

Modèle linéaire

Moh A. Sedki

Université Paris-Sud

17 octobre 2017

Contenu

Données publicitaires

Les données publicitaires affichent des ventes (en milliers d'unités) pour un produit particulier en fonction des budgets publicitaires (en milliers de dollars) pour la télévision, la radio et les journaux. Supposons que, dans notre rôle de consultant en statistique, nous souhaitons mettre en place un plan de marketing pour l'année prochaine qui aboutira à des ventes élevées du produit. Quelles informations seraient utiles pour fournir un tel plan ? Voici quelques questions importantes que nous pourrions aborder :

Question

Existe-t-il une relation entre le budget publicitaire et les ventes ?

Notre premier objectif est de déterminer si les données témoignent d'une association entre les dépenses publicitaires et les ventes. Si la preuve est faible, on pourrait recommander qu'aucun budget ne devrait être consacré à la publicité !

Questions

- ▶ Quelle est la relation entre le budget publicitaire et les ventes ?
- ▶ En supposant qu'il existe une relation entre la publicité et les ventes, nous aimerions connaître la force de cette relation. Autrement dit, compte tenu d'un certain budget publicitaire, pouvons-nous prédire les ventes avec un haut niveau de précision ?
- ▶ Cette relation est-elle forte ou une prédiction des ventes basées uniquement sur le budget publicitaire serait légèrement mieux qu'une hypothèse aléatoire (relation faible) ?
- ▶ Quels médias contribuent aux ventes ? et comment ?
- ▶ La relation est-elle linéaire ? S'il y a approximativement une relation linéaire entre les dépenses publicitaires dans les différents médias et les ventes, alors la régression linéaire est appropriée. Sinon, il est encore possible de transformer les variables pour que la régression linéaire puisse être utilisée.
- ▶ A-t-on des interactions entre les différentes dépenses publicitaires ?

De quoi s'agit-il ?

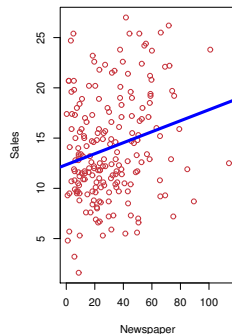
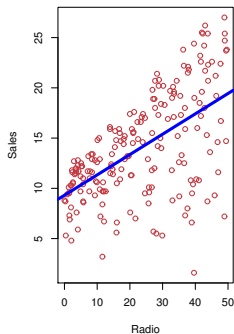
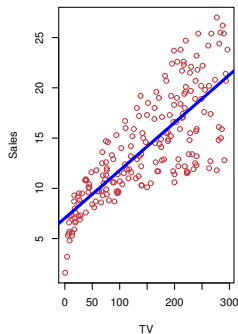
Une régression (*regression analysis* en anglais) est une méthode pour étudier la relation fonctionnelle entre deux variables. Dans la partie régression linéaire multiple, nous allons étudier la modélisation de la relation entre plusieurs variables (> 2).

De quoi s'agit-il ?

Une régression (*regression analysis* en anglais) est une méthode pour étudier la relation fonctionnelle entre deux variables. Dans la partie régression linéaire multiple, nous allons étudier la modélisation de la relation entre plusieurs variables (> 2).

Nous considérons la modélisation de la relation entre deux variables par une ligne droite, c'est à dire lorsque une variable Y est modélisée comme une fonction linéaire d'une autre variable X .

Modèle de régression pour les données publicitaires



Le nuage de points peut parfois nous donner une idée du type de relation entre les variables (linéaire, quadratique, exponentielle, ...)

Modèle de régression linéaire simple

Le jeu de données est noté

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

- ▶ x_i est la $i^{\text{ème}}$ valeur de la variable X dite *explicative* ou *covariable*.
- ▶ y_i est la $i^{\text{ème}}$ valeur de la variable Y dite *réponse* ou *variable à expliquer*.

Modèle de régression linéaire simple

Le jeu de données est noté

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

- ▶ x_i est la $i^{\text{ème}}$ valeur de la variable X dite *explicative* ou *covariable*.
- ▶ y_i est la $i^{\text{ème}}$ valeur de la variable Y dite *réponse* ou *variable à expliquer*.

Deux situations pour la collecte des valeurs x_1, x_2, \dots, x_n associées variable X

- ▶ Elles sont observées comme dans le cas des données de production.
- ▶ Quand il s'agit d'une étude, elles peuvent être fixées par cette étude.

Régression linéaire : formalisme

La régression linéaire est typiquement utilisée pour modéliser la relation entre deux variables Y et X telle que pour une certaine valeur spécifique de $X = x$, on peut prédire la valeur de Y . De manière formelle, la régression d'une variable aléatoire Y sur une variables aléatoire X est

$$\mathbb{E}(Y \mid X = x),$$

la valeur moyenne de Y lorsque $X = x$.

Par exemple,

- ▶ X un budget publicité TV
- ▶ Y les ventes du produit.

La régression de Y sur X représente la vente moyenne (l'espérance des ventes) pour un budget publicité donné.

Régression linéaire : formalisme

La régression de Y sur X est linéaire si

$$\mathbb{E}(Y \mid X = x) = \beta_0 + \beta_1 x$$

où β_0 et β_1 désignent l'*intercept* et la *pente* de la droite de régression.

Supposons que Y_1, Y_2, \dots, Y_n sont des réalisations indépendantes de la variable aléatoire Y observées aux points x_1, x_2, \dots, x_n de la variables aléatoires X . Si la régression de Y sur X est linéaire, alors pour $i = 1, 2, \dots, n$

$$Y_i = \mathbb{E}(Y \mid X = x) + e_i = \beta_0 + \beta_1 x + e_i$$

où les e_i sont des erreurs aléatoires associées aux Y_i telles que $\mathbb{E}(e_i \mid X) = 0$.

Régression linéaire : le terme d'erreur

- ▶ L'ajout du terme d'erreur e_i est du à la nature aléatoire de Y . Il y a certainement une certaine variation dans Y qui ne peut être prédite ou expliqué.
- ▶ En d'autres termes, toute variation inexpliquée est appelée erreur aléatoire. Ainsi, le terme d'erreur aléatoire e_i ne dépend pas de x et ne contient aucune information sur Y (sinon on parlera de terme d'erreur systématique).

On supposera que,

$$\text{Var}(Y \mid X = x) = \sigma^2.$$

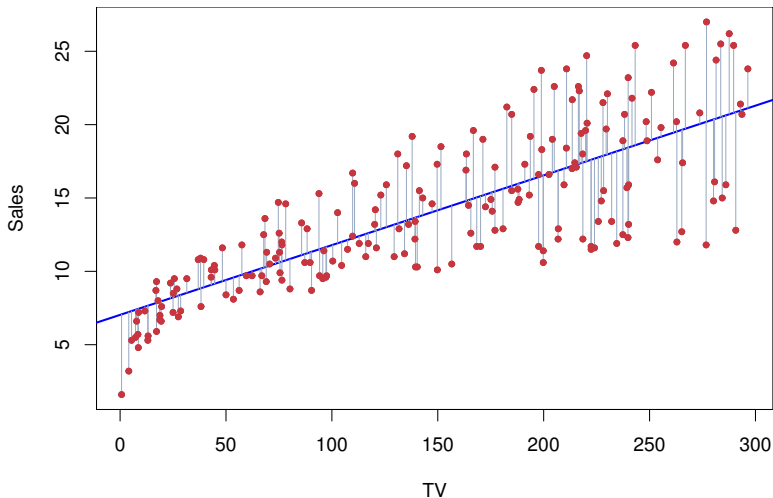
Estimation de β_0 et β_1 de la population

- ▶ Supposons que X est la taille, Y est le poids d'un individu choisi aléatoirement dans une population.
- ▶ Dans un modèle de régression linéaire simple, le poids moyen d'un individu pour une taille donnée est une fonction linéaire de la taille.
- ▶ En pratique, nous n'avons qu'un échantillon de données (un certain nombre de couples (x_i, y_i)) au lieu de la population entière.
- ▶ Les valeurs de β_0 et β_1 sont évidemment inconnues.
- ▶ On souhaite utiliser les données pour estimer l'intercept (β_0) et la pente (β_1).
- ▶ Cela peut être obtenu via une droite qui ajuste au mieux nos données, c'est à dire des valeurs b_0 et b_1 telles que $\hat{y}_i = b_0 + b_1 x_i$ soit aussi proche que possible de y_i .
- ▶ La notation \hat{y}_i est utilisée pour noter la valeur donnée par la droite ajustée pour la distinguer de la valeur observée y_i . On dit que \hat{y}_i est la valeur *prédite* ou *ajustée* de y_i .

