Analyse de données pour le génie industriel

Régression linéaire simple

Iragaël Joly Pierre Lemaire

Grenoble INP – Génie Industriel, 2A

2015–2016

Bibliographie

- P.-A. Cornillon et al. *Statistiques avec R*, PU Rennes, 2012.
- S. Tufféry. *Data Mining et statistique décisionnelle*, Technip, 2012.
- G. Saporta. *Probabilités et analyse de données*, Technip, 2011.

1. Régression linéaire simple

Régression (1)

- les données sont une matrice $n \times (p+1)$
  - un échantillon de n observations
  - chaque observation est constituée de $p+1$ variables
  - une réponse (ou variable dépendante ou à expliquer) : y
  - des variables explicatives (ou indépendantes) : $x_1, x_2, \ldots, x_p$

$$\left(\begin{array}{c|cccc} y_1 & x_{11} & x_{21} & \cdots & x_{p1} \\ y_2 & x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & & & & \vdots \\ y_n & x_{1n} & x_{2n} & \cdots & x_{pn} \end{array}\right)$$

Régression

une régression de y sur $\mathbf{x}$ est un modèle de la forme

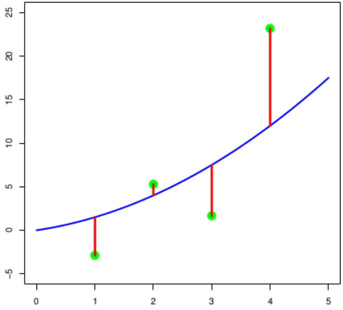
$$y = f(\mathbf{x}) + \varepsilon$$

Régression (2)

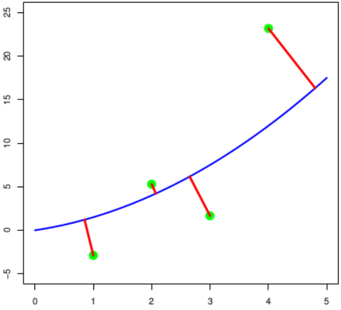
- f caractérise l'effet de $\mathbf{x}$ sur y ; la forme de f est une hypothèse
- ε représente l'effet des autres facteurs (non pris en compte)

Erreurs de régression

erreur «ordinaire»
x est parfaitement connu, toute l'erreur est sur y



erreur orthogonale
l'erreur est partagée entre x et y



- modèle $y = f(x) + \varepsilon$ implique une erreur ordinaire
- l'erreur ordinaire est une hypothèse à valider

Agrégation des erreurs de régression

- enjeu : minimiser les erreurs, donc la norme de ε
- norme L_1 : $||\varepsilon||_1 = \sum_{i=1}^{n} |\varepsilon_i|$
  - poids des erreurs proportionnel à la valeur
  - non dérivable en zéro, donc peu adaptée à l'optimisation
- norme L_2 : $||\varepsilon||_2 = \sqrt{\sum_{i=1}^{n} \varepsilon_i^2}$
  - privilégie plusieurs petites erreurs à une grosse
  - très bien adaptée à l'optimisation
  - «moindres carrés»
- norme L_∞ : $||\varepsilon||_\infty = \max_{i=1}^{n} |\varepsilon_i|$
  - erreur «égale» pour tous (idéalement), contrôle du pire cas
  - peu adaptée à l'optimisation
- nous nous restreindrons aux moindres carrés ordinaires

Régression linéaire simple (1)

- on se limite à une unique variable explicative x
- on se limite à une fonction linéaire

Régression linéaire simple

Le modèle de régression linéaire simple est défini par

$$y_i = \beta_1 x_i + \beta_0 + \varepsilon_i, i = 1..n$$

- β_1 et β_0 sont les paramètres inconnus à estimer
- ε_i est l'erreur ou résidu pour l'observation i

Régression linéaire simple (2)

- d'un point de vue statistique (deuxième partie du cours) :
  - y_i est la réalisation d'une variable aléatoire Y_i
  - il y a des hypothèses fortes sur les résidus : indépendants, de même loi, centrés et de même variance σ^2
    - le modèle donne la valeur *moyenne* de Y : $E[Y_i] = E[\beta_1 x_i + \beta_0 + \varepsilon_i] = \beta_1 x_i + \beta_0$
    - σ^2 représente le bruit ou poids des autres facteurs : $var[Y_i] = var[\beta_1 x_i + \beta_0 + \varepsilon_i] = \sigma^2$
- enjeux :
  - estimer β_1 et β_0 ; faire des tests d'hypothèses sur ces paramètres
  - faire des tests d'hypothèses sur la validité du modèle, sur les prédictions

Calcul de β_1 et β_0

- on veut minimiser l'erreur quadratique (moyenne) :

$$\delta^2 = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0)^2$$

- [... quelques calculs ...]

