Sélection de variables

Analyse multidimensionnelle appliquée

Léo Belzile

HEC Montréal

automne 2022

Modèles prédictifs

Objectif: bâtir un modèle pour une variable réponse Y en fonction de variables explicatives $\mathrm{X}_1, \ldots, \mathrm{X}_p$.

On s'intéresse à

$$\underbrace{f(\mathrm{X}_1, \ldots, \mathrm{X}_p)}_{\text{vraie moyenne inconnue}}.$$

L'analyste détermine

$$\underbrace{\widehat{f}(\mathrm{X}_1, \ldots, \mathrm{X}_p)}_{\text{approximation}},$$

une fonction des variables explicatives.

On spécifie que la **moyenne** de la variable réponse Y est une fonction linéaire des variables explicatives $\mathsf{X}_1, \ldots, \mathsf{X}_p$, soit

$$\underbrace{\mathsf{E}(Y \mid \mathbf{X})}_{\text{moyenne théorique}} = \underbrace{\beta_0 + \beta_1 \mathsf{X}_{i1} + \cdots + \beta_p \mathsf{X}_{ip}}_{\text{somme pondérée des variables explicatives}}.$$

en supposant que l'écart entre les observations et cette moyenne est constant,

$$\mathsf{Va}(Y \mid \mathbf{X}) = \sigma^2.$$

Représentation alternative

Pour la ie observation,

$$\underbrace{Y_i}_{\text{réponse}} = \underbrace{\beta_0 + \beta_1 \mathsf{X}_{i1} + \cdots + \beta_p \mathsf{X}_{ip}}_{\text{prédicteur linéaire}} + \underbrace{\varepsilon_i}_{\text{aléa}}.$$

- L'aléa ε_i représente la distance **verticale** entre la vraie pente et l'observation
- Autant d'aléas que d'observations (n), variable aléatoire inconnue...

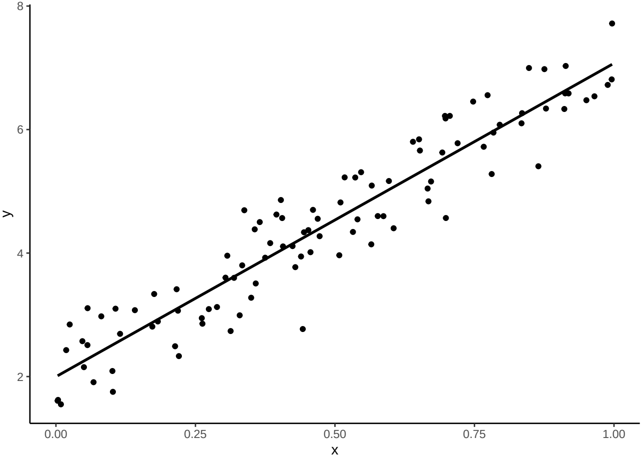
Postulats

- L'aléa ε_i représente l'erreur, soit la différence entre la valeur observée et la moyenne de la population pour les même valeurs des variables explicatives.
- On suppose que le modèle pour la moyenne est correctement spécifié: l'aléa a une moyenne théorique nulle, $\mathsf{E}(\varepsilon_i) = 0$.
- On suppose que les observations sont indépendantes les unes des autres.

Régression linéaire en deux dimensions

Si $\mathsf{E}(Y) = \beta_0 + \beta_1 \mathsf{X}$, alors

- β_0 représente l'ordonnée à l'origine (valeur quand $\mathsf{X} = 0$.)
- β_1 est la pente



Résidus ordinaires

L'estimation des paramètres $\widehat{\beta}_0, \cdots, \widehat{\beta}_p$ nous donne

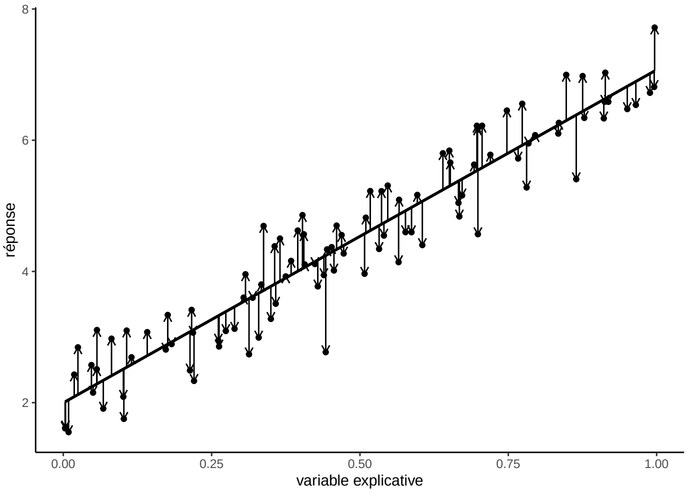
$$\underbrace{\widehat{Y}_i}_{\text{prédiction}} = \widehat{\beta}_0 + \widehat{\beta}_1 \mathsf{X}_{i1} \cdots + \widehat{\beta}_p \mathsf{X}_{ip}.$$

On peut approximer l'aléa à l'aide du **résidu ordinaires**, soit

$$\underbrace{e_i}_{\text{résidu ordinaire}} = \underbrace{Y_i}_{\text{observation}} - \underbrace{\widehat{Y}_i}_{\text{prédiction}}.$$

- par construction, la moyenne des e_i est zéro.
- le résidu ordinaire est la distance verticale entre l'observation et la "droite" **ajustée**

Illustration des résidus ordinaires



Erreur quadratique moyenne

L'erreur quadratique moyenne théorique est

$$\mathsf{E}\left[\left\{(Y - \widehat{f}(\mathsf{X}_1, \ldots, \mathsf{X}_p)\right\}^2\right],$$

la moyenne de la différence au carré entre la vraie valeur de Y et la valeur prédite par le modèle.