Les résidus

En pratique, on veut minimiser la différence entre les valeurs de $y(y_i)$ et les valeurs prédites $\hat{y}(\hat{y}_i)$. Cette différence est appelée le résidu, notée \hat{e}_i ,

$$\hat{e}_i = y_i - \hat{y}_i.$$



Meilleur ajustement au sens des moindres carrées

La méthode couramment utilisée pour choisir les valeurs b_0 et b_1 s'appelle ***méthode des moindres carrés***. Comme son nom l'indique, b_0 et b_1 sont choisis pour minimiser la somme des résidus notée RSS (*residual sum of squares*),

$$\text{RSS} = \sum_{i=1}^n \hat{e}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2.$$

Meilleur ajustement au sens des moindres carrées

La méthode couramment utilisée pour choisir les valeurs b_0 et b_1 s'appelle **méthode des moindres carrés**. Comme son nom l'indique, b_0 et b_1 sont choisis pour minimiser la somme des résidus notée RSS (*residual sum of squares*),

$$\text{RSS} = \sum_{i=1}^n \hat{e}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2.$$

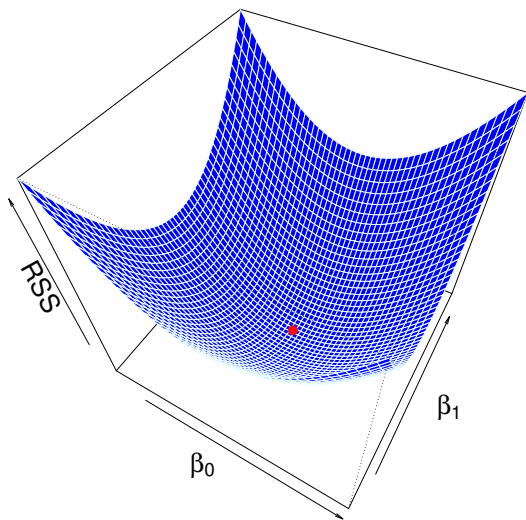
Pour que la somme des moindres carrés RSS soit minimale, il faut que b_0 et b_1 vérifient les équations nous avons les équations

$$\frac{\partial \text{RSS}}{\partial b_0} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) = 0$$

et

$$\frac{\partial \text{RSS}}{\partial b_1} = -2 \sum_{i=1}^n x_i (y_i - b_0 - b_1 x_i) = 0$$

Solution des moindres carrés



Solution des moindres carrés

En réarrangeant les équations précédentes, on obtient

$$\sum_{i=1}^n y_i = b_0 n + b_1 \sum_{i=1}^n x_i$$

et

$$\sum_{i=1}^n x_i y_i = b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2$$

dites *équations normales*.

Solution des moindres carrés

En réarrangeant les équations précédentes, on obtient

$$\sum_{i=1}^n y_i = b_0 n + b_1 \sum_{i=1}^n x_i$$

et

$$\sum_{i=1}^n x_i y_i = b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2$$

dites *équations normales*.

La solution du système composé des deux équations précédentes en b_0 et b_1 nous donne les **estimateurs par les moindres carrés** de l'intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

et la pente

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{XY}}{S_{XX}}.$$

Estimation de la variance du terme d'erreur aléatoire

Revenons au modèle de régression linéaire à variance constante décrit précédemment,

$$Y_i = \beta_0 + \beta_1 x_i + e_i \quad (i = 1, 2, \dots, n)$$

où le l'erreur aléatoire e_i est de moyenne 0 et de variance σ^2 . On veut estimer $\sigma^2 = \text{Var}(e)$. Notons que

$$e_i = Y_i - (\beta_0 + \beta_1 x_i) = Y_i - \text{la droite de régression en } x_i \quad \text{inconnue}$$

Puisque β_0 et β_1 sont inconnues, le mieux qu'on puisse faire, c'est de remplacer ces paramètres inconnus par $\hat{\beta}_0$ et $\hat{\beta}_1$ pour obtenir

$$\hat{e}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = Y_i - \text{la droite de régression estimée au point } x_i.$$

Estimateur sans biais de σ^2

Les résidus peuvent être utilisés pour estimer σ^2 . En effet, on peut montrer que

$$S^2 = \frac{\text{RSS}}{n-2} = \frac{1}{n-2} \sum_{i=1}^n \hat{e}_i^2$$

est un estimateur sans biais de σ^2 .

Estimateur sans biais de σ^2

Les résidus peuvent être utilisés pour estimer σ^2 . En effet, on peut montrer que

$$S^2 = \frac{\text{RSS}}{n-2} = \frac{1}{n-2} \sum_{i=1}^n \hat{e}_i^2$$

est un estimateur sans biais de σ^2 .

Deux remarques à noter

- ▶ $\bar{\hat{e}} = 0$ (puisque $\sum_{i=1}^n \hat{e}_i = 0$ car la droite des moindres carrés minimise $\text{RSS} = \sum_{i=1}^n \hat{e}_i^2$)
- ▶ Dénominateur dans S^2 est $n-2$ (nous avons estimé deux paramètres β_0 et β_1)

Plan

Dans cette partie, nous allons étudier des méthodes pour

- ▶ Calculer des intervalles de confiance
- ▶ Construire des tests d'hypothèses sur la pente et l'intercept

Hypothèses minimales sur le modèle de régression

Dans cette partie, on considère les hypothèses suivantes

- ▶ Y est lié à x par un modèle de régression linéaire simple

$$Y_i = \beta_0 + \beta_1 x_i + e_i (i = 1, \dots, n), \text{ i.e. } \mathbb{E}(Y \mid X = x_i) = \beta_0 + \beta_1 x_i$$

- ▶ Les erreurs e_1, e_2, \dots, e_n sont indépendantes l'une de l'autre
- ▶ Les erreurs e_1, e_2, \dots, e_n ont la même variance σ^2
- ▶ Les erreurs sont gaussiennes de moyenne 0 et variance σ^2 , i.e.

$$e \mid X \sim \mathcal{N}(0, \sigma^2)$$

Hypothèses minimales sur le modèle de régression

Dans cette partie, on considère les hypothèses suivantes

- ▶ Y est lié à x par un modèle de régression linéaire simple

$$Y_i = \beta_0 + \beta_1 x_i + e_i (i = 1, \dots, n), \text{ i.e. } \mathbb{E}(Y \mid X = x_i) = \beta_0 + \beta_1 x_i$$

- ▶ Les erreurs e_1, e_2, \dots, e_n sont indépendantes l'une de l'autre
- ▶ Les erreurs e_1, e_2, \dots, e_n ont la même variance σ^2
- ▶ Les erreurs sont gaussiennes de moyenne 0 et variance σ^2 , i.e.

$$e \mid X \sim \mathcal{N}(0, \sigma^2)$$

Nous verrons (prochaines séances) des techniques de diagnostics pour vérifier ces hypothèses. De plus, puisque le modèle de régression est conditionnel à X . Les covariables x_1, x_2, \dots, x_n peuvent être supposées constantes.

Une expression commode de $\hat{\beta}_1$

Rappelons que l'estimateur de β_1 au sens des moindres carrés,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{XY}}{S_{XX}}$$

Or on sait que $\sum_{i=1}^n (x_i - \bar{x}) = 0$, ainsi

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x})y_i - \bar{y} \sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n (x_i - \bar{x})y_i.$$

Une expression commode de $\hat{\beta}_1$

Rappelons que l'estimateur de β_1 au sens des moindres carrés,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{SXY}{SXX}$$

Or on sait que $\sum_{i=1}^n (x_i - \bar{x}) = 0$, ainsi

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x})y_i - \bar{y} \sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n (x_i - \bar{x})y_i.$$

On obtient l'expression de $\hat{\beta}_1$

$$\hat{\beta}_1 = \sum_{i=1}^n c_i y_i \text{ où } c_i = \frac{x_i - \bar{x}}{SXX}$$

Espérance et variance de $\hat{\beta}_1$

Sous les hypothèses précédentes, nous avons

$$\mathbb{E}(\hat{\beta}_1 \mid X) = \beta_1$$

et

$$\text{Var}(\hat{\beta}_1 \mid X) = \frac{\sigma^2}{\text{SXX}}$$

Loi de $\hat{\beta}_1$

Sous les hypothèses précédentes, nous avons

$$\hat{\beta}_1 \mid X \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\text{SXX}}\right)$$


Notons que la variance de $\hat{\beta}_1$ diminue lorsque la somme des carrés SXX augmente (*i.e.* lorsque la variabilité des X augmente). Cela souligne la nécessité d'avoir un large choix de valeurs de X .

