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 $\mathsf{X}_1, \dots, \mathsf{X}_p$.

On s'intéresse à

$$\underbrace{f(\mathsf{X}_1, \dots, \mathsf{X}_p)}_{\text{vraie moyenne inconnue}} \quad .$$

L'analyste détermine

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

une fonction des variables explicatives.

Rappels sur la régression linéaire

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

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

- 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...

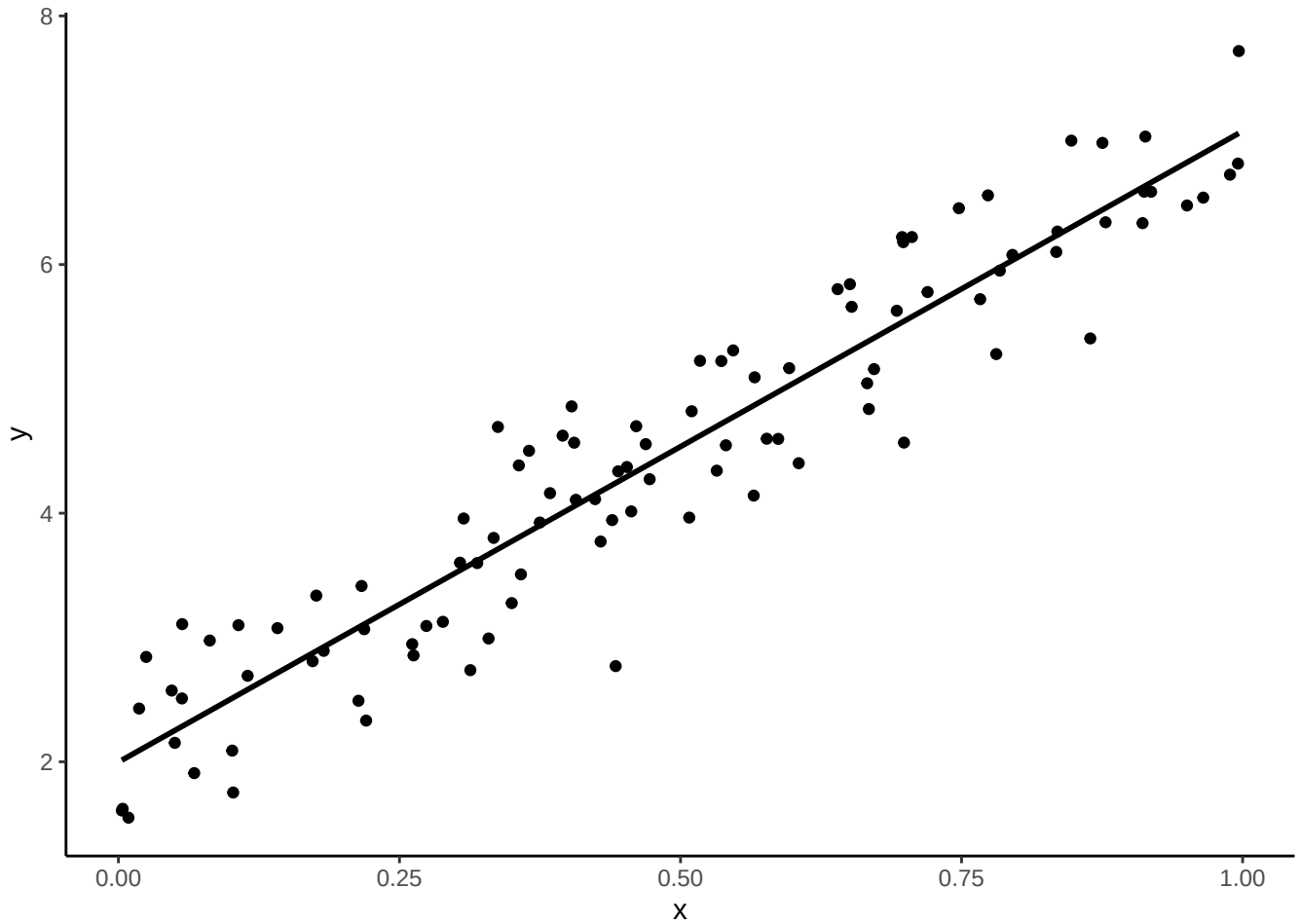
Postulats

- 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 $\mathrm{E}(Y) = \beta_0 + \beta_1 \mathrm{X}$, alors