En pratique, on remplace la moyenne théorique par une moyenne empirique obtenue à partir d'un échantillon aléatoire.

Estimation des paramètres

Comment estimer les paramètres $\beta_0, \ldots, \beta_p$?

Optimisation: trouver les valeurs qui minimisent l'erreur quadratique moyenne **empirique** avec l'échantillon des n observations, soit

$$\frac{e_1^2 + \cdots + e_n^2}{n}$$

Solution explicite $\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top Y$!

Estimation dans **R**

La fonction `lm` calcule l'ajustement du modèle linéaire.

Arguments:

- `formula`: formule de type `reponse ~ variables explicatives`, où les variables explicatives sont séparées par un signe +
- `data`: base de données

```r
modlin <- lm(mpg ~ hp + wt,
             data = mtcars)
summary(modlin)
```

Sortie de summary

```
Call:
lm(formula = mpg ~ hp + wt, data = mtcars)

Residuals:
   Min     1Q Median     3Q    Max
-3.941 -1.600 -0.182  1.050  5.854

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
hp          -0.03177    0.00903  -3.519  0.00145 **
wt          -3.87783    0.63273  -6.129 1.12e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.593 on 29 degrees of freedom
Multiple R-squared:  0.8268,    Adjusted R-squared:  0.8148
F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12
```

Tableau de sortie

- Formule de l'appel
- Statistiques descriptives des résidus ordinaires $e_1, \ldots, e_n$.
- Tableau des estimations
  - Coefficients $\widehat{\beta}_j$
  - Erreurs-types, $\mathsf{se}(\widehat{\beta}_j)$
  - Statistique du test-*t* pour $\mathscr{H}_0 : \beta_j = 0$, soit $t = \widehat{\beta}_j/\mathsf{se}(\widehat{\beta}_j)$
  - Valeur-*p* selon loi nulle $\mathsf{St}(n-p-1)$
- Estimation de l'écart-type $\widehat{\sigma}$ et degrés de liberté $n-p-1$
- Estimations du coefficient de détermination, R^2 et R^2 ajusté
- Statistique F d'ajustement global et valeur-p de $\mathsf{F}(p, n-p-1)$
  - $\mathscr{H}_a$: modèle linéaire
  - $\mathscr{H}_0$: modèle avec uniquement ordonnée à l'origine (chaque observation prédite par la moyenne des réponses, $\overline{Y}$)

Quelques méthodes pour `lm`

- `resid` pour les résidus ordinaires e_i
- `fitted` pour les valeurs ajustées $\widehat{Y}_i$
- `coef` pour les estimations des paramètres $\widehat{\beta}_0, \ldots, \widehat{\beta}_p$
- `plot` pour des diagnostics graphiques d'ajustement
- `anova` pour la comparaison de modèles emboîtés
- `predict` pour les prédictions (avec nouvelles données)
- `confint` pour intervalles de confiance pour les paramètres.

Variables catégorielles

- Les facteurs (`<factor>`) sont traités adéquatement par **R**.
- Si la variable a K valeurs possibles (niveaux), le modèle inclut $K-1$ indicatrices 0/1.
- Par défaut dans **$**, la catégorie de référence est la plus petite en ordre alphanumérique.

Encodage des variables catégorielles

Considérons une variable catégorielle `cat` avec niveaux 1, 2, et 3.

`cat`	`cat2`	`cat3`
1	0	0
2	1	0
3	0	1

La catégorie de référence est associée à l'ordonnée à l'origine (quand `cat2=0` et `cat3=0`).

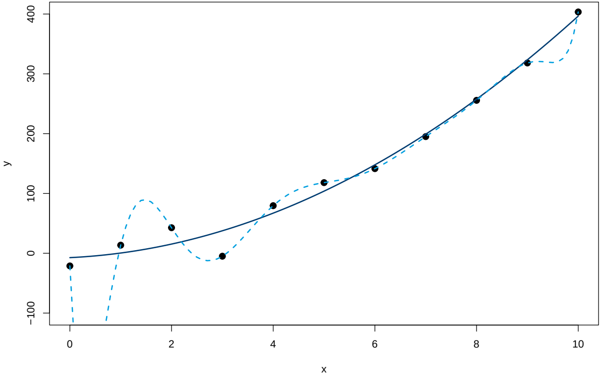
Prédiction vs inférence

Sélection de variables

- Comment choisir quelles variables inclure?
- Quel est la spécification adéquate pour $f(\mathsf{X}_1, \ldots, \mathsf{X}_p)$?
  - régression, réseaux de neurone, forêts aléatoires, etc.
  - transformations de variables, $\texttt{age}^2$, $\ln(\texttt{age})$, etc.

Notre but sera de sélectionner un **bon** modèle, selon les objectifs de l'étude

Illustration du surajustement



Séparation des données

Ne pas utiliser les données employés pour ajuster un modèle pour **prédire la performance**

- échantillons d'apprentissage/validation/test (fixes)
- validation croisée (avec $K = 5, 10$ groupes), mais *résultat aléatoire*