Loi de $\hat{\beta}_1$

Sous les hypothèses précédentes, nous avons

$$\hat{\beta}_1 \mid X \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\text{SXX}}\right)$$

Notons que la variance de $\hat{\beta}_1$ diminue lorsque la somme des carrés SXX augmente (*i.e.* lorsque la variabilité des X augmente). Cela souligne la nécessité d'avoir un large choix de valeurs de X .

Centrage + réduction 

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{\text{SXX}}} \sim \mathcal{N}(0, 1)$$

Une loi commode de $\hat{\beta}_1$

Si σ est connu, on peut utiliser la statistique Z pour construire un test et déduire un intervalle de confiance pour β_1 . Comme σ est souvent inconnu en pratique, il est remplacé par S , l'écart type des résidus dans

$$T = \frac{\hat{\beta}_1 - \beta_1}{S / \sqrt{\text{SXX}}} = \frac{\hat{\beta}_1 - \beta_1}{\text{se}(\hat{\beta}_1)}$$

où $\text{se}(\hat{\beta}_1) = S / \sqrt{\text{SXX}}$ est l'estimateur de l'écart type de $\hat{\beta}_1$, calculé automatiquement par le logiciel R.

Statistique de test pour $\hat{\beta}_1$

On peut montrer sous les hypothèses précédentes que T est de loi de student à $n - 1$ degrés de liberté

$$T = \frac{\hat{\beta}_1 - \beta_1}{\text{se}(\hat{\beta}_1)} \sim t_{n-2}$$

Notons que le degré de liberté provient de la formule

degré de liberté = taille de l'échantillon - nombre de paramètres estimés.

Test d'égalité à 0 de $\hat{\beta}_1$

Pour tester $H_0 : \beta_1 = \beta_1^0$, la statistique de test est

$$T = \frac{\hat{\beta}_1 - \beta_1^0}{\text{se}(\hat{\beta}_1)} \sim t_{n-2} \text{ lorsque } H_0 \text{ est vraie.}$$

R calcul la valeur de T et la p -value associé au test $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$. Pour l'exemple des données de publicité, $T = 3.985$, et la p -value est 9.48×10^{-5} .

Intervalle de confiance pour β_1

Un intervalle de confiance à $100(1 - \alpha)\%$ de β_1 est donné par

$$[\hat{\beta}_1 - t(\alpha/2, n - 2)\text{se}(\hat{\beta}_1), \hat{\beta}_1 + t(\alpha/2, n - 2)\text{se}(\hat{\beta}_1)]$$

où $t(\alpha/2, n - 2)$ est le quantile d'ordre $100(1 - \alpha/2)$ de la loi de student à $n - 2$ degrés de liberté.

Espérance, variance et loi de $\hat{\beta}_0$

Rappelons que l'estimateur par les moindres carrés de β_0

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

On montre sous les hypothèses précédentes que


$$\mathbb{E}(\hat{\beta}_0 \mid X) = \beta_0,$$

$$\text{Var}(\hat{\beta}_0 \mid X) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\text{SXX}} \right),$$

et

$$\hat{\beta}_0 \mid X \sim \mathcal{N} \left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\text{SXX}} \right) \right),$$

Statistique de test pour β_0

Centrer et réduire 

$$Z = \frac{\hat{\beta}_0 - \beta_0}{\sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\text{SXX}}}} \sim \mathcal{N}(0, 1)$$

Estimation σ par S

$$T = \frac{\hat{\beta}_0 - \beta_0}{S \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\text{SXX}}}} = \frac{\hat{\beta}_0 - \beta_0}{\text{se}(\hat{\beta}_0)} \sim t_{n-2}$$

où $\text{se}(\hat{\beta}_0) = S \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\text{SXX}}}$ est l'écart type estimé de $\hat{\beta}_0$ directement donné dans R.

Test d'égalité à 0 de $\hat{\beta}_0$

Pour tester $H_0 : \beta_0 = \beta_0^0$, la statistique de test est

$$T = \frac{\hat{\beta}_0 - \beta_0^0}{\text{se}(\hat{\beta}_0)} \sim t_{n-2} \text{ lorsque } H_0 \text{ est vraie.}$$

R calcul la valeur de T et la p -value associé au test $H_0 : \beta_0 = 0$ contre $H_1 : \beta_0 \neq 0$. Pour l'exemple des données de publicité, $T = 7.787$, et la p -value est 3.76×10^{-13} .

Intervalle de confiance pour β_0

Un intervalle de confiance à $100(1 - \alpha)\%$ de β_0 est donné par

$$[\hat{\beta}_0 - t(\alpha/2, n - 2)\text{se}(\hat{\beta}_0), \hat{\beta}_0 + t(\alpha/2, n - 2)\text{se}(\hat{\beta}_0)]$$

où $t(\alpha/2, n - 2)$ est le quantile d'ordre $100(1 - \alpha/2)$ de la loi de student à $n - 2$ degrés de liberté.

Prévision

Un des buts de la régression est de proposer des prévisions pour la variable à expliquer Y . Soit x_{n+1} une nouvelle valeur de la variable X , nous voulons prédire Y_{n+1} . Le modèle indique

$$Y_{n+1} = \beta_0 + \beta_1 x_{n+1} + e_{n+1},$$

où $\mathbb{E}(e_{n+1}) = 0$, $\text{Var}(e_{n+1}) = \sigma^2$ et $\text{Cov}(e_{n+1}, e_i) = 0$ pour $1 \leq i \leq n$

Nous avons

$$\hat{y}_{n+1}^p = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}$$

de variance

$$\text{Var}(\hat{y}_{n+1}^p) = \sigma^2 \left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\text{SXX}} \right]$$

- ▶ Notation \hat{y}_{n+1}^p pour insister sur la notion de prévision (non ajustée).
- ▶ Cette variance nous donne une idée de la stabilité de l'estimation.

Erreur de prévision

En prévision, on s'intéresse à l'erreur que l'on commet entre la vraie valeur à prévoir Y_{n+1} et celle que l'on prévoit \hat{y}_{n+1}^p . Cette erreur permet de quantifier la capacité du modèle à prévoir.

$$\mathbb{E}(Y_{n+1} - \hat{y}_{n+1}^p) = 0$$

et

$$\text{Var}(Y_{n+1} - \hat{y}_{n+1}^p) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\text{SXX}} \right]$$

$$Y_{n+1} - \hat{y}_{n+1}^p \sim \mathcal{N} \left(0, \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\text{SXX}} \right] \right)$$

Intervalle de prévision

Estimation de σ par S

$$T = \frac{Y_{n+1} - \hat{y}_{n+1}^p}{S \sqrt{\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\text{SXX}}\right)}} \sim t_{n-2}$$

Intervalle de prévision

Estimation de σ par S

$$T = \frac{Y_{n+1} - \hat{y}_{n+1}^p}{S \sqrt{\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\text{SXX}}\right)}} \sim t_{n-2}$$

Un intervalle de prévision à $100(1 - \alpha)\%$ pour Y_{n+1} , la valeur de Y au point $X = x_{n+1}$ est donné par

$$\hat{y}_{n+1}^p \pm t(\alpha/2, n-2) S \sqrt{\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\text{SXX}}\right)}$$

Décomposition de la variabilité

Montrons que

$$\sum_{i=1}^n \hat{e}_i = 0.$$

Décomposition de la variabilité

Montrons que

$$\sum_{i=1}^n \hat{e}_i = 0.$$

On note

- ▶ La somme des carrés totale $SST = SYY = \sum_{i=1}^n (y_i - \bar{y})^2$
- ▶ La somme des carrés résiduelle $RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$
- ▶ La somme des carrés expliquée par la régression
 $SSreg = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$

Montrons que

$$SST = SSreg + RSS$$

Plan et objectif

L'objectif principal de cette partie est de comprendre ce qui se passe réellement lorsque les hypothèses standards du modèle de régression ne sont pas vérifiées, et ce qui devrait être fait en réponse à chaque situation.

4 jeux de données

- ▶ On considère les 4 jeux de données construits par *Anscombe(1973)*¹
- ▶ Nous montrons qu'il ne suffit pas de lire la sortie (R ou SAS etc.) d'une régression.
- ▶ De fausses conclusions et un faux modèle (très éloigné de la réalité).

