Apprendimento statistico e machine learning

Selezione del modello lineare

Leonardo Egidi

Ottobre 2024

Università di Trieste

Selezione modello lineare

Modello lineare

- Ricordiamo il modello lineare:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p Xp + \epsilon.$$

- Nelle lezioni che seguono, considereremo alcuni approcci per estendere la struttura del modello lineare. Considereremo per l'appunto effetti *non lineari* ma ancora *additivi*.
- Più avanti considereremo invece veri e propri modelli generali *non lineari*, come ad esempio la *regressione polinomiale* e le *splines*.

Lode al modello lineare!

- Nonostante la sua apparente semplicità, il modello lineare ha grandi vantaggi dal punto di vista dell' *interpretabilità* e spesso produce buone *performance predittive*.
- Discutiamo qui alcuni modi in cui il modello lineare può essere migliorato, sostituendo il metodo di stima dei minimi quadrati ordinari con altre procedure di stima.

Perché considerare alternative ai metodi quadrati?

- *Accuratezza predittiva*: specialmente quando $p > n$, per controllare la varianza.
- *Interpretabilità del modello*: rimuovendo covariate/fattori non rilevanti - ovverossia, ponendo le stime dei corrispondenti coefficienti uguali a zero - possiamo ottenere un modello più facile da interpretare. Presenteremo quindi degli approcci per fare automaticamente una *selezione delle covariate*.
- **Glossario**: in questo corso useremo interscambiabilmente i seguenti vocaboli: predittori, covariate, variabili esplicative, variabili indipendenti, *features*,...

Tre classi di metodi

- *Selezione subset*: identifichiamo un sottoinsieme tra i p predittori/covariate che crediamo siano associate alla variabile risposta. Usiamo poi il metodo di stima dei minimi quadrati ordinari su questo sottoinsieme di predittori.
- *Regolarizzazione (Shrinkage)*: stimiamo un modello con tutti i p predittori/covariate, ma le stime dei coefficienti vengono 'tirate' (shrunken) verso zero rispetto alla stima dei minimi quadrati ordinari. Questo metodo di *shrinkage* ha l'effetto di ridurre la varianza e simultaneamente performare una selezione delle covariate rilevanti.
- *Riduzione delle dimensioni (Dimension Reduction)*: proiettiamo i p predittori/covariate su un sottospazio di dimensione M, dove $M < p$. Questo si ottiene calcolando M combinazioni lineari, o *proiezioni*, delle variabili. Dopodiché queste M proiezioni vengono usate come predittori per stimare una regressione lineare con i minimi quadrati ordinari.

Subset selection

Selezione di modelli *best subset*

Best subset selection

1. Indichiamo con $\mathcal{M}_0$ il modello nullo, che non ha predittori, ma solo l'intercetta. Questo modello prevede/assegna semplicemente la media campionaria ad ogni osservazione.
2. Per $k = 1, 2, \ldots, p$:

   - stimare tutti i $\binom{p}{k}$ modelli che contengono esattamente k predittori;
   - selezionare il migliore tra questi $\binom{p}{k}$ modelli, $\mathcal{M}_k$. Qui per migliore si intende il modello avente la più piccola somma dei residui al quadrato (RSS), o equivalentemente il più alto R^2.

3. Selezionare un unico modello migliore dalla collezione $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ usando errori di previsione cross-validati, C_p, AIC, BIC o l'R^2 aggiustato.