Droite des moindres carrés

Droite des moindres carrés

La droite des moindres carrées est $y = \beta_1 x + \beta_0$ avec

$$\beta_1 = \frac{C_{xy}}{s_x^2} \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (moyenne de x)
- $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ (moyenne de y)
- $s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$ (variance de x)
- $C_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$ (covar. entre x et y)
- β_1 et β_0 sont estimateurs sans biais des variables aléatoires correspondantes

- $y = \bar{y} - \frac{C_{xy}}{s_x^2}(x - \bar{x})$
- la droite passe par le barycentre $(\bar{x}; \bar{y})$

1. Régression linéaire simple

TP : Échauffement

1. Générez `x` $= (1; 2; \ldots; 100)$ et `y` de la forme $y = f(x) + \varepsilon$ on prendra $y = 0.05x^2 - 3x + 7 + \mathcal{N}(\mu = 0, \sigma = 50)$
2. Tracez `y` en fonction de `x` ; donnez un nom aux axes et au graphique.
3. Calculez la variance de `x` et la covariance de `x` et `y` (faire vos propres fonctions). Comparez aux résultats des fonctions `var` et `cov`.
4. *Astuce* : savez-vous afficher la valeur de la variance de `x` avec exactement 3 décimales (`sprintf`) ?

TP : Calculs de base

1. Calculez β_1 et β_0 (faire des fonctions)
2. Définissez deux fonctions qui permettent respectivement de calculer une régression linéaire et de l'évaluer sur des nouvelles données.
3. Calculez la régression linéaire de `y` en fonction de `x`
4. Ajoutez la droite de régression sur le graphique (`abline`).
5. *Astuce* : savez-vous modifier la couleur, l'épaisseur ou le style d'un tracé (`col`, `lwd`, `lty`) ?

TP : Calculs des erreurs

1. Calculez le vecteur des erreurs.
2. Calculez l'erreur quadratique moyenne. Calculez l'erreur moyenne, l'erreur absolue moyenne, l'erreur relative moyenne, l'erreur relative absolue moyenne.
3. Déterminez pour combien et pour quelles valeurs de `x` l'erreur absolue est plus grande que 100.
4. Tracez la courbe $N(e)$ où $N(e)$ représente le nombre de valeurs de valeurs de `x` pour lesquelles l'erreur est supérieure à e.
5. *Astuce* : connaissez-vous `sapply` ?

2. Qualité d'une régression

Évaluation empirique des erreurs

y : vecteur des valeurs mesurées ; $\hat{y}$: vecteur des valeurs prédites

- somme des erreurs quadratiques : $\sum_i (y_i - \hat{y}_i)^2$ (norme $L2$) (moindres carrés)
- somme des erreurs absolues : $\sum_i |y_i - \hat{y}_i|$ (norme $L1$)
- plus grande erreur absolue : $\max_i |y_i - \hat{y}_i|$ (norme L_∞)

- somme des erreurs relatives : $\sum_i \left|\frac{y_i - \hat{y}_i}{y_i}\right|$
- plus grande erreur relative : $\max_i \left|\frac{y_i - \hat{y}_i}{y_i}\right|$

- ...

Coefficient de corrélation linéaire

Coefficient de corrélation

Le coefficient de corrélation linéaire (empirique) entre les x_i et les y_i est

$$r_{xy} = \frac{C_{xy}}{s_x s_y}$$

où C_{xy} est la covariance entre x et y, s_x et s_y sont les écarts-types de x et y

- $\delta^2 = s_y^2(1 - r_{xy}^2)$: minimiser δ est équivalent à maximiser r_{xy}^2
- $r_{xy} = \pm 1$ ssi les points sont parfaitements alignés
- $r_{xy} \in [-1; 1]$

Décomposition de la variance

- Soit $\hat{y}_i = \beta_1 x_i + \beta_0$, et donc $y_i = \hat{y}_i + \varepsilon_i$
- On peut décomposer la variance :

$$\underbrace{y_i - \bar{y}}_{\text{variance}} = \underbrace{\hat{y}_i - \bar{y}}_{\text{expliquée}} + \underbrace{\varepsilon_i}_{\text{non expliquée}}$$