1. Anscombe(1973) Graphs in statistical analysis. The American Statistician

4 jeux de données

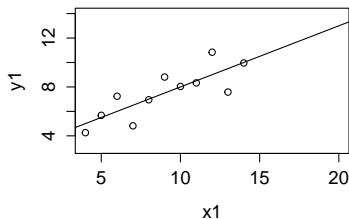
- ▶ On considère les 4 jeux de données construits par *Anscombe(1973)*¹
- ▶ Nous montrons qu'il ne suffit pas de lire la sortie (R ou SAS etc.) d'une régression.
- ▶ De fausses conclusions et un faux modèle (très éloigné de la réalité).

case	x_1	x_2	x_3	x_4	y_1	y_2	y_3	y_4
1	10	10	10	8	8.04	9.14	7.46	6.58
2	8	8	8	8	6.95	8.14	6.77	5.76
3	13	13	13	8	7.58	8.74	12.74	7.71
4	9	9	9	8	8.81	8.77	7.11	8.84
5	11	11	11	8	8.33	9.26	7.81	8.47
6	14	14	14	8	9.96	8.10	8.84	7.04
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

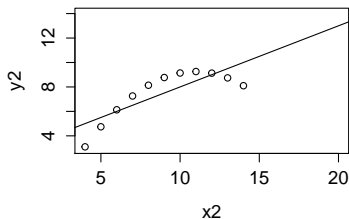
1. Anscombe(1973) Graphs in statistical analysis. The American Statistician

Les nuages de points

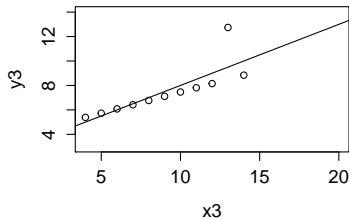
Data Set 1



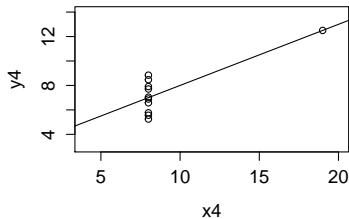
Data Set 2



Data Set 3



Data Set 4



Remarques

Sur les 4 jeux de données, on a la même droite de régression

$$\hat{y} = 3.0 + 0.5x.$$

Remarques

Sur les 4 jeux de données, on a la même droite de régression

$$\hat{y} = 3.0 + 0.5x.$$

- ▶ Ces exemples montrent qu'une sortie numérique d'un logiciel ne suffit pas et doit être accompagnée d'une analyse qui vérifie que le modèle convient aux données.
- ▶ Les outils graphiques peuvent être utilisés.
- ▶ Plus d'une covariable dans le modèle → développement d'autres outils.

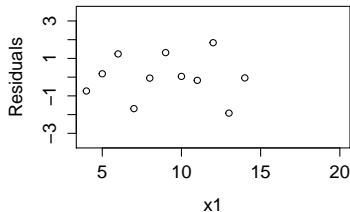
Un outil basé sur les résidus

- ▶ Visualisation des résidus (ou résidus *standardisés* que nous allons définir).
- ▶ Le nombre de covariables du modèle importe peu.

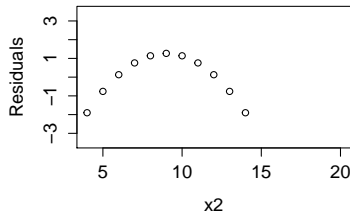
Un outil basé sur les résidus

- Visualisation des résidus (ou résidus *standardisés* que nous allons définir).
- Le nombre de covariables du modèle importe peu.

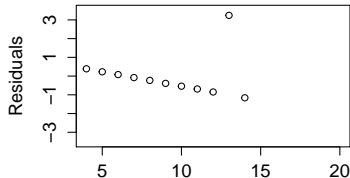
Data Set 1



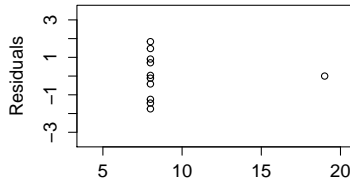
Data Set 2



Data Set 3



Data Set 4



Points influents *leverage points*

Un ***point influent*** est un point dont la valeur de x est éloignée des valeurs des abscisses dans le reste du jeu de données.

Points influents *leverage points*

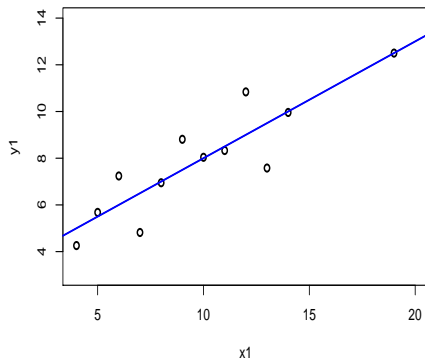
Un ***point influent*** est un point dont la valeur de x est éloignée des valeurs des abscisses dans le reste du jeu de données.

- ▶ Un ***mauvais*** point influent est un point influent dont la valeur de y ne suit pas la tendance du nuage de points. Autrement dit, un mauvais point influent est un point influent aberrant (*outlier*).
- ▶ Un ***bon*** point influent est un point influent dont la valeur de y suit de près la tendance du nuage de points. Autrement dit, un bon point influent est un point influent non aberrant.

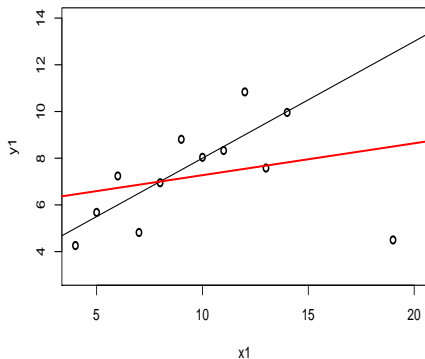
Exemple de bon et mauvais points influents

Revenons au jeu de données 1 de Anscombe(1973)

Good leverage point



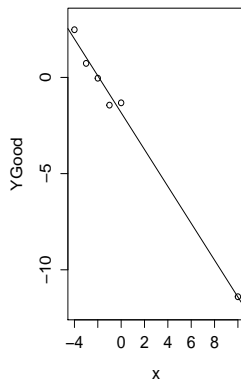
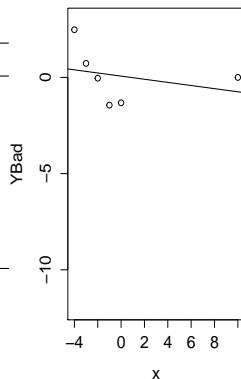
Bad leverage point



Exemple de *Huber(1981, pp. 153-155)*

Cet exemple a été construit pour illustrer les deux situations² précédentes

x	YBad	x	YGood
-4	2.48	-4	2.48
-3	0.73	-3	0.73
-2	-0.04	-2	-0.04
-1	-1.44	-1	-1.44
0	-1.32	0	-1.32
10	0.00	10	-11.40



Voir les différences sur les sorties R!!!

Identification des points influents

On veut établir une règle numérique qui permet d'identifier les points influents (voire de grande influence). Cette règle sera basée sur

- ▶ La distance entre x_i et la grande masse des x .
- ▶ La connaissance de l'endroit où la droite de régression est attirée par un point donné.

Identification des points influents

Rappelons que

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

où $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ et $\hat{\beta}_1 = \sum_{j=1}^n c_j y_j$ où $c_j = \frac{x_j - \bar{x}}{\text{SXX}}$. Nous avons

$$\begin{aligned}\hat{y}_i &= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i \\ &= \bar{y} + \hat{\beta}_1 (x_i - \bar{x}) \\ &= \frac{1}{n} \sum_{j=1}^n y_j + \sum_{j=1}^n \frac{(x_j - \bar{x})}{\text{SXX}} y_j (x_i - \bar{x}) \\ &= \sum_{j=1}^n \left[\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\text{SXX}} \right] y_j \\ &= \sum_{j=1}^n h_{ij} y_j \quad \text{où } h_{ij} = \left[\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\text{SXX}} \right]\end{aligned}$$

Identification des points influents

Notons que

$$\sum_{j=1}^n h_{ij} = \sum_{j=1}^n \left[\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\text{SXX}} \right] = \frac{n}{n} + \frac{(x_i - \bar{x})}{\text{SXX}} \sum_{j=1}^n [x_j - \bar{x}] = 1$$

car $\sum_{j=1}^n [x_j - \bar{x}] = 0$. On peut ainsi exprimer la valeur prédite \hat{y}_i comme

$$\hat{y}_i = h_{ii}y_i + \sum_{j \neq i} h_{ij}y_j,$$

Identification des points influents

Notons que

$$\sum_{j=1}^n h_{ij} = \sum_{j=1}^n \left[\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\text{SXX}} \right] = \frac{n}{n} + \frac{(x_i - \bar{x})}{\text{SXX}} \sum_{j=1}^n [x_j - \bar{x}] = 1$$

car $\sum_{j=1}^n [x_j - \bar{x}] = 0$. On peut ainsi exprimer la valeur prédite \hat{y}_i comme

$$\hat{y}_i = h_{ii}y_i + \sum_{j \neq i} h_{ij}y_j,$$

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2}$$

h_{ii} est communément appelé *l'influence* du $i^{\text{ème}}$ point.