Esempio - Dati sui crediti

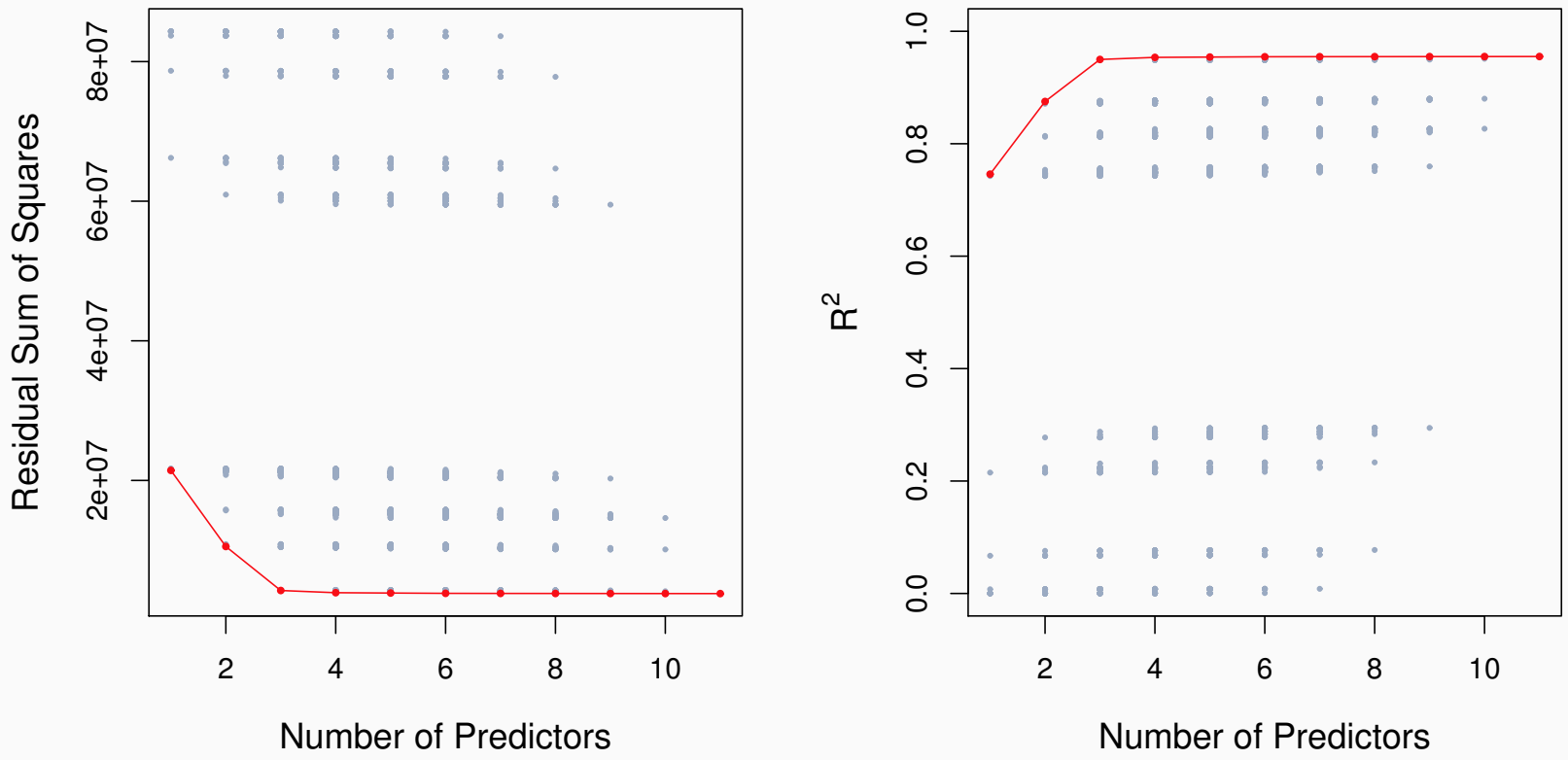

Figure 1: RSS e R^2 per ogni possibile modello con un sottoinsieme dei dieci predittori nel dataset `Credit`. La spezzata rossa segnala il miglior modello per un numero dato di predittori. Il dataset contiene solamente dieci predittori, tuttavia l'asse x va da 1 a 11 siccome una delle variabili è categorica e assume tre possibili valori, il che porta alla creazione di due variabili dummy.

Commenti sulla figura precedente

- L'adattamento ai dati migliora al crescere del numero di predittori, come atteso (RSS e R^2 si comportano monotonicamente con il numero di predittori inclusi). *Provatelo!*
- Non possiamo limitarci a scegliere un modello sulla base dello step 2. dell'algoritmo nella slide 6, che valuta il *training error*, ma dobbiamo poi cercare il modello che minimizza il *test error* tramite il punto 3.
- Ogni punto nella figura precedente indica una stima ai minimi quadrati ordinari usando un diverso sottoinsieme degli 11 predittori del dataset.

Estensioni ad altri modelli

- Nonostante la selezione di tipo subset sia stata presentata per modelli lineari con stima dei minimi quadrati, questo tipo di selezione si applica anche ad altre classi di modelli, quali ad esempio la regressione logistica.
- La *devianza* - due volte la log-verosimiglianza massimizzata - gioca il ruolo di RSS per un'ampia classe di modelli, quali la regressione logistica.