- on appelle variance totale $SST = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$
- on appelle variance expliquée $SSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- on appelle variance résiduelle $SSR = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \delta^2_{\min}$

Décomposition de la variance

$$SST = SSE + SSR$$

Coefficient de détermination (1)

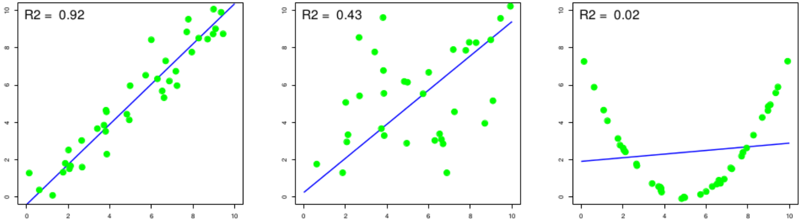
Coefficient de détermination

Le coefficient de détermination de la régression linéaire est

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

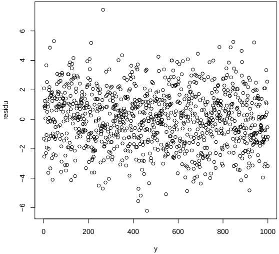
- R^2 est la proportion de la variance de y expliquée par x
  - R^2 proche de 1 : x et y *semblent* en relation linéaire
  - R^2 proche de 0 : x et y ne sont pas liés *linéairement*
- propriété : $R^2 = r_{xy}^2$

Coefficient de détermination (2)

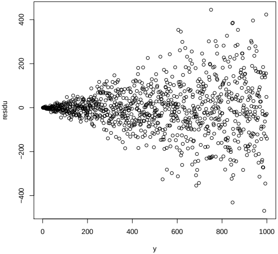


Analyse des résidus

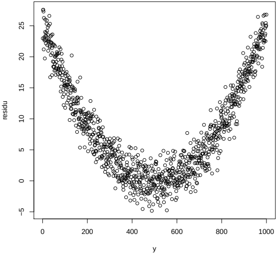
- résidus en fonction de x (ou y ou $\hat{y}$)



normal

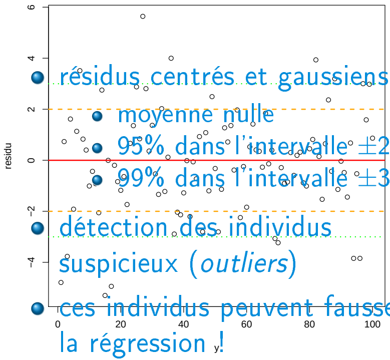


hétéroscédasticité



composante non linéaire

Observations suspicieuses



- résidus centrés et gaussiens
  - moyenne nulle
  - 95% dans l'intervalle $\pm 2\sigma$
  - 99% dans l'intervalle $\pm 3\sigma$
- détection des individus suspicieux (*outliers*)
- ces individus peuvent fausser la régression !

2. Qualité d'une régression

Évaluation de la qualité d'une régression
Travaux pratiques

TP : Qualité de la régression

1. Calculez le coefficient de corrélation linéaire (faire une fonction).
2. Calculez la variance totale, la variance expliquée, la variance résiduelle (faire des fonctions). Vérifiez la décomposition de la variance.
3. Calculez le coefficient de détermination (faire une fonction).
4. *Astuce* : savez-vous ajouter ces grandeurs sur le graphique (`text`) ?

TP : Résidus

1. Tracez les résidus en fonction de x.
2. Ces résidus vous semblent-ils «normaux» ? Y a-t-il des points suspicieux ?
3. Calculez la régression $y = a\,x^2 + b$.
4. Est-elle meilleure ?
5. Parmi les régression de la forme $y = a\,x^k + b$, laquelle vous semble la meilleure ?

3. Régression linéaire multiple

Généralisation de la régression linéaire
Travaux pratiques

TP : `lm` (et `predict`)

1. Récupérez le fichier `003_regression_linaire_lm.R`
2. Exécutez-le pas à pas, en prenant le temps de comprendre chaque instruction proposée.
3. Vérifiez que vous savez
   - importer ou générer des données, tracer des graphiques pour les représenter ;
   - calculer une régression linéaire avec une ou plusieurs variables, transformer les variables ;
   - déterminer les caractéristiques essentielles d'une régression linéaire (coefficients, R^2, résidus, ...) ;
   - calculer des prédictions pour les régressions calculées.