Identification des points influents

Nous avons l'expression

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2}$$

- ▶ Le terme $(x_i - \bar{x})^2$ mesure la distance entre x_i et la moyenne des x
- ▶ h_{ii} mesure l'apport de y_i à \hat{y}_i . Par exemple si $h_{ii} \approx 1$, les autres h_{ij} sont proche de zéro (car $\sum_{j=1}^n h_{ij} = 1$), ainsi

$$\hat{y}_i = 1 \times y_i + \text{les autres termes} \approx y_i.$$

- ▶ Il est clair que

$$\text{average}(h_{ii}) = \frac{2}{n} (i = 1, 2, \dots, n).$$

Identification des points influents

Nous avons l'expression

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2}$$

- ▶ Le terme $(x_i - \bar{x})^2$ mesure la distance entre x_i et la moyenne des x
- ▶ h_{ii} mesure l'apport de y_i à \hat{y}_i . Par exemple si $h_{ii} \approx 1$, les autres h_{ij} sont proche de zéro (car $\sum_{j=1}^n h_{ij} = 1$), ainsi

$$\hat{y}_i = 1 \times y_i + \text{les autres termes} \approx y_i.$$

- ▶ Il est clair que

$$\text{average}(h_{ii}) = \frac{2}{n} (i = 1, 2, \dots, n).$$

On identifie (ou on déclare) un point x_i comme ***influant*** si

$$h_{ii} > 2 \times \text{average}(h_{ii}) = 2 \times 2/n = 4/n.$$

Retour à l'exemple de *Huber(1981)*

i	x_i	Leverage h_{ii}
1	-4	0.2897
2	-3	0.2359
3	-2	0.1974
4	-1	0.1744
5	0	0.1667
6	10	0.9359

- ▶ Les valeurs des h_{ii} sont les mêmes pour YGood et YBad.
- ▶ Notons que

$$h_{66} = 0.9359 > 2 \times \text{average}(h_{ii}) = 4/n = 4/6 = 0.67.$$

- ▶ Le point 6 est le seul à être identifié comme influent et les autres sont < 0.67 .
- ▶ Rappelons que les mauvais points influents sont aberrants.
- ▶ Nous allons voir comment détecter les points aberrants et donc les mauvais points influents.

Stratégies

- ▶ **Suppression des points influents aberrants** : une solution simple consiste à supprimer les points influents identifiés et réajuster la droite de régression sur le nouveau jeu de données réduit.
- ▶ **Ajuster un autre modèle de régression** : On peut ajuster un autre modèle de en incluant une nouvelle covariable (des termes de polynômes) ou en transformant Y et/ou x . Par exemple, dans le cas de l'exemple de *Huber(1981)* avec un mauvais point influent, un modèle quadratique ajuste parfaitement le jeu de données.

Régression quadratique dans l'exemple de Huber

Le modèle quadratique est donné par

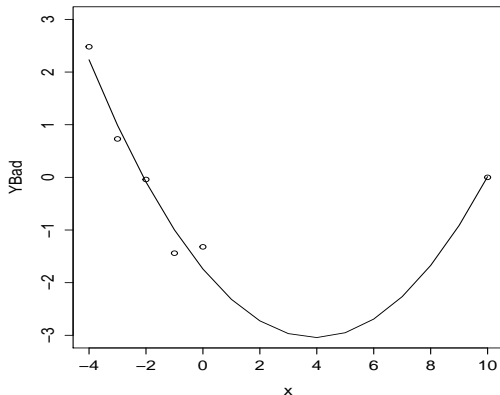
$$\mathbb{E}(Y \mid X = x)\beta_0 + \beta_1 x + \beta_2 x^2.$$

Régression quadratique dans l'exemple de Huber

Le modèle quadratique est donné par

$$\mathbb{E}(Y \mid X = x) \beta_0 + \beta_1 x + \beta_2 x^2.$$

Voir le code R correspondant !!!



Pourquoi

- ▶ Rappelons que nous faisons appel au graphique des résidus pour valider le modèle.
- ▶ On peut montrer que

$$\text{Var}(\hat{e}_i) = \sigma^2 [1 - h_{ii}]$$

où

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{j=1}^n (x_j - \bar{x})^2} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\text{SXX}}.$$

- ▶ Si $h_{ii} \approx 1$ alors le point $i^{\text{ème}}$ est influent, ainsi, le résidu correspondant \hat{e}_i possède une petite variance. Cela est cohérent car si $h_{ii} \approx 1$ alors $\hat{y}_i \approx y_i$.
- ▶ On peut montrer aussi que $\text{Var}(\hat{y}_i) = \sigma^2 h_{ii}$. Ce qui paraît cohérent aussi car lorsque $h_{ii} \approx 1$ alors $\hat{y}_i \approx y_i$ et donc

$$\text{Var}(\hat{y}_i) = \sigma^2 h_{ii} \approx \sigma^2 = \text{Var}(y_i).$$

Les résidus standardisés

Le $i^{\text{ème}}$ *résidu standardisé* est donné par

$$r_i = \frac{\hat{e}_i}{s\sqrt{1 - h_{ii}}}$$

où $s = \sqrt{\frac{1}{n-2} \sum_{j=1}^n \hat{e}_j^2}$ est l'estimateur de σ .

- ▶ Lorsque des points à forte influence existent dans le jeu de données, le graphique des résidus standardisés est meilleur que le graphique des résidus.
- ▶ Lorsqu'il n'y a pas de points à forte influence, il n'y a pas de grandes différences entre les deux graphiques.

Avantages des résidus standardisés

- ▶ L'autre avantage des résidus standardisés est dans l'information qu'ils nous apportent sur l'écart type de chaque point du modèle de régression.
- ▶ Par exemple, supposons qu'un point possède un résidu standardisé de 4.3. Si les erreurs ont une loi normale, l'observation d'un point avec un résidu standardisé de 4.3 semble inhabituel.

Identifier les points aberrants et les mauvais influents

- ▶ Une règle couramment utilisée pour identifier les valeurs aberrantes dans les jeu de données de tailles petites et moyennes consiste à déclarer comme aberrants les points dont le résidu standardisé est hors de l'intervalle $[-2, 2]$.
- ▶ Pour les jeux de données de très grandes tailles (n très grand), on l'utilise l'intervalle $[-4, 4]$.
- ▶ Parmi les points influents, les mauvais points sont ceux dont le résidu standardisé se trouve hors de l'intervalle $[-2, 2]$.

Correlation entre les résidus

En dépit de l'hypothèse d'indépendance des erreurs, Il existe un peu de corrélation entre les résidus

$$\text{Cov}(\hat{e}_i, \hat{e}_j) = -h_{ij}\sigma^2 (i \neq j)$$

et donc

$$\text{Corr}(\hat{e}_i, \hat{e}_j) = \frac{-h_{ij}}{\sqrt{(1-h_{ii})(1-h_{jj})}} (i \neq j).$$

En pratique, cette corrélation est négligeable!!!

Recommandations

- ▶ Les points ne doivent pas être systématiquement supprimés à partir d'une analyse juste parce qu'ils ne correspondent pas au modèle. Les valeurs aberrantes et les points influents peuvent être révélateurs de l'invalidité du modèle.
- ▶ Certains points aberrants dans un modèle ne le sont pas dans un autre.

Normalité des erreurs

Montrons que

$$\hat{e}_i = e_i - \sum_{j=1}^n h_{ij} e_j$$

Normalité des erreurs

Montrons que

$$\hat{e}_i = e_i - \sum_{j=1}^n h_{ij} e_j$$

$$\begin{aligned}\hat{e}_i &= y_i - \hat{y}_i \\ &= y_i - h_{ii} y_i - \sum_{j \neq i} h_{ij} y_j \\ &= y_i - \sum_{j=1}^n h_{ij} y_j \\ &= \beta_0 - \beta_1 x_i + e_i - \sum_{j=1}^n h_{ij} (\beta_0 + \beta_1 x_j + e_j) \\ &= e_i - \sum_{j=1}^n h_{ij} e_j\end{aligned}$$

Voir l'exemple des données de production sous R!!!