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



Résidus ordinaires

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

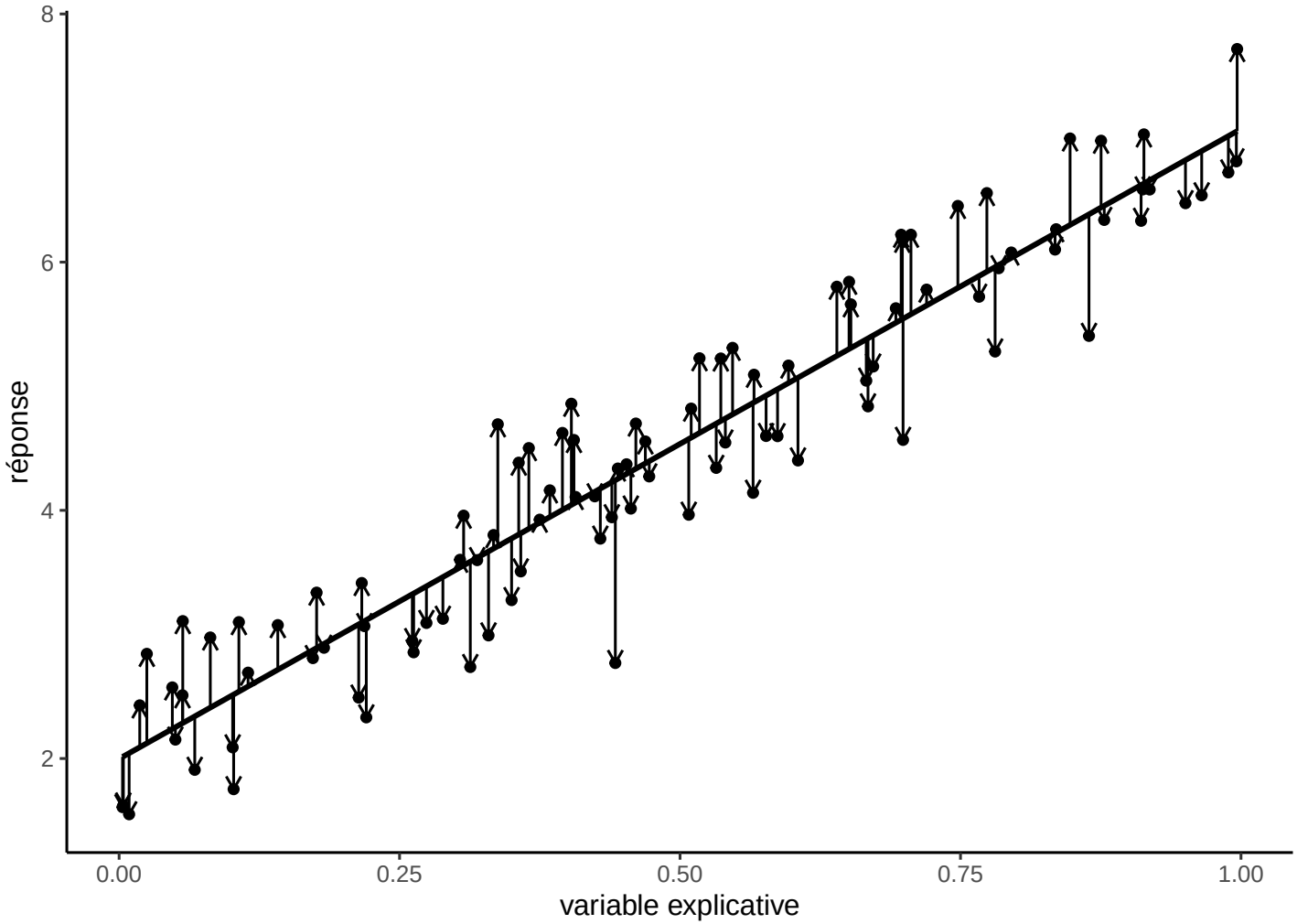
$$\underset{\text{prédiction}}{\hat{Y}_i} = \hat{\beta}_0 + \hat{\beta}_1 \mathsf{x}_{i1} \cdots + \hat{\beta}_p \mathsf{x}_{ip}.$$

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

$$\underset{\text{résidu ordinaire}}{e_i} = \underset{\text{observation}}{Y_i} - \underset{\text{prédiction}}{\hat{Y}_i}.$$

- 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

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

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 $\beta_0, \dots, \beta_p$?

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

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

Il existe une solution explicite au problème d'optimisation!