Selezione Stepwise

- Per ragioni computazionali, la selezione di un sottoinsieme ottimale (best subset) non si può applicare per un p molto alto.
- Difatti, più grande è p, più largo è lo spazio della ricerca, e maggiore è la probabilità di reperire modelli che si adattano bene ai dati di *training*, sebbene questi possano non avere alcuna efficacia predittiva per dati futuri.
- Uno spazio di ricerca troppo ampio può portare a *overfitting* e a un'alta varianza delle stime dei coefficienti.
- Per entrambe queste ragioni, i cosiddetti metodi *stepwise*, che esplorano un insieme di modelli decisamente più ristretto, rappresentano una valida alternativa ai metodi di selezioni di tipo *best subset*.

Selezione stepwise in avanti (forward stepwise)

- La selezione di tipo *stepwise* (graduale) in avanti inizia con un modello senza predittori, e aggiunge mano a mano un predittore alla volta, fino a quando tutti i predittori sono inclusi nel modello.
- In particolare, ad ogni passo la variabile che apporta il più cospicuo miglioramento alla stima rispetto alle altre è aggiunta automaticamente al modello.

Forward stepwise selection

Selezione di modelli *stepwise*

Forward Stepwise selection

1. Indichiamo con $\mathcal{M}_0$ il modello nullo, che non ha predittori, ma solo l'intercetta.
2. Per $k = 0, 1, , \ldots, p-1$:

   - considerare tutti i $p-k$ modelli che aumentano i predittori in $\mathcal{M}_k$ con un predittore aggiuntivo;
   - selezionare il migliore tra questi $p-k$ modelli, e chiamiamolo $\mathcal{M}_{k+1}$. Qui per migliore si intende il modello avente la più piccola somma dei residui (RSS), o equivalentemente il più alto R^2.

3. Selezionare un unico modello migliore dalla collezione $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ usando errori di previsione cross-validati, C_p, AIC, BIC o l'R^2 aggiustato.

Di più sulla selezione stepwise in avanti

- Il vantaggio computazionale rispetto ai metodi best subset è ovvio.
- Non è garantita la reperibilità del migliore tra tutti i 2^p modelli possibili contenenti sottoinsiemi dei p predittori. *Perché no? Diamo un esempio*.

Dati sui crediti

# variables	best subset	forward stepwise
one	rating	rating
two	rating, income	rating, income
three	rating, income, student	rating, income, student
four	cards, income, student, limit	rating, income, student, limit

Nella tabella qui sopra, i primi 4 modelli migliori secondo i metodi best subset e stepwise in avanti sul dataset `Credit`. I primi tre modelli coincidono, ma il quarto differisce nei due metodi.

Stepwise selection all'indietro (backward)

- Come nel caso della selezione in avanti, la *selezione stepwise all'indietro* fornisce una valida alternativa (computazionalmente meno onerosa) rispetto ai metodi best subset.
- Tuttavia, diversamente da quanto avviene nella selezione in avanti, la selezione all'indietro inizia con il modello contenente tutti e p i predittori, e poi iterativamente rimuove il meno utile dei predittori, uno alla volta.

Backward stepwise selection

Selezione di modelli *stepwise*

Backward Stepwise selection

1. Indichiamo con $\mathcal{M}_p$ il modello pieno, con tutti e p i predittori.
2. Per $k = p, p-1, \ldots, 1$:

- considerare tutti i k modelli che contengono tutti tranne uno i predittori in $\mathcal{M}_k$, per un totale di $k-1$ predittori;
- scegliere il migliore tra questi k modelli, chiamiamolo $\mathcal{M}_{k-1}$, dove per migliore si intende il modello avente la più piccola somma dei residui (RSS), o equivalentemente il più alto R^2.

3. Selezionare un unico modello migliore dalla collezione $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ usando errori di previsione cross-validati, C_p, AIC, BIC o l'R^2 aggiustato.

Di più sulla selezione stepwise all'indietro

- Come per i metodi in avanti, i metodi di ricerca all'indietro ricercano tra soli $1 + p(p+1)/2$ modelli, e possono quindi essere applicati solo in quei casi in cui p è troppo grande per implementare la ricerca di tipo best subset.
- Come per i metodi in avanti, i metodi di ricerca di tipo backward non garantiscono per forza il reperimento del *miglior* modello contenente p predittori.
- I metodi stepwise di tipo backward richiedono che *il numero di unità statistiche n sia molto più grande del numero di predittori p* (in modo tale che il modello pieno possa venir stimato). Viceversa, i metodi stepwise in avanti possono essere utilizzati anche se $n < p$, e per questa ragione i metodi forward stepwise sono gli unici metodi idonei quando p è grande.