Normalité des erreurs(suite)

car $\sum_{j=1}^n h_{ij} = 1$, et

$$\begin{aligned}\sum_{j=1}^n x_j h_{ij} &= \sum_{j=1}^n \left[\frac{x_j}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})x_j}{\text{SXX}} \right] \\ &= \bar{x} + \frac{(x_i - \bar{x})\text{SXX}}{\text{SXX}} = x_i.\end{aligned}$$

Normalité des erreurs(suite)

Nous avons,

$$\hat{e}_i = e_i - \sum_{j=1}^n h_{ij} e_j. \quad (1)$$

On a les deux remarques suivantes

- ▶ Lorsque n est petit, le terme $\sum_{j=1}^n h_{ij} e_j$ domine dans l'expression (1) et résidus peuvent donner l'impression qu'ils viennent d'une loi normale!!!
- ▶ Lorsque n est grand, le premier terme domine dans (1). Ce qui montre que les résidus peuvent être utilisés pour vérifier la normalité des erreurs.

Normalité des erreurs(suite)

Nous avons,

$$\hat{e}_i = e_i - \sum_{j=1}^n h_{ij} e_j. \quad (1)$$

On a les deux remarques suivantes

- ▶ Lorsque n est petit, le terme $\sum_{j=1}^n h_{ij} e_j$ domine dans l'expression (1) et résidus peuvent donner l'impression qu'ils viennent d'une loi normale!!!
- ▶ Lorsque n est grand, le premier terme domine dans (1). Ce qui montre que les résidus peuvent être utilisés pour vérifier la normalité des erreurs.

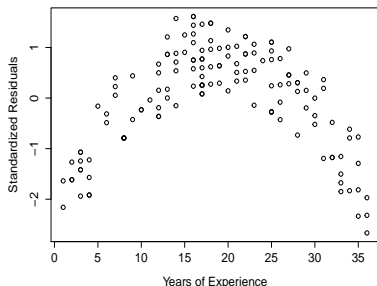
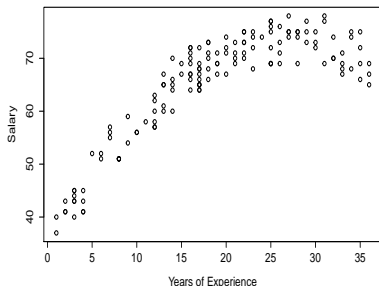
Souvent, pour vérifier la normalité des erreurs, on trace le *normal Q-Q plot* des résidus standardisés.

Plan

Régression polynomiale

Commençons par le cas particulier d'une régression multiple, connue sous le nom de la régression polynomiale. Dans ce cas, les covariables proviennent de l'observation d'une unique covariable x et ses puissances (x^2, x^3, \dots). En régression polynomiale, on peut tracer le résultat sur un graphique.

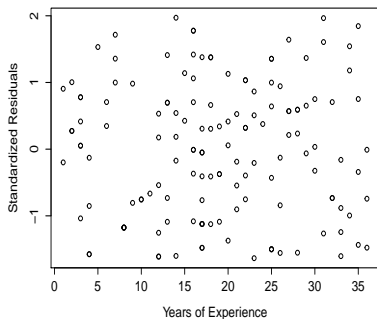
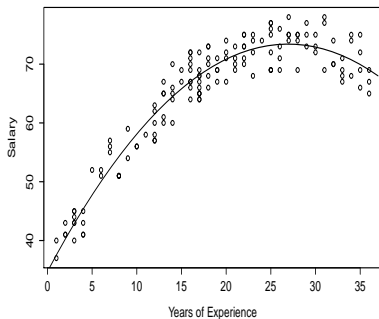
Modélisation du salaire en fonction des années d'expérience :
une droite ?



Régression polynomiale

La tendance dans le nuage de point suggère un modèle de régression polynomiale

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$$



Modèle linéaire multiple

Dans le modèle de régression multiple

$$\mathbb{E}(Y \mid X_1 = x_1, X_2 = x_2, \dots, X_p = x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p.$$

Ainsi,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + e_i,$$

où e_i est l'erreur aléatoire dans Y_i telle que $\mathbb{E}(e_i \mid X) = 0$. Ici, la variable réponse Y (à expliquer) est expliquée par p variables explicatives X_1, X_2, \dots, X_p et la relation entre Y et X_1, X_2, \dots, X_p est linéaire en $\beta_0, \beta_1, \dots, \beta_p$. Dans l'exemple précédent, Y est le salaire, $x_1 = x$ est le nombre d'années d'expérience et $x_2 = x^2$.

Estimateurs des moindres carrés

Les estimateurs des moindres carrés de $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ sont les valeurs de $b_0, b_1, b_2, \dots, b_p$ qui minimisent la somme des carrés des résidus

$$\text{RSS} = \sum_{i=1}^n \hat{e}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \dots - b_p x_{ip})^2.$$

Pour que le RSS soit minimale en $b_0, b_1, b_2, \dots, b_p$,

$$\frac{\partial \text{RSS}}{\partial b_0} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \dots - b_p x_{ip}) = 0$$

$$\frac{\partial \text{RSS}}{\partial b_1} = -2 \sum_{i=1}^n x_{i1} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \dots - b_p x_{ip}) = 0$$

$$\vdots$$

$$\frac{\partial \text{RSS}}{\partial b_p} = -2 \sum_{i=1}^n x_{ip} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \dots - b_p x_{ip}) = 0$$

Écriture matricielle des moindres carrés

Pour pouvoir étudier les propriétés des estimateurs des moindres carrés qu'on notera $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$, nous adoptons des notations vectorielles et matricielles. On définit

- ▶ Le vecteur \mathbf{Y} de dimension $(n \times 1)$.
- ▶ La matrice \mathbf{X} de taille $n \times (p + 1)$.
- ▶ Le vecteur de paramètres inconnus (coefficients de régression) $\boldsymbol{\beta}$ de dimension $(p + 1) \times 1$.
- ▶ Le vecteur des erreurs aléatoires \mathbf{e} de dimension $(n \times 1)$.

Tels que

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \text{ et } \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

Écriture matricielle des moindres carrés

On peut écrire le modèle de régression multiple sous forme d'équation matricielle

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}.$$

De plus, si on note \mathbf{x}'_i la $i^{\text{ème}}$ ligne de la matrice \mathbf{X} . Alors

$$\mathbf{x}'_i = (1, x_{i1}, x_{i2}, \dots, x_{ip})$$

est un vecteur ligne $1 \times (p + 1)$ et on peut écrire

$$\mathbb{E}(Y \mid X = x_i) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} = \mathbf{x}'_i \boldsymbol{\beta}.$$

La somme des carrés des résidus est une fonction de $\boldsymbol{\beta}$ qui s'écrit sous la forme

$$\text{RSS}(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

Solution des moindres carrés

Rappelons que $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$, nous avons donc

$$\begin{aligned}\text{RSS}(\boldsymbol{\beta}) &= \mathbf{Y}'\mathbf{Y} + (\mathbf{X}\boldsymbol{\beta})'\mathbf{X}\boldsymbol{\beta} - \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} - (\mathbf{X}\boldsymbol{\beta})'\mathbf{Y} \\ &= \mathbf{Y}'\mathbf{Y} + \boldsymbol{\beta}'(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} - 2\mathbf{Y}'\mathbf{X}\boldsymbol{\beta}.\end{aligned}$$

Solution des moindres carrés

Rappelons que $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$, nous avons donc

$$\begin{aligned}\text{RSS}(\beta) &= \mathbf{Y}'\mathbf{Y} + (\mathbf{X}\beta)' \mathbf{X}\beta - \mathbf{Y}'\mathbf{X}\beta - (\mathbf{X}\beta)' \mathbf{Y} \\ &= \mathbf{Y}'\mathbf{Y} + \beta' (\mathbf{X}'\mathbf{X}) \beta - 2\mathbf{Y}'\mathbf{X}\beta.\end{aligned}$$

On sait que pour $v \in \mathbb{R}^d$, $a \in \mathbb{R}^d$ et une matrice $M \in \mathbb{R}^{d \times d}$, nous avons

$$\frac{\partial(v'a)}{\partial v} = \frac{\partial(a'v)}{\partial v} = a, \text{ et } \frac{\partial(v'Mv)}{\partial v} = (M + M')v.$$

Solution des moindres carrés

Rappelons que $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$, nous avons donc

$$\begin{aligned}\text{RSS}(\boldsymbol{\beta}) &= \mathbf{Y}'\mathbf{Y} + (\mathbf{X}\boldsymbol{\beta})'\mathbf{X}\boldsymbol{\beta} - \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} - (\mathbf{X}\boldsymbol{\beta})'\mathbf{Y} \\ &= \mathbf{Y}'\mathbf{Y} + \boldsymbol{\beta}'(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} - 2\mathbf{Y}'\mathbf{X}\boldsymbol{\beta}.\end{aligned}$$