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

```
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 $e_1, \dots, e_n$.
- Tableau des estimations
  - Coefficients $\hat{\beta}_j$
  - Erreurs-types, se$(\hat{\beta}_j)$
  - Statistique du test-t pour $\mathscr{H}_0 : \beta_j = 0$, soit $t = \hat{\beta}_j / \text{se}(\hat{\beta}_j)$
  - Valeur-p selon loi nulle St$(n - p - 1)$
- Estimation de l'écart-type $\hat{\sigma}$ 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 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 $\hat{Y}_i$
- `coef` pour les estimations des paramètres $\hat{\beta}_0, \dots, \hat{\beta}_p$
- `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 **R**, 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).

Sélection de variables et de modèles

- Comment choisir quelles variables inclure?
- Quel est la spécification adéquate pour $f(\mathsf{X}_1, \dots, \mathsf{X}_p)$?
  - 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

Prédiction vs inférence

Prédiction: obtenir une estimation de $\hat{Y}$: on veut un modèle performant

Inférence: estimer l'effet de variables explicatives, effectuer des tests d'hypothèse

Pour l'inférence, il est préférable de spécifier le modèle dès le départ (devis expérimental) selon des considérations scientifiques et de s'y tenir.

Spécification adéquate d'un modèle

Démonstration **R**

Omettre des termes importants mène à un **modèle biaisé**.

Ajouter des termes superflus augmente la **variabilité**.

Évaluer la performance d'un modèle

On peut calculer l'erreur quadratique moyenne (EQM) sur l'ensemble des données qui servent à ajuster le modèle.

Q: Est-ce que c'est un marqueur fiable de la performance du modèle?

Exemple avec la régression polynomiale

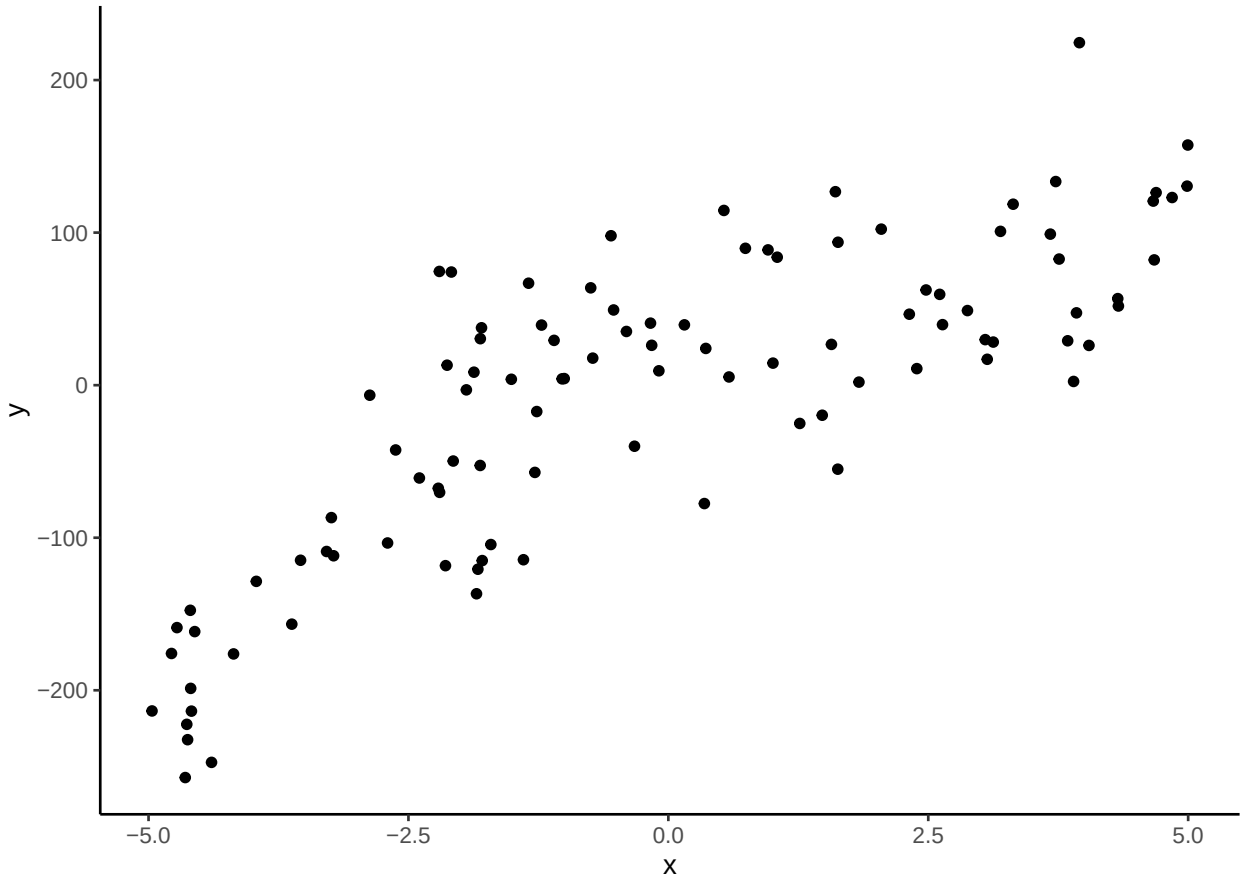
On a simulé des observations d'un modèle polynomiale d'ordre K, de la forme

$$\mathsf{E}(Y \mid \mathbf{X}) = \beta_0 + \beta_1 \mathsf{X} + \cdots + \beta_K \mathsf{X}^K.$$

On vise à estimer l'ordre du polynôme en effectuant des régressions linéaires.

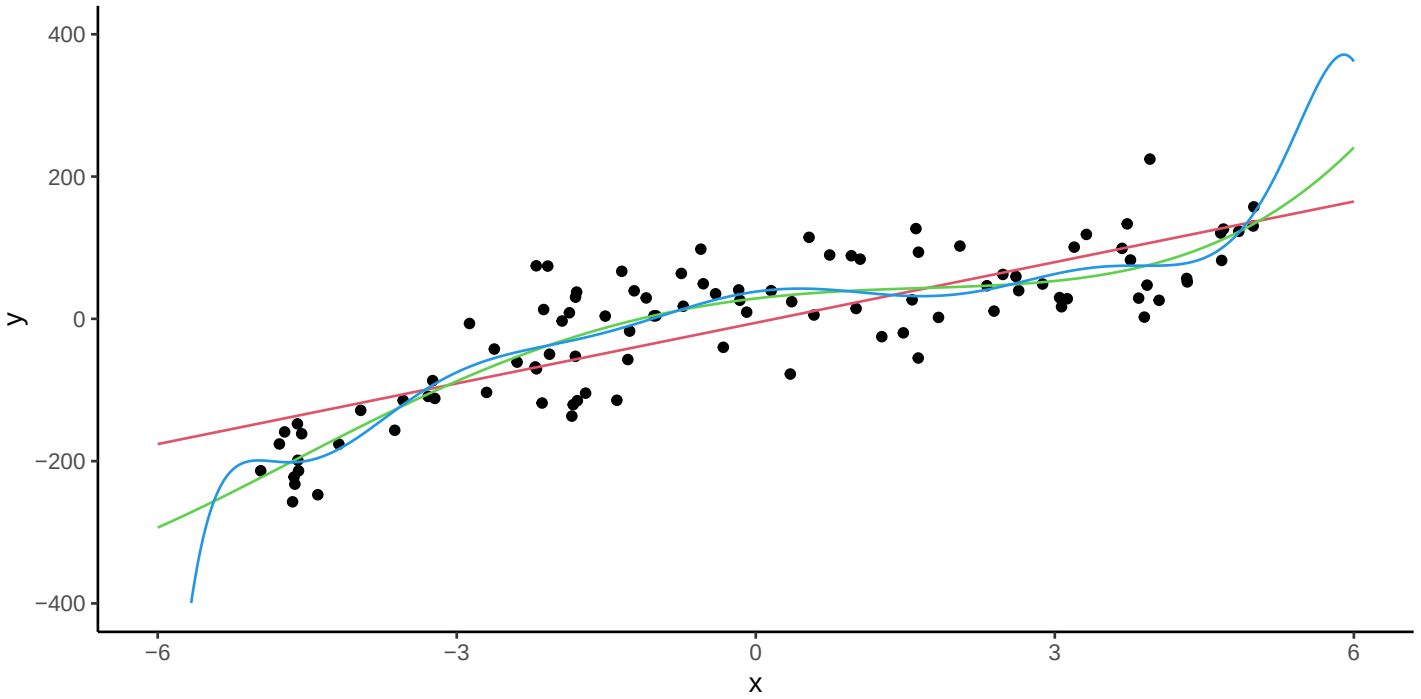
```
data(polynome, package = "hecmulti")
k <- 4L # degré du polynôme
lm(y ~ poly(x, degree = k),
   data = polynome)
```

Aperçu des données



D'après vous, quelle est la vraie valeur de K?

Ajustement de polynômes



Ajustement pour des polynômes de degré $K = 1, 4, 10$.

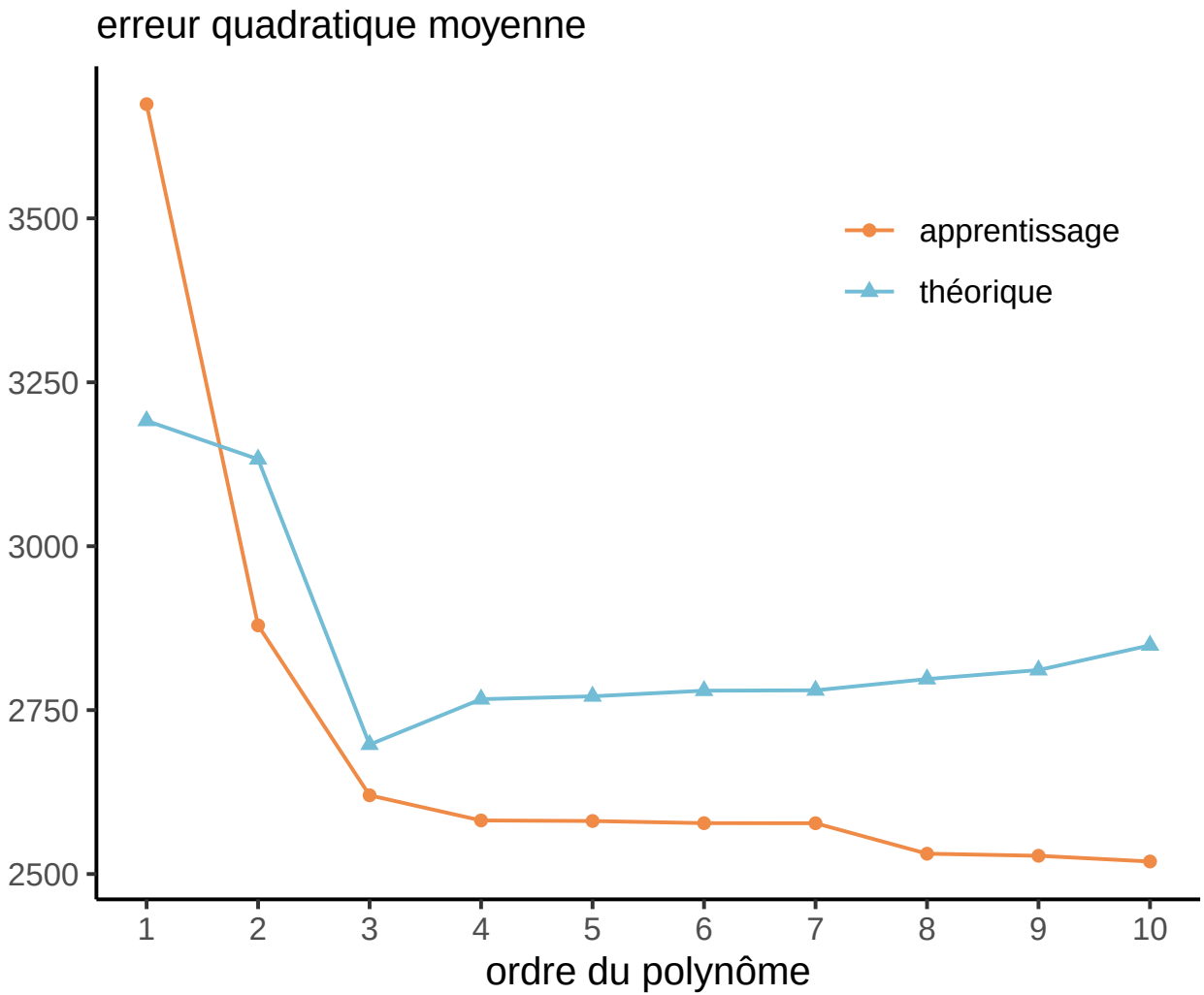
Évaluation de la performance

Pour $K = 1, \ldots, 10$:

- Ajuster le modèle linéaire
- Obtenir les résidus ordinaires
- Calculer l'erreur quadratique moyenne