Régression linéaire multiple (1)

- on étend la régression linéaire à p variables explicatives
- on se limite à des fonctions linéaires (en chaque variable)

Régression linéaire multiple

Le modèle de régression linéaire simple est défini par

$$y_i = \beta_p x_{p,i} + \ldots + \beta_1 x_{1,i} + \beta_0 + \varepsilon_i, i = 1..n$$

- les β_i sont les paramètres inconnus à estimer
- ε_i est le résidu pour l'observation i

Régression linéaire multiple (2)

- hypothèse : les résidus sont centrés
  - si les résidus sont indépendants, de même loi et même variance σ^2, on parle de moindres carrées ordinaires
  - sinon on parle de moindres carrées généralisés
- il n'y a pas d'hypothèse sur les régresseurs (en particulier, ils n'ont pas à être indépendants)
- écriture matricielle : $Y = X\beta + \varepsilon$

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{11} & \dots & x_{p1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \dots & x_{pn} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Calcul des paramètres

- on veut minimiser l'erreur quadratique (moyenne) : $||Y - X\beta||_2$
- dans le cas des moindres carrés ordinaires β minimise $||Y - X\beta||_2$ ssi

$$X^T X\beta = X^T Y$$

  («équations normales» de la régression linéaires)
- dans le cas des moindres carrés généralisés, il faut faire des hypothèses sur les résidus...

Coefficients de corrélation linéaire

Coefficient de corrélation

Le coefficient de corrélation linéaire multiple entre la variable à expliquer y et les régresseurs x_i est

$$r = \sup_{a_1,\dots,a_p} r_{y,\sum a_j x_j}$$

- r est la valeur maximale prise par le coefficient de corrélation linéaire entre y et une combinaison linéaire des x_j.
- $r \in [0;1]$

- conséquence de la définition : ajouter des variables augmente r (sauf colinéarité)

Coefficient de détermination

Coefficient de détermination

Le coefficient de détermination de la régression linéaire est

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- identique à la régression linéaire simple
- « *multiple R-squared* » sous R
- propriété : $R^2 = r^2$
  corollaire : R^2 augmente avec le nombre de variables...

Coefficient de détermination ajusté

- enjeux :
  - maximiser R^2 n'est pas un bon critère de qualité, car ajouter des variables aléatoires serait «bénéfique»
  - on ne peut pas comparer deux modèles avec un nombre de variables différent
- idée : pénaliser le R^2 en fonction du nombre de variables

Coefficient de détermination ajusté

Le coefficient de détermination ajusté de la régression linéaire est

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

- $n-1$: degrés de liberté de SST
- $n-p-1$: degrés de liberté de SSE

Conclusion partielle

- nous savons calculer des modèles de régression linéaire
- nous savon évaluer ces modèles, dans une certaine mesure

- encore beaucoup à dire sur la construction et l'interprétation de modèles linéaires
  - autres critères de qualité (que le R^2_{adj})
  - tests statistiques sur la significativité d'une variable
  - méthodes de construction et de sélection des variables
  - analyse des erreurs, intervalles de confiance des prédictions
  - …

3. Régression linéaire multiple

Généralisation de la régression linéaire
Travaux pratiques

TP : Consommation de véhicules (1)

Jeu de données `mpg.dat` (UCI Machine Learning Repository) :

- `mpg` (*miles per galon*) : la distance parcourue
- `cylinders` : le nombre de cylindres.
- `displacement` (pouces-cube) : la cylindrée
- `horsepower` : puissance du moteur
- `weight` (livres) : poids du véhicule
- `acceleration` : secondes pour atteindre 100 m/h
- `model year` : année du modèle.
- `origin` : lieu (1 - USA ; 2 - EU ; 3 - Japon).
- `name` : nom du modèle (unique).

TP : Consommation de véhicules (2)

TP noté !

- à faire en binôme.
- à rendre sur Chamilo le dimanche 6 mars 2016.
- rendu : script R commenté.
  - Les commandes permettant de répondre à une question doivent être fournies, à l'exclusion de toutes autres.
  - Les graphiques doivent être soignés, clairs et complets.
  - Une réponse explicite, pour chaque question, est requise. Cette réponse sera rédigée convenablement (hormis fonctions ou graphiques demandés).