On sait que pour $v \in \mathbb{R}^d$, $a \in \mathbb{R}^d$ et une matrice $M \in \mathbb{R}^{d \times d}$, nous avons

$$\frac{\partial(v'a)}{\partial v} = \frac{\partial(a'v)}{\partial v} = a, \text{ et } \frac{\partial(v'Mv)}{\partial v} = (M + M')v.$$

Sachant que $\mathbf{X}'\mathbf{X}$ est symétrique, montrons que la solution des moindres carrés qui minimise $\text{RSS}(\boldsymbol{\beta})$ est

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Valeurs ajustées et résidus

Les valeurs ajustées ou prédites sont

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}.$$

Le vecteur des résidus est

$$\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}.$$

Cas particulier : régression linéaire simple

Revenons à l'équation de régression

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}.$$

Dans le cas d'une régression linéaire simple, \mathbf{X} est donnée par

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

Nous avons

$$\begin{aligned} \mathbf{X}'\mathbf{X} &= \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} \\ &= n \begin{pmatrix} 1 & \bar{x} \\ \bar{x} & \frac{1}{n} \sum_{i=1}^n x_i^2 \end{pmatrix} \end{aligned}$$

Cas particulier : régression linéaire simple

L'inverse de $\mathbf{X}'\mathbf{X}$

$$\begin{aligned}(\mathbf{X}'\mathbf{X})^{-1} &= \frac{1}{n \left(\frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x})^2 \right)} \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \\ &= \frac{1}{\text{SXX}} \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}\end{aligned}$$

Cas particulier : régression linéaire simple

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\&= \frac{1}{\text{SXX}} \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix} \\&= \frac{1}{\text{SXX}} \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n x_i y_i - \bar{x} \sum_{i=1}^n y_i \end{pmatrix} \\&= \begin{pmatrix} \frac{\bar{y} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) - \bar{x} \left(\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} \right)}{\text{SXX}} \\ \frac{\text{SXY}}{\text{SXX}} \end{pmatrix} = \begin{pmatrix} \bar{y} - \frac{\text{SXY}}{\text{SXX}} \bar{x} \\ \frac{\text{SXY}}{\text{SXX}} \end{pmatrix}\end{aligned}$$

Biais

Rappelons que

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \text{ où } \text{Var}(\mathbf{e}) = \sigma^2 \mathbf{I}_n,$$

et l'estimateur des moindres carrés de $\boldsymbol{\beta}$ est donné par

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Biais

Rappelons que

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \text{ où } \text{Var}(\mathbf{e}) = \sigma^2 \mathbf{I}_n,$$

et l'estimateur des moindres carrés de $\boldsymbol{\beta}$ est donné par

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

$$\begin{aligned}\mathbb{E}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}(\mathbf{Y}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}.\end{aligned}$$

Variance

$$\begin{aligned}\text{Var}(\hat{\beta}) &= \text{Var}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{Var}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}_n\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

Variance

$$\begin{aligned}\text{Var}(\hat{\beta}) &= \text{Var}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{Var}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}_n\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

Vérification : nous avons vu que dans le cas simple

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{\text{SXX}} \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

Donc,

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\text{SXX}}$$

Variance des erreurs

La somme des carrés des résidus est

$$\text{RSS} = \text{RSS}(\hat{\beta}) = (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta}) = \mathbf{\hat{e}}'\mathbf{\hat{e}} = \sum_{i=1}^n \hat{e}_i^2.$$

On peut montrer que

$$S^2 = \frac{\text{RSS}}{n - p - 1} = \frac{1}{n - p - 1} \sum_{i=1}^n \hat{e}_i^2$$

est un estimateur sans biais de σ^2 .

Intervalles de confiances et tests

Supposons que les erreurs sont de loi normale de variance constante σ^2 , pour tout $j = 0, 1, \dots, p$

$$T_i = \frac{\hat{\beta}_j - \beta_j}{\text{se}(\hat{\beta}_j)} \sim t_{n-p-1},$$

où $\text{se}(\hat{\beta}_j)$ est l'écart type estimé de $\hat{\beta}_j$ en remplaçant σ par S .

R calcule automatiquement $\text{se}(\hat{\beta}_j)$.

Petit rappel :

degré de liberté = n – nombre de paramètres estimés

Test l'existence d'un lien ...

- ▶ On veut savoir dit s'il y a un lien linéaire entre Y et toutes les covariables (ou un sous-ensemble) X_1, X_2, \dots, X_p si

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + e$$

où certains (ou tous) $\beta_i \neq 0$, $i = 1, \dots, p$. Cela revient à tester

$$\begin{cases} H_0 & : \beta_1 = \beta_2 = \dots = \beta_p = 0 \\ H_A & : \text{il existe au moins un des } \beta_i \neq 0 \end{cases}$$

Encore du vocabulaire

- ▶ La somme des carrés totale $SST = SY = \sum_{i=1}^n (y_i - \bar{y})^2$
- ▶ La somme des carrés résiduelle $RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$
- ▶ La somme des carrés expliquée par la régression
 $SS_{\text{reg}} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$

Décomposition de la variabilité

On sait que

$$\text{SST} = \text{SSreg} + \text{RSS}$$

ou

variabilité totale = variabilité expliquée + variabilité inexpliquée

Construction du test

Si

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + e \text{ et au moins un } \beta_i \neq 0,$$

alors RSS doit être *proche* de 0 et SSreg doit être *proche* de SST.

On doit donc tester

$$\begin{cases} H_0 & : \beta_1 = \beta_2 = \dots = \beta_p = 0 \\ H_A & : \text{il existe au moins un des } \beta_i \neq 0 \end{cases}$$

Pour peut faire appel à la statistique de test

$$F = \frac{\text{SSreg}/p}{\text{RSS}/(n - p - 1)}$$

car RSS est de $(n - p - 1)$ degrés de liberté et SSreg est de p degrés de liberté. Puisque les erreurs théoriques e_1, \dots, e_n sont supposées indépendantes et gaussiennes, on peut montrer que sous H_0 , la statistique F suit une loi de Fisher à $(p, n - p - 1)$ degrés de liberté.

Le R^2 et le R^2_{adj}

- ▶ On fait souvent appel à *coefficient de détermination* R^2 donné par

$$R^2 = \frac{\text{SSreg}}{\text{SST}} = 1 - \frac{\text{RSS}}{\text{SST}}$$

et qui mesure la proportion de variabilité expliquée par le modèle.

- ▶ Inclure des covariables non pertinentes dans le modèle de régression augmente le R^2 . Pour compenser cela, on fait appel au coefficient de détermination ajusté noté R^2_{adj}

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

Notons que $S^2 = \frac{\text{RSS}}{n - p - 1}$ est un estimateur sans biais de

$\sigma^2 = \text{Var}(e_i) = \text{Var}(Y_i)$ alors que $\frac{\text{SST}}{n - 1}$ est un estimateur sans biais de $\sigma^2 = \text{Var}(Y_i)$ lorsque $\beta_1 = \dots = \beta_p = 0$.

Le R^2 et le R^2_{adj}

- ▶ Lorsqu'on compare des modèles avec des différents nombres de covariables, on utilise R^2_{adj} au lieu de R^2 .
- ▶ Le test de Fisher (ou F -test) est utilisé pour tester l'existence d'un lien linéaire entre Y et **certaines** covariables.
- ▶ Si le test de Fisher est significatif (on rejette H_0),

Quelles sont les covariables liées à Y ?

- ▶ On peut répondre à cette question en proposant de faire p **tests de student séparément!!!**
- ▶ Il est difficile d'interpréter les tests lorsque les covariables sont fortement corrélées!!!

Tester un sous-ensemble de coefficients

Supposons qu'on veut tester

$$\begin{cases} H_0 & : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \text{ où } k < p \\ \text{i.e.,} & Y = \beta_0 + \beta_{k+1}x_{k+1} + \cdots + \beta_px_p + e \text{ **modèle réduit** } \end{cases}$$

contre

$$\begin{cases} H_A & : H_0 \text{ **est fausse** } \\ \text{i.e.,} & Y = \beta_0 + \beta_1x_1 + \cdots + \beta_px_p + e \text{ **modèle complet** } \end{cases}$$

Tester un sous-ensemble de coefficients

Supposons qu'on veut tester

$$\begin{cases} H_0 & : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \text{ où } k < p \\ \text{i.e.,} & Y = \beta_0 + \beta_{k+1}x_{k+1} + \cdots + \beta_px_p + e \text{ **modèle réduit** } \end{cases}$$

contre

$$\begin{cases} H_A & : H_0 \text{ **est fausse** } \\ \text{i.e.,} & Y = \beta_0 + \beta_1x_1 + \cdots + \beta_px_p + e \text{ **modèle complet** } \end{cases}$$

On peut utiliser le F -test comme statistique de test

Construction du test

- ▶ On note $\text{RSS}(\text{full})$ la somme des carrés des résidus sous le modèle complet (*i.e.* le modèle impliquant toutes les covariables *i.e.*, H_A)
- ▶ On note $\text{RSS}(\text{reduced})$ la somme des carrés des résidus sous le modèle réduit (*i.e.* le modèle impliquant les covariables avec des coefficients différents de zéro *i.e.*, H_0).