```
data(polynome, package = "hecmulti")
eqm <- vector(length = 10L, mode = "numeric")
for(K in seq_len(10)){
 mod <- lm(y ~ poly(x, degree = K), data = polynome)
 eqm[K] <- mean(resid(mod)^2)
}
```

Estimation de l'erreur quadratique moyenne



Plus le modèle est complexe, plus l'erreur quadratique moyenne de l'échantillon d'apprentissage est petite!

Suroptimisme

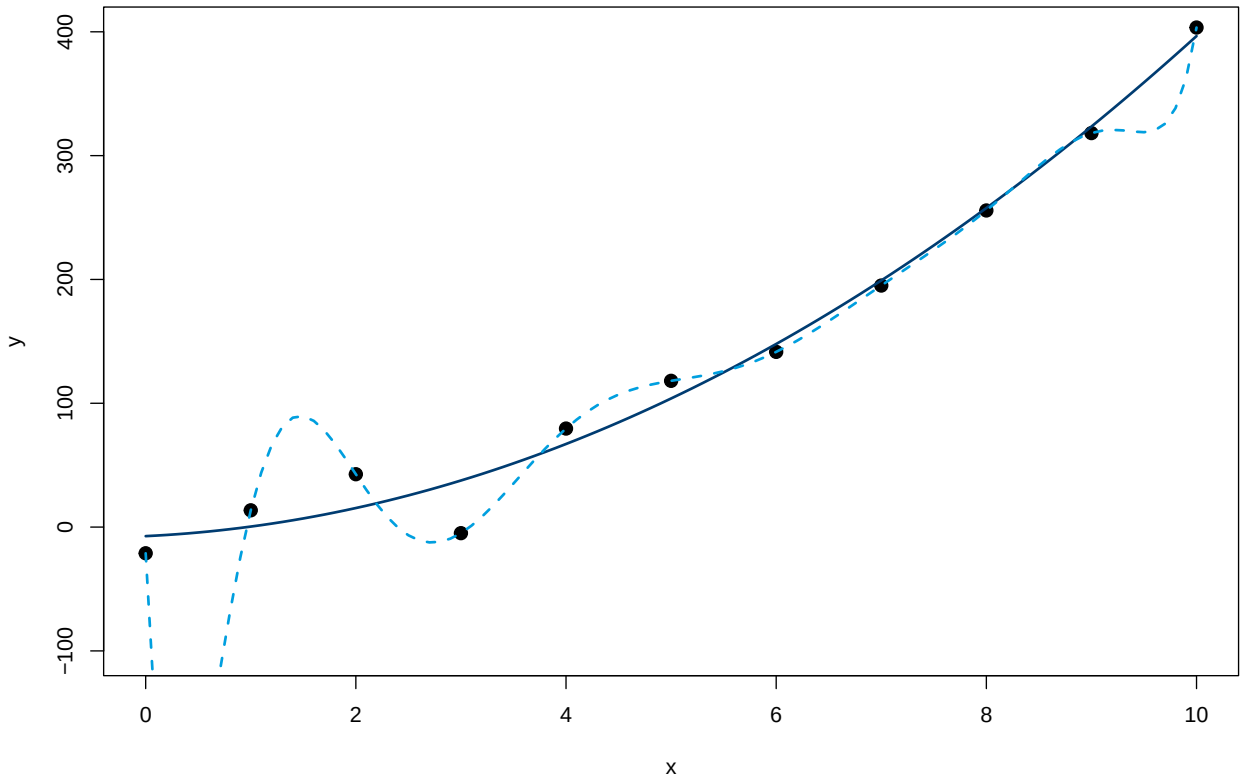
L'erreur quadratique moyenne décroît mécaniquement à chaque fois qu'on ajoute une variable au modèle de régression.

C'est pourquoi on ne peut pas l'utiliser comme outil de sélection de variables, autrement on va surestimer la performance (**suroptimisme**).

Surajustement

Un modèle plus complexe va toujours mieux s'ajuster aux données de l'échantillon.

En revanche, il se **généralise** moins bien (notre objectif en prédiction).



Morale de l'histoire

Nous nous sommes rendus coupables d'un péché capital en utilisant deux fois les données

- une fois pour l'ajustement,
- une fois pour la validation.

C'est comme tremper deux fois un biscuit dans le verre de lait...

Estimation fiable de la performance

- Pénalisation de la complexité du modèle
  - critères d'information
  - pénalité sur les coefficients (*ridge*, LASSO)
- Validation et entraînement avec des données différentes
  - validation externe
  - validation croisée

Pénalisation

Utiliser toutes les données (échantillon de validation), mais ajouter une pénalité.

Si le modèle est ajusté avec la méthode du maximum de vraisemblance, alors on a accès aux critères d'information.

Lien entre régression linéaire et vraisemblance

Si on suppose que les aléas sont indépendants et suivent une loi normale de variance σ^2 constante, alors la log-vraisemblance s'écrit

$$\begin{aligned}\ell(Y; \mathbf{X}, \beta, \sigma) = &-\frac{n}{2}\ln(2\pi) - n\ln(\sigma) \\ &- \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \cdots - \mathrm{X}_{ip}\beta_p)^2.\end{aligned}$$

Ainsi, l'estimateur des moindres carrés, $\widehat{\beta}$, correspond au maximum de vraisemblance des coefficients.

Critères d'information

Les critères d'information sont de la forme

$$\begin{aligned} \text{IC} &= -2\ell(\hat{\beta}, \hat{\sigma}^2) + \text{pénalité} \times \text{nb param} \\ &= n \ln(\widehat{\text{EQM}}) + \text{pénalité} \cdot \text{nb param} + \text{constante}, \end{aligned}$$

Le critère d'Akaike (AIC) utilise une pénalité de 2, le critère bayésien (BIC) de $\ln(n)$.

Utiliser les critères d'information

- Plus la valeur du critère d'information, meilleure est l'adéquation (et donc la performance).
- La pénalité assure que les modèles plus simples ne sont pas systématiquement retournés.
- Le BIC retourne toujours des modèles plus parcimonieux (simples) que le AIC.
- Génériques `logLik`, `AIC` et `BIC` en **R**

Illustration

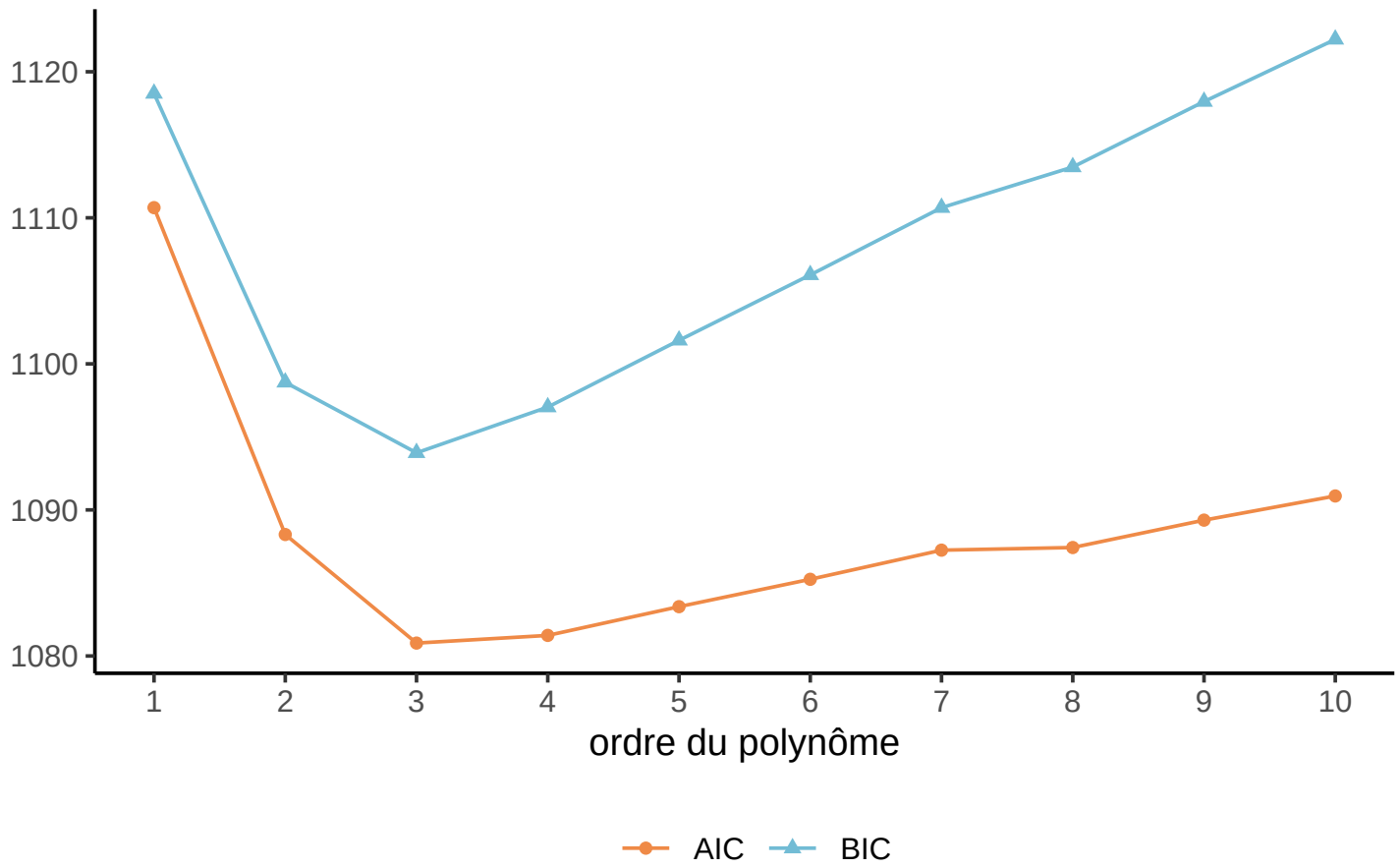


Figure 1: Critères d'information en fonction de l'ordre du polynôme.

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

Validation externe

Séparer l'ensemble d'observations en plusieurs groupes

- l'échantillon d'apprentissage pour l'ajustement du modèle
- l'échantillon de validation pour l'estimation de la performance
- l'échantillon test pour l'inférence (optionnel)

En pratique, il faut beaucoup d'observations!

Avantages et inconvénients de la validation externe

Applicable en toute généralité peu importe le type de modèle.

Cette approche, qui compartimente les échantillons, n'est pas sans faille.

- On obtient un résultat différent selon la division
- Gaspillage potentiel
- Choix d'ordinaire aléatoire, mais choix particuliers de fenêtre selon données (par ex., séries chronologiques)

Validation croisée

Une autre méthode de rééchantillonage

Diviser l'échantillon en K groupes d'observations de taille moyenne égale

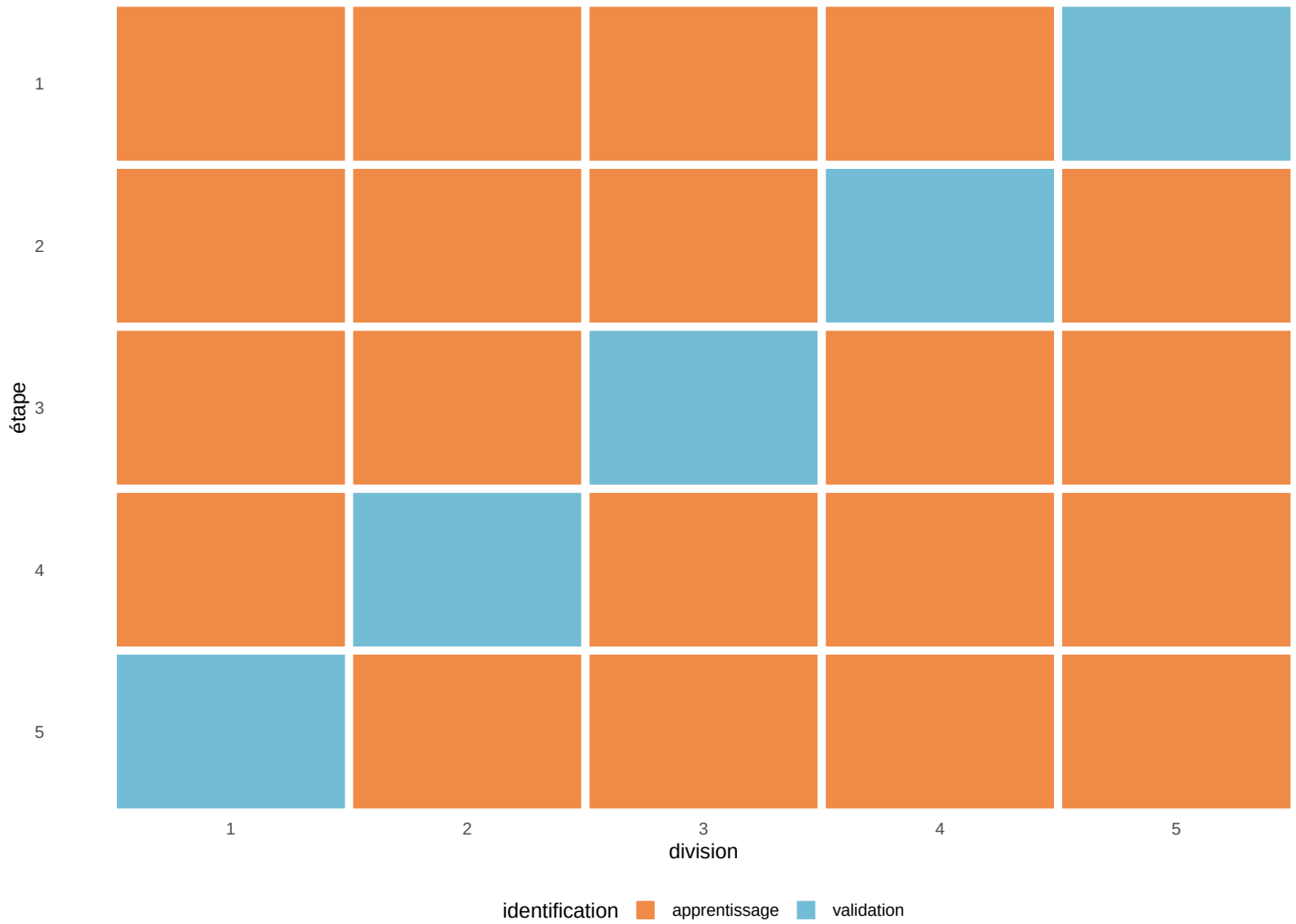
- utiliser $K-1$ groupes pour l'ajustement
- estimer la performance avec les données du dernier groupe

Les regroupements sont aléatoires, tout comme la mesure finale de performance!

Étapes de la validation croisée à K groupes

1. Diviser l'échantillon au hasard en K parties $P_1, P_2, \ldots, P_K$ contenant toutes à peu près le même nombre d'observations.
2. Pour $j = 1$ à K,
   i. Enlever la partie j.
   ii. Estimer les paramètres du modèle en utilisant les observations des $K - 1$ autres parties combinées.
   iii. Calculer la mesure de performance (par exemple la somme du carré des erreurs $\{Y_i - \hat{Y}_i\}^2$) de ce modèle pour le groupe P_j.
3. Combiner les K estimations de performance pour obtenir une mesure de performance finale.

Illustration de la validation croisée



Combien de groupes K pour la validation croisée?

La validation croisée est plus coûteuse parce qu'on doit ajuster K fois le modèles

- Pour la régression linéaire, choisir $K = n$ (*leave-one-out cross-validation*) permet d'éviter de réajuster le modèle grâce à un artifice de calcul.
- On recommande habituellement de prendre $K = \min\{n^{1/2}, 10\}$ groupes.
- les choix de $K = 5$ ou $K = 10$ sont ceux qui revient le plus souvent en pratique.

Voir Davison & Hinkley (1997), *Bootstrap methods and their applications*, Section 6.4 pour une discussion plus étoffée et l'algorithme 6.5 pour une meilleure implémentation.

Validation croisée = résultat aléatoire!

Soit $\widehat{\text{EQM}}_{\text{VC}}$ l'estimation de l'erreur quadratique moyenne obtenue par validation croisée à K plis pour un modèle $\mathcal{M}$ donné.

Si on a un nombre similaire d'observations dans chaque groupe, on peut plutôt

- calculer l'erreur quadratique moyenne de chaque pli, disons $\widehat{\text{EQM}}_{\text{VC},k}$
- si K est suffisamment grand (disons $K > 10$), on peut estimer l'écart-type empirique de cette moyenne via

$$\text{sd}(\widehat{\text{EQM}}_{\text{VC}}) = \frac{1}{K-1}\sum_{k=1}^{K}(\widehat{\text{EQM}}_{\text{VC},k} - \widehat{\text{EQM}}_{\text{VC}})^2$$

Règle d'un écart-type

Suivant Breiman, on choisit le modèle le plus simple parmi un ensemble $\mathcal{M}_0 \subset \cdots \subset \mathcal{M}_m$ qui satisfasse

$$\widehat{\mathsf{EQM}}_{\mathsf{VC}}(\mathcal{M}_i) \leq \min_{m=i+1}^{M} \widehat{\mathsf{EQM}}_{\mathsf{VC}}(\mathcal{M}_m) + \mathsf{sd}\{\widehat{\mathsf{EQM}}_{\mathsf{VC}}(\mathcal{M}_m)\}$$

Choisir le modèle le plus simple qui soit à au plus un écart-type de la performance des modèles plus compliqués.

Validation croisée en **R**

Le paquet `caret` a une fonction pour faire la validation croisée.

