}(\hat{\beta}) = \sigma^2 \left({}^tXX \right)^{-1}$$

ISEN ALL IS DIGITAL! OUEST yncréa

Régression linéaire multiple

- Estimateur de σ^2:

$$\hat{\sigma}^2 = \frac{||Y - \hat{Y}||^2}{n-p}$$

- Cette statistique est un estimateur sans biais de σ^2

ISEN ALL IS DIGITAL! OUEST yncréa

Etude de cas

- Etude de la concentration d'ozone (O) dans l'atmosphère en fonction de la température (T), du vent (V) (phénomène d'advectance) et de la nébulosité (N)
  - 10 données journalières de température, vent, nébulosité et ozone

T	V	N	O
23,8	9,25	5	115,4
16,3	-6,15	7	76,8
27,2	-4,92	6	113,8
7,1	11,57	5	81,6
25,1	-6,23	2	115,4
27,5	2,76	7	125
19,4	10,15	4	83,6
19,8	13,5	6	75,2
32,2	21,27	1	136,8
20,7	13,79	4	102,8

ISEN ALL IS DIGITAL! OUEST | yncréa

Résultats

```
> a <- lm(O3 ~ T12+Vx+Ne12,data=DONNEE)
> summary(a)
Call:
lm(formula = O3 ~ T12 + Vx + Ne12, data = DONNEE)

 Residuals:
                 Min        1Q    Median       3Q      Max
            -29.0441   -8.4833    0.7857   7.7011  28.2919

 Coefficients :
              Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)   84.5483     13.6065    6.214  1.38e-07   ***
 T12            1.3150      0.4974    2.644   0.01118   *
 Vx             0.4864      0.1675    2.903   0.00565   **
 Ne12          -4.8935      1.0270   -4.765  1.93e-05   ***

--
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 13.91 on 46 degrees of freedom
Multiple R-Squared: 0.6819, Adjusted R-squared: 0.6611
F-statistic: 32.87 on 3 and 46 DF, p-value: 1.663e-11
```

ISEN ALL IS DIGITAL! OUEST | yncréa

Question?