Construction du test

- ▶ On note $\text{RSS}(\text{full})$ la somme des carrés des résidus sous le modèle complet (*i.e.* le modèle impliquant toutes les covariables *i.e.*, H_A)
- ▶ On note $\text{RSS}(\text{reduced})$ la somme des carrés des résidus sous le modèle réduit (*i.e.* le modèle impliquant les covariables avec des coefficients différents de zéro *i.e.*, H_0).

La statistique F est donnée par

$$\begin{aligned} F &= \frac{(\text{RSS}(\text{reduced}) - \text{RSS}(\text{full})) / (\text{df}_{\text{reduced}} - \text{df}_{\text{full}})}{\text{RSS}(\text{full}) / \text{df}_{\text{full}}} \\ &= \frac{(\text{RSS}(\text{reduced}) - \text{RSS}(\text{full})) / k}{\text{RSS}(\text{full}) / (n - p - 1)}, \end{aligned}$$

car le modèle réduit possède $p - k + 1$ covariables et

$$[n - (p + 1 - k)] - [n - (p + 1)] = k.$$

Évaluer un sous ensembles de covariables

On supposera dans la suite que nous avons m covariables au total et on entend par modèle un sous-ensemble de ces covariables de taille p .

Évaluer un sous ensembles de covariables

On supposera dans la suite que nous avons m covariables au total et on entend par modèle un sous-ensemble de ces covariables de taille p .

Rappelons que

$$R^2 = \frac{\text{SSreg}}{\text{SST}} = 1 - \frac{\text{RSS}}{\text{SST}}$$

et

$$R_{\text{adj}}^2 = 1 - \frac{\text{RSS} / (n - p - 1)}{\text{SST} / (n - 1)}$$

où p est le nombre de variables du modèle.

- On sélectionne le modèle avec le R_{adj}^2 le plus élevé, cela revient à choisir le sous-ensemble de variables qui minimise

$$S^2 = \frac{\text{RSS}}{n - p - 1}$$

où p est le nombre de variables du modèle.

Choix basé sur R_{adj}^2

- ▶ Souvent le choix basé sur R_{adj}^2 montre un phénomène de *over-fitting*.
- ▶ Supposons que la valeur maximale de $R_{\text{adj}}^2 = 0.692$ pour un sous-ensemble $p = 10$ de covariables, $R_{\text{adj}}^2 = 0.691$ pour $p = 9$ et $R_{\text{adj}}^2 = 0.541$ pour un sous-ensemble de $p = 8$ covariables.
- ▶ Il est clairement préférable de choisir le modèle à $p = 10$ covariables.

Choix basé sur R_{adj}^2

- ▶ Souvent le choix basé sur R_{adj}^2 montre un phénomène de *over-fitting*.
- ▶ Supposons que la valeur maximale de $R_{\text{adj}}^2 = 0.692$ pour un sous-ensemble $p = 10$ de covariables, $R_{\text{adj}}^2 = 0.691$ pour $p = 9$ et $R_{\text{adj}}^2 = 0.541$ pour un sous-ensemble de $p = 8$ covariables.
- ▶ Il est clairement préférable de choisir le modèle à $p = 10$ covariables.

On va faire appel à un critère(s) basé(s) sur la vraisemblance.

La vraisemblance

Rappelons que d'après les hypothèses du modèle linéaire

$$Y_i \mid x_{i1}, \dots, x_{ip} \sim \mathcal{N}(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}, \sigma^2).$$

Et

$$f(y_i \mid x_{i1}, \dots, x_{ip}) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[-\frac{\left(y_i - \{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}\} \right)^2}{2\sigma^2} \right]$$

La vraisemblance

Rappelons que d'après les hypothèses du modèle linéaire

$$Y_i \mid x_{i1}, \dots, x_{ip} \sim \mathcal{N}(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}, \sigma^2).$$

Et

$$f(y_i \mid x_{i1}, \dots, x_{ip}) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[-\frac{\left(y_i - \{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}\} \right)^2}{2\sigma^2} \right]$$

- Écrire la vraisemblance

$$L(\beta_0, \beta_1, \dots, \beta_p, \sigma^2; y_1, \dots, y_n)$$

- Écrire la log-vraisemblance

$$\ell(\beta_0, \beta_1, \dots, \beta_p, \sigma^2; y_1, \dots, y_n)$$

Mesurer l'ajustement ou l'adéquation du modèle

L'estimateur par maximum de vraisemblance de σ^2 est donné par

$$\hat{\sigma}_{\text{MLE}}^2 = \frac{\text{RSS}}{n}.$$

$$\ell(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p, \hat{\sigma}_{\text{MLE}}^2; y_1, \dots, y_n) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln\left(\frac{\text{RSS}}{n}\right) - \frac{n}{2}$$

Critère d'information

Un critère d'information est une quantité qui réalise un compromis entre l'ajustement aux données (*la vraisemblance par exemple*) et la complexité du modèle.

Donc

- ▶ On peut chercher le modèle qui **maximise**
Ajustement - Pénalité
- ▶ On peut chercher le modèle qui **minimise**
- Ajustement + Pénalité

Critère d'information C_p de Mallows

Le critère d'information noté C_p de Mallows associé au modèle à p covariables est donné par

$$C_p = \frac{\text{RSS}_p}{S^2} + 2p - n$$

où RSS_p est la somme des carrés des résidus du modèle en question et S^2 est l'estimateur de σ^2 dans le modèle complet.

Critère d'information AIC

Le critère d'information noté AIC (*Akaike's Information Criterion*) associé à un modèle à p covariables est donné par

$$\text{AIC} = -2\ell(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p, \hat{\sigma}_{\text{MLE}}^2; y_1, \dots, y_n) + 2K$$

où $K = p + 2$ est nombre de paramètre du modèle.

On peut montrer que

$$\text{AIC} = n \log \hat{\sigma}_{\text{MLE}}^2 + 2p + \text{const.}$$

Ce critère est préférable pour la ***prédiction***.

Critère d'information BIC

Le critère d'information noté BIC (*Bayesian Information Criterion*) associé à un modèle à p covariables est donné par

$$\text{BIC} = -2\ell(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p, \hat{\sigma}_{\text{MLE}}^2; y_1, \dots, y_n) + K \ln(n)$$

où $K = p + 2$ est nombre de paramètre du modèle.

- ▶ Ce critère possède de *bonnes propriétés théoriques*.
- ▶ Ce critère est préférable pour *l'explication*.

Sélection de modèle

L'idée de chercher le modèle (un sous-ensemble de covariables) qui minimise un des trois critères précédents.

Comment procéder ?

Sélection de modèle

L'idée de chercher le modèle (un sous-ensemble de covariables) qui minimise un des trois critères précédents.

Comment procéder ?

Recherche exhaustive : calculer la valeur du critère pour chaque modèle.

Sélection de modèle

L'idée de chercher le modèle (un sous-ensemble de covariables) qui minimise un des trois critères précédents.

Comment précéder ?

Recherche exhaustive : calculer la valeur du critère pour chaque modèle.

Problème combinatoire : $2^{\text{le nombre de covariables}}$ modèles en compétition !!

Sélection de modèle en forward

Le modèle de départ de la procédure de sélection *forward* est le modèle avec la constante seulement. La procédure consiste à

1. Ajouter séparément chaque variable au modèle actuel et calculer le critère d'intérêt (BIC, AIC, ou C_p).
2. Si aucun des nouveaux modèles n'améliore le critère, alors : **stop**.
3. Mettre à jour le modèle en incluant la covariable qui apporte la meilleure amélioration au sens du critère. Aller à 1.

Sélection de modèle backward

Le point de départ de la procédure d'élimination *backward* est le modèle complet incluant toutes les covariables. La procédure consiste à

1. Si aucune élimination d'une covariable n'améliore le critère alors : **stop**.
2. Mettre à jour le modèle en éliminant la covariable qui réalise la meilleure amélioration du critère. Aller à **1**.