Scegliere il modello ottimale

- Il modello contenente tutti i predittori avrà sempre la più piccola RSS e il più grande R^2, siccome queste quantità sono intrinsecamente legate all'errore di previsione sul training set.
- Vorremmo poter selezionare un modello con un piccolo errore di previsione sul test set, non un modello con un basso errore sul training set. Ricordiamo infatti che l'errore sul training set è solitamente una stima poco efficace dell'errore sul test set.
- Inoltre, RSS e R^2 non sono adatti per scegliere il migliore tra una serie di modelli con diverso numero di predittori ciascuno.

Stimare il test error: due possibili approcci

- Possiamo stimare il test error *indirettamente*, apportando un aggiustamento dell'errore di training per controllare/abbassare il bias dovuto all'overfitting.
- Possiamo *direttamente* stimare il test error, usando o un approccio su un set di validazione, oppure un metodo di cross-validazione.
- Illustreremo entrambi i metodi.
- Quando parliamo di *training error* intendiamo una misura del tipo:

$$\text{err}_{train} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\text{RSS},$$

basata per l'appunto su una qualche funzione dei residui di regressione (in questo caso la somma dei quadrati).

C_p, AIC, BIC e R^2 aggiustato.

- Queste misure *aggiustano* il training error tramite indicatori di ampiezza del modello, e possono essere utilizzate per selezionare tra una serie di modelli con diverso numero di predittori.
- La figura seguente illustra C_p, BIC e l' R^2 aggiustato per il miglior modello di ciascuna dimensione prodotto da una selezione di tipo *best subset* sul dataset `Credit`.