```
cv_caret <-
  caret::train(form = formula(y ~ poly(x, degree = 3)),
               data = polynome,
               method = "lm",
               trControl = caret::trainControl(
                 method = "cv",
                 number = 10)) #nb plis
reqm_cv <- cv_caret$results$RMSE # racine EQM
reqm_sd_cv <- cv_caret$results$RMSESD
```

Aussi `boot::cv.glm()` qui inclut une correction de biais pour les modèles linéaires généralisé.

Résultats de la validation croisée

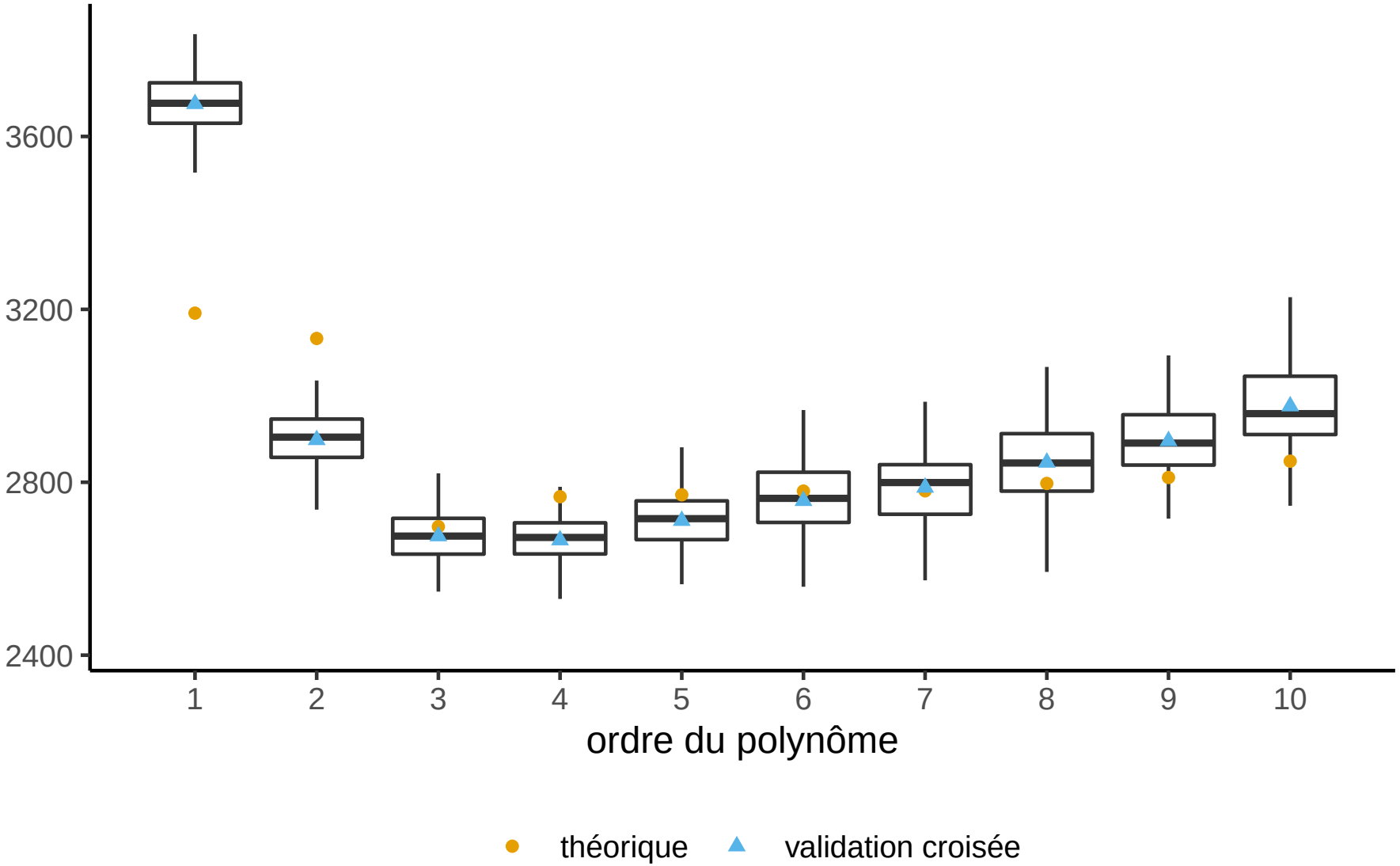


Figure 2: Boîtes à moustaches des 100 estimations de l'erreur quadratique moyenne obtenues par validation croisée à 10 plis.

Récapitulatif

- Le choix de la complexité d'un modèle est un compromis entre
  - le biais (modèle trop simple, mal spécifié) et
  - la variance (même nombre d'observations/budget, plus de paramètres à estimer, estimations moins fiables)
- L'erreur quadratique moyenne est la mesure usuelle de la performance d'un modèle linéaire

Récapitulatif

Si on estime la performance avec les mêmes données qui ont servi à l'ajustement, on surestime la performance

- l'erreur quadratique moyenne calculée sur les mêmes données qui ont servi à l'entraînement est biaisée
- cela mène à du **surajustement**

Récapitulatif

Trois méthodes pour estimer de manière plus objective la performance d'un modèle

- Critères d'information (pénalisation)
- Validation externe
- Validation croisée

Vous devez être en mesure de nommer les forces et faiblesses et d'expliquer le fonctionnement (avec du pseudocode).