Dati sui crediti

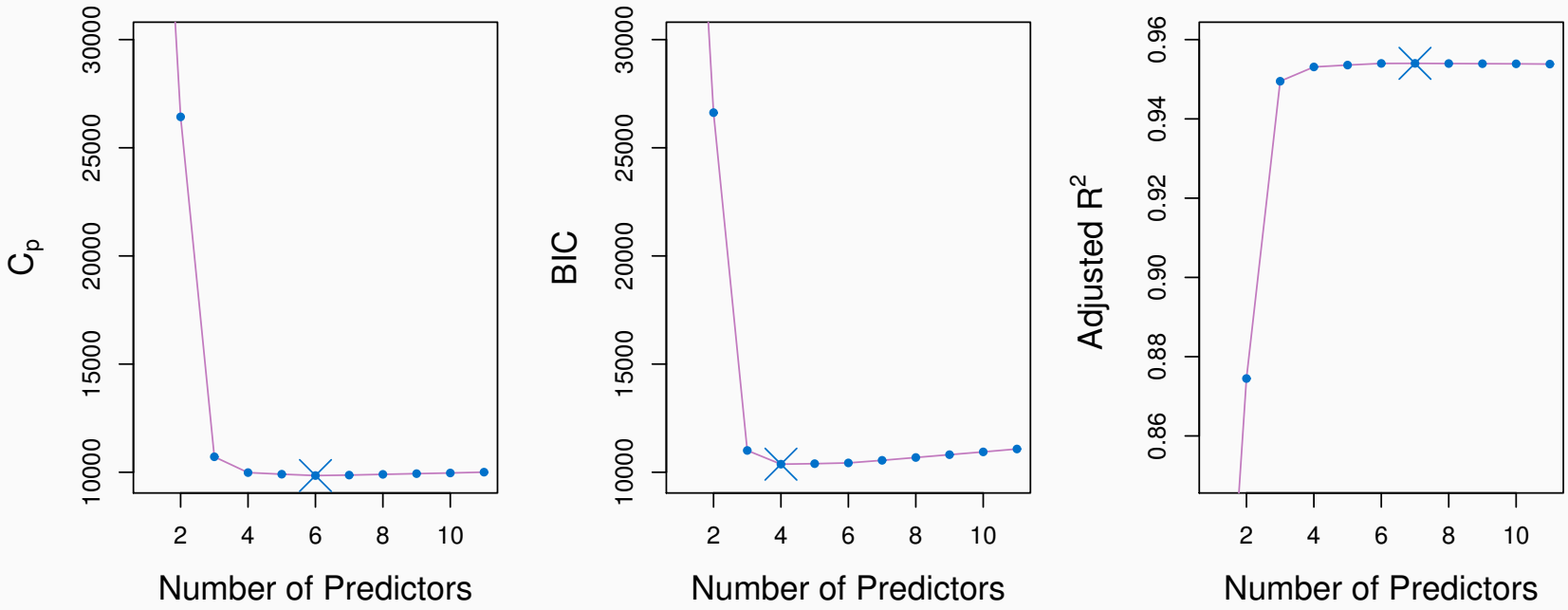

Figure 2: Come si evince, metodi diversi portano a selezione di modelli diversi

Dettagli su queste misure

- *Mallow's C_p*:

$$C_p = \frac{1}{n}(\text{RSS} + 2d\hat{\sigma}^2),$$

dove d indica il numero totale di parametri usati, mentre $\hat{\sigma}^2$ indica una stima della varianza dell'errore ϵ associata ad ogni misurazione della variabile risposta.

- Criterio AIC: definito per un'ampia classi di modelli stimati con massima verosimiglianza:

$$\text{AIC} = -2\log L + 2d,$$

dove L denota la verosimiglianza massimizzata per il modello stimato.

- Nei modelli lineari con errori gaussiani, minimi quadrati e stima di verosimiglianza si equivalgono, così come AIC e C_p. *Provatelo come esercizio!*

Dettagli sul BIC

- BIC: criterio utilizzabile per ampia classe di modelli:

$$\text{BIC} = -2 \log L + 2 \log n = \frac{1}{n}(\text{RSS} + \log(n) d \hat{\sigma}^2).$$

- Come per C_p, il BIC tenderà ad assumere valori piccoli per modelli con basso errore di previsione sul test set, per questo motivo generalmente selezioniamo il modello con il BIC più basso.
- Notiamo che il BIC sostituisce la *penalizzazione* $2d\hat{\sigma}^2$ del C_p con $\log(n)d\hat{\sigma}^2$, dove n è il numero totale di osservazioni (unità statistiche).
- Siccome $\log(n) > 2$ per $n > 7$, il BIC generalmente penalizza di più modelli molto parametrizzati, e risulta quindi un criterio che seleziona modelli più *parsimoniosi* di quanto non faccia il C_p (dare un'occhiata alla slide 21 per avere conferma).

R^2 aggiustato

- Per un modello stimato con i minimi quadrati e avente d predittori, l'R^2 aggiustato è calcolato come:

$$\text{adj}R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)},$$

  ove TSS indica la somma totale dei residui al quadrato. (Ricorda che TSS $=$ RSS $+$ SSR, quindi $R^2 = 1 - RSS/TSS$).
- A differenza di C_p, AIC e BIC per i quali un valore piccolo indica un basso errore di previsione, nel caso di adjR^2 un valore alto indica un basso errore di previsione.
- Massimizzare adjR^2 equivale a minimizzare $\text{RSS}/(n-d-1)$. Mentre RSS decresce sempre all'aumentare del numero di predittori, la quantità $\text{RSS}/(n-d-1)$ invece può crescere o decrescere, data la presenza di d al denominatore.
- A differenza di R^2, adjR^2 *paga un prezzo* per l'inclusione di predittori non necessari nel modello (guardare slide 21).

Validazione e cross-validazione

- Ognuna di queste procedure restituisce una sequenza di modelli $\mathcal{M}_k$ indicizzati dalle dimensioni $k = 0, 1, 2, \ldots$. Il nostro scopo qui è selezionare $\hat{k}$. Una volta selezionato, restituiremo il modello $\mathcal{M}_k$.
- Calcoliamo l'errore sul set di validazione o l'errore di cross-validazione per ogni modello $\mathcal{M}_k$ considerato, e poi selezioniamo il k per il quale l'errore stimato di previsione sul test set è il più piccolo.
- Questa procedura ha un vantaggio rispetto ad AIC, BIC, C_p, e adjR^2, in quanto fornisce una stima diretta dell'errore di previsione sul test set, e non richiede una stima della varianza dell'errore σ^2.
- Può essere usata anche in un'ottica più ampia di selezione dei modelli, anche in casi in cui è difficile individuare il numero di gradi di libertà del modello (per esempio, il numero di predittori nel modello) o quando è arduo stimare la varianza dell'errore σ^2.

Dati sui crediti

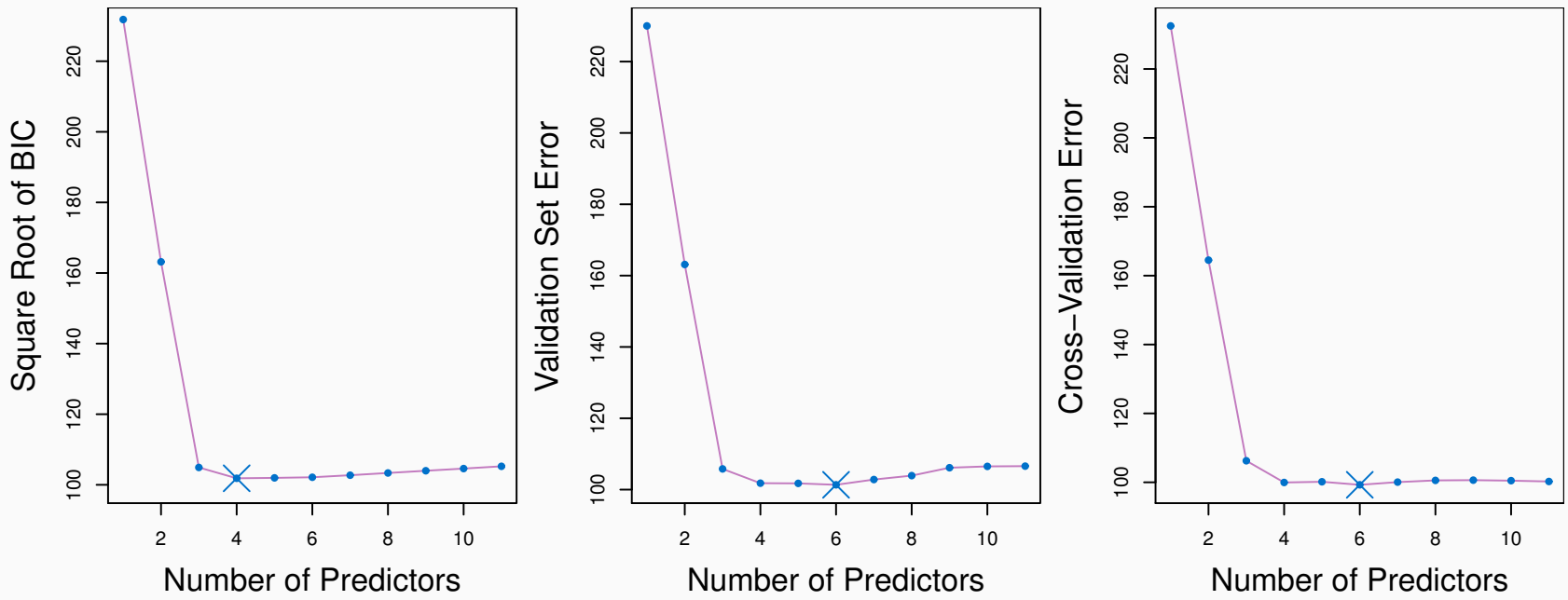

Figure 3: Come si evince, metodi diversi portano a selezione di modelli diversi

Dettagli sui grafici precedenti

- Gli errori sul set di validazione sono stati calcolati selezionando randomicamente tre-quarti delle osservazioni come set di training, e il rimanente quarto come set di validazione.
- Gli errori di cross-validazione sono stati utilizzati usando $k = 10$ gruppi (*folds*). In questo caso, entrambi i metodi di validazione e cross-validazione suggeriscono un modello con sei variabili.
- Tuttavia, tutti e tre i metodi suggeriscono che i modelli con quattro, cinque e sei variabili sono approssimativamente equivalenti per quanto riguarda l'errore di previsione.
- In questo contesto, possiamo selezionare i modelli usando la *regola di uno standard error*: prima calcoliamo lo standard error di ogni test MSE stimato per ogni ampiezza dei modelli, e poi selezioniamo il più piccolo modello per il quale l'errore stimato di previsione rientra in un errore standard dal punto più basso della curva. *Quale è la motivazione per qusta